Batch Effects are Causal Effects: Applications in Human Connectomics

Eric W. Bridgeford[1,†], Michael Powell[1], Gregory Kiar[2], Ross Lawrence[1], Brian Caffo[1], Michael Milham[2], Joshua T. Vogelstein[1]

Abstract. Batch effects, undesirable sources of variance across multiple experiments, present a substantial hurdle for scientific and clinical discoveries. Specifically, the presence of batch effects can create both spurious discoveries and hide veridical signals, contributing to the ongoing reproducibility crisis. Typical approaches to dealing with batch effects conceptualize 'batches' as an associational effect, rather than a causal effect, despite the fact that the sources of variance that comprise the batch – potentially including experimental design and population demographics – causally impact downstream inferences. We therefore cast batch effects as a causal problem rather than an associational problem. This reformulation enables us to make explicit the assumptions and limitations of existing approaches for dealing with batch effects. We therefore develop causal batch effect strategies—*Causal* DCORR for discovery of batch effects and *Causal ComBat* for mitigating batch effects – which build upon existing statistical associational methods by incorporating modern causal inference techniques. We apply these strategies to a large mega-study of human connectomes assembled by the Consortium for Reliability and Reproducibility, consisting of 24 batches including over 1700 individuals to illustrate that existing approaches create more spurious discoveries (false positives) and miss more veridical signals (true positives) than our proposed approaches. Our work therefore introduces a conceptual framing, as well as open source code, for combining multiple distinct datasets to increase confidence in claims of scientific and clinical discoveries.

1 Introduction

The 21st century has seen the advent of high-throughput techniques for acquiring data on an unprecedented scale. Collection of these datasets often occurs through consortia using a range of technologies across different sites, requiring data to be brought together by aggregation services and hosted by data banks for storage and sharing. These "mega-studies" comprising numerous individual studies serve the useful purpose of facilitating inference on individuals with unprecedented diversity; researchers can subvert traditional sample size and generalizability issues through the use of data from a more diverse sample of the population than could otherwise be collected.

Unfortunately, aggregating data across diverse datasets introduces an undesirable variability, known as a batch effect. Several strategies have been proposed to define a batch effect. Lazar et al. [14] summarize many of these approaches succinctly: "the batch effect represents the systematic technical differences when samples are processed and measured in different batches and which are unrelated to any biological variation." Unfortunately, this unified description provides limited information about how batch effects manifest, nor how they can be formally characterized and investigated. We have been unable to find a satisfactory technical definition of batch effects in the literature.

Some approaches model the batch collection or measurement process as a nuisance variable which is to be removed [11, 15, 16, 24, 25, 39, 42], which implies batch effect is a particular component of an associational model. These approaches typically make strong, potentially unjustified, and often inappropriate parametric assumptions about the data. Others use ComBat or Conditional ComBat batch effect correction procedures [11]. These approaches have demostrated empirical utility in certain genomics and neuroimaging contexts [25, 43]. However, it remains unclear when these approaches will be successful and when they will fail. Specifically, they could either remove biofidelic variability or fail to remove nuisance variability.

In this work, we develop a unified and systematic approach to the definition, detection, estimation, and mitigation of batch effects. Our main conceptual advance is appreciating that one can view batch effects as causal effects rather than associational effects. Given this understanding, we introduce a formal definition of causal batch effects. This formal definition reveals the limitations of (typically inappropriate) assumptions implicit in existing approaches [28, 29, 33]. Methodologically, we introduce

[1] Johns Hopkins University [2] Child Mind Institute, [†] Corresponding author: Eric W. Bridgeford (ebridge2@jhu.edu).

two new approaches. First, we developed *Causal* DCORR [35]—building on modern non-parametric statistical methods—to estimate and detect the presence of batch effects. Second, we developed *Causal* ComBat—an augmentation of the ComBat procedure—which uses causal methods to remove batch effects while limiting the removal of variation due to known demographic variables. Our main empirical contribution is to apply these methods to a large neuroimaging mega-study assembled by the Consortium for Reliability and Reproducibility (CoRR) [44], consisting of more than 1700 connectome measurements across 24 disparate studies. Our investigations reveal that previous strategies fail to differentiate between veridical biological signal and sources of variability due to other experimental design factors. Further, previous strategies for removing batch effects remove more apparently veridical biological signal than *Causal* ComBat.

We believe that this investigation raises significant questions for the validity of downstream inference with or without applying previously proposed batch effect correction procedures to mega-studies. We hope this work contributes to the ongoing effort to improve the validity and reliability of inferences in past and future mega-studies.

2 Methods

For further details on methodological considerations, see Section 5.

2.1 Formally defining batch effects from a causal perspective

Throughout this paper, we use P_X to denote a generic distribution function and f_X to denote a generic probability density or mass function for a random variable X, which we abbreviate P and f for simplicity. We use $P_{X|y}$ to denote the conditional distribution function of the random variable X conditioned on $Y = y$. The corresponding conditional density/mass function is $f(Y = y|X = x)$, which we abbreviate $f(y|x)$. We let Y denote the $y \in \mathcal{Y}$-valued random variable denoting the outcome, T denote the $t \in \{1, \ldots, K\} = [K]$ categorical-valued random variable denoting the batch identity assignment (the *exposure*), x the $x \in \mathcal{X}$-valued random variable denoting measured covariates, and Z the $z \in \mathcal{Z}$-valued random variable denoting unmeasured covariates. We let $\mathbb{P}[T = t|X = x, Z = z]$ denote the probability mass function of the random variable T taking the value t, conditional on the measured and unmeasured covariates X and Z taking values of x and z, respectively. Throughout, we assume that (Y, T, X, Z) are sampled identically and independentally from some true but unknown distribution. Together, all possible confounding and indirect associations between Y and T can be accounted for by the tuple (X, Z); meaning that X alone may not be sufficient to account entirely for confounding (X and Z *close backdoor paths*). $f(y|t, x, z)$ denotes the true outcome model; note the true outcome model is conditional on both measured *and* unmeasured covariates. In our investigation, the outcome Y is an fMRI connectome, the batch T is the study in which the connectome was acquired, and X are covariates regarding the individual which are known (age, sex, and continent of measurement).

We are interested in estimating the effect of different exposures on the outcome of interest, which can be quantified using the backdoor formula under the assumption that X and Z close all backdoor paths [17, 23, 24, 27]:

$$\forall\, t \in \{1, \ldots, K\} : \qquad f_{x,z}(y|t) = \int_{\mathcal{X}\times\mathcal{Z}} f(y|t, x, z) f(x, z)\, \mathrm{d}(x, z). \tag{1}$$

So our causal estimands will be functions of the set of $\{f_{x,z}(y|t)\}_{t\in\{1,\ldots,K\}}$. Note that the above equation integrates over the *unconditional* distribution of X and Z, $f(x, z)$. This quantity can therefore be thought of as the distribution of the true outcome averaged over *all* measured and unmeasured covariates. This contrasts from the notationally similar, but practically distinct, quantity $f(y|t)$, e.g.:

$$f(y|t) = \int_{\mathcal{X}\times\mathcal{Z}} f(y|t, x, z) f(x, z|t)\, \mathrm{d}(x, z)$$

The key difference is that unlike Equation (1), $f(y|t)$ averages the true outcome distribution, $f(y|t, x, z)$, over the *conditional* distribution of the measured and unmeasured covariates, $f(x, z|t)$. This has the

implication that if $f(x, z|t)$ is not the same as $f(x, z|t')$, $f(y|t)$ represents an average over a different covariate distribution than $f(y|t')$.

Batch Effect Given the above, we formally define the **batch effect** as a causal effect using the back-door formula (1).

Definition 1 (Batch Effect). *Suppose the setup described above. A batch effect between batches t and t' exists if:*

$$f_{x,z}(y|t) \neq f_{x,z}(y|t').$$

In the below sections, we will elucidate the specific sets of progressively less restrictive assumptions about (Y, T, X, Z) that will enable us to estimate batch effects. The causal claims are only valid insofar as the assumptions upon which they rest adequately characterize the data.

Associational Effect We only observe the pairs (y_i, t_i) for $i \in [n]$. Therefore we will only be able estimate functions of (Y, T).

Definition 2 (Associational Site Effect). *An associational site effect exists between batches t and t' if:*

$$f(y|t) \neq f(y|t'). \tag{2}$$

When would an associational site effect be equivalent to a causal batch effect? A sufficient assumption for the equivalence of those two effects is that the measured and unmeasured covariates (X, Z) are non-confounding (which is satisfied, for example, whenever (Y, T) is independent of (X, Z)). This assumption, however, is not often a reasonable one to make. Consider, for instance, if in one batch t, we tend to measure older individuals, whereas in another batch t', we tend to measure younger individuals. If age is related to the measurement we obtained, then the differences between $f(y|t)$ and $f(y|t')$ could be due to age or batch identity, and we have no way of differentiating whether the effect is a *bona fide* batch effect versus an associational effect.

Conditional Effect We observe triples (y_i, t_i, x_i) for $i \in [n]$; thus, we will only be able estimate functions of (Y, T, X).

Definition 3 (Conditional Site Effect). *A conditional site effect exists between batches t and t' if:*

$$f(y|t, x) \neq f(y|t', x).$$

When would a conditional site effect be equivalent to a causal batch effect? To answer this question, we first define the propensity score $e(t|x, z)$ [28, 29]; that is, the probability of selection into a particular exposure given one's covariates: $e(t|x, z) = \mathbb{P}[T = t_i|X = x_i, Z = z_i]$. With this definition in hand, a sufficient assumption for the equivalence of those two effects is that both of the following conditions must hold:

1. Conditional on the measured and unmeasured covariates, (X, Z), the exposure is independent of the outcome: $T \perp\!\!\!\perp Y|X, Z$.
2. The covariate distributions overlap; that is, for some $c > 0$, and all $(x, z) \in \mathcal{X} \times \mathcal{Z}$, $c \leq e(t|x, z) \leq 1 - c$, for all $t \in [K]$.

Together, these two conditions are referred to as strong ignorability [28]. If the unmeasured covariates Z are non-confounding, they can be ignored in the above. Note that if the covariate distributions fail to overlap for a pair of exposures, it is impossible to differentiate whether the differences between the outcomes are due to the different covariate distributions or different batches.

In practice, we never know whether the true unmeasured propensities overlap; we can only estimate them from the data. If our estimated propensities do not overlap given finite data, we again cannot

reasonably differentiate between differences in the two groups being due to *bona fide* batch effects or empirical differences in the covariate distributions. This motivates a third approach.

Adjusted Effect We again observe triples (y_i, t_i, x_i) for $i \in [n]$. Prior to assessing the equality of any distributions, however, we weight the observations such that the observed covariate distributions are rendered approximately equal; that is, $\widetilde{f}(x|t) \approx \widetilde{f}(x|t')$ for a given $t \neq t'$. These sample adjustments yield adjusted conditional distributions $\left(\widetilde{f}(y|t,x), \widetilde{f}(y|t',x)\right)$ for all pairs $t \neq t'$.

Definition 4 (Adjusted Site Effect). *An adjusted site effect exists between batches t and t' if:*

$$\tilde{f}(y|t,x) \neq \tilde{f}(y|t',x).$$

When would an adjusted site effect be equivalent to a causal batch effect? The overlap assumption stated above is satisfied by construction. Thus, the exposures must be independent of the outcome conditional on any confounding covariates, X and/or Z. Again, if everything is independent of Z, this detail need not be considered. The adjusted site effect also has limitations. For example, imagine that one batch has more individuals that have had a stroke, but stroke history is unmeasured. Stroke history would be critical to observe if it impacts the outcome. Thus, such effects are only causal under reasonably strong assumptions, specifically that there is no unmeasured confounding. To this end, we develop Causal `DCORR` to estimate adjusted effects and Causal `ComBat` to remove adjusted effects, both of which are discussed below.

Crossover Site Effect Here we observe outcomes, exposures, and measured covariates again, but we obtain multiple measurements for each participant, or more specifically, one measurement for each exposure. Thus, we observe for each participant, $\{(y_i^{(t)}, t, x_i^{(t)})\}$ for all $t \in [K]$.

Definition 5 (Crossover Effect). *A crossover effect exists between batches t and t' if:*

$$f\left(y^{(t)}|t, x^{(t)}\right) \neq f\left(y^{(t')}|t', x^{(t')}\right).$$

When would a crossover effect be equivalent to a causal batch effect? We are certain that any *traits* of the participants (i.e., variables that are constant for a given participant, such as genome) are the same across the two groups since the group members are identical. However, *states* may differ as they may be functions of location or time. Therefore, for a crossover effect to be equivalent to a causal batch effect there must be no unmeasured confounding states. For example, if being measured impacts subsequent states, then a crossover effect may not be indicative of a causal batch effect without further assumptions and/or experimental design considerations (such as randomizing exposure order across participants).

2.2 Batch effects in biological datasets

The four effects discussed above have immediate practical implications on measuring batch effects in modern biological datasets. To this end, we demonstrate the applicability of these effect types in the context of a neuroimaging mega-study, in which measurements are collected from individuals across various batches (the *exposures*). The logical process, however, extends beyond this context to many other biological data modalities, such as genomics. Figure 1A shows a directed acyclic graph, or DAG, which underlies a typical mega-study. The exposure is the batch in which participants' data are collected, and the outcomes are the measurements themselves. We seek to estimate the *causal estimand*, which is the causal effect of the exposure to a particular batch on the outcome (the *measurement*). Numerous covariates, both known and unknown, impart potentially confounding biases, in which the variables can influence both the exposure and the outcome. For instance, one batch collected in a children's hospital may tend to over-sample younger individuals relative other batches, and younger children may have distinct patterns of brain connectivity

from older children. Unmeasured covariates are often prominent confounders in biological datasets, making proper causal analyses troublesome. *Participant history* involves variables related to specific life experiences, both environmental and societal, that the individual has been exposed to. *Participant traits* are characteristics of a person that remain relatively constant over time. *Participant states* refer to characteristics of a person that change over an individual's lifetime. *Exposure determinants* are factors which impact batch membership for a particular individual, but do not impact the measurement itself. *Outcome determinants* are factors which do not impact the batch assignment, but impact the outcome.

Figure 1B illustrates the implicit DAGs that are sufficient such that if they characterized the data, then each of the four effect types discussed above become valid causal batch effect estimands. If no covariates exist (observed or unobserved), then an associational effect is a causal effect (Figure 1B.I). Crossover effects can account for many types of measured and unmeasured covariates which are confounding (Figure 1B.III). Because of the crossover property, the batches will be balanced across certain measured and unmeasured covariates, specifically patient history and traits. Hence, many of the confounding variables from Figure 1A are deconfounded. However, unmeasured changes in participant state between different exposure groups (such as measuring individuals only during the daytime in one batch and only during the night time in another batch) remain potential confounders if not carefully balanced across exposure groups through either randomization or direct experimental intervention.

3 Results

3.1 The CoRR Studies share demographic characteristics Figure 2A explores the demographic characteristics for the individuals in the CoRR mega-study. Many of the studies have a narrow age range, and several studies only include females. Because sex [10, 31, 40], age [7, 30, 37], and continent [18] are variables that have been associated with brain connectivity, they serve as conditional covariates used in our investigation.

Figure 2B investigates connectomes from individual measurements for different demographic sets within our study. Each connectome exhibits homotopy—in which edges incident parcels of the same region share greater connectivity than edges incident parcels of disparate regions—as indicated by the off-diagonal band within the individual connectomes. Also, it is clear that individual connectomes vary widely within a single study.

3.2 Detecting batch effects using Causal Dcorr At present, literature approaches for detecting the presence of batch effects are limited. Many of the more direct types of detectable effects, such as associational and conditional effects, fail to adaequately account for confounding biases present in the data. We instead propose the use of Causal DCORR, in which a conditional two-sample test is performed on pairs of observational studies which are balanced against measured demogrphic covariates.

Causal DCORR is performed as follows. Given measurements and covariates from n individuals across 2 batches, each with n_l individuals:

1. Identify the smallest dataset k' as the exposure dataset and the other dataset k as the control dataset.
2. Match the largest integer less than $n_k/n_{k'}$ control individuals using nearest-neighbor matching to individuals in the exposure dataset. Methods 5 discusses the adjustment procedure in more detail.
3. Discard individuals from the exposure dataset who do not have matches in the control dataset, yielding the exposure cohort.
4. Discard individuals from the control dataset who are not matched to individuals in the exposure cohort.
5. Perform Conditional DCORR [34, 35] on the measurements of the matched exposure and matched control individuals, conditioned on the measured covariates X. The null hypothesis is that the two batches have the same distribution, against the alternative hypothesis that

the two batches do not have the same distribution, conditional on the measured covariates X. In the event that the conditioning set closes backdoor paths [23, 24], the adjusted effect detected by Causal DCORR is a causal batch effect, and does not require extrapolation assumptions like Conditional DCORR [9, 28, 29, 33]. If the conditioning set does not close backdoor paths, the effect detected is a conditional effect and may yield the detection of demographic effects.

Figure 3A explores the four effects defined above in fMRI connectomes from the CoRR mega-study. The majority of associational and conditional effects (>90$\%$) are significant (partial distance correlation two-sample test, BH correction, $\alpha = 0.05$). However, these approaches are overly simplistic for concluding that a causal batch effect is indeed present because they fail to account for non-overlap of the covariate distributions between the subjects.

Effects that account for demographic distributions are examined in Figure 3B. Many pairs of studies could not have an adjusted effect estimated due to poor demographic alignment, which resulted in no suitable matches existing between the studies, indicated by a red X (279 of 406 pairs of studies). Notably, adjusted conditional effects are estimable between all pairs of the American Clique (a group of separate studies consisting of similar demographics collected in North America). After adjustment for the sample characteristics of the demographics associated with individual studies, a majority of adjusted effects (>90$\%$) remain significant (partial distance correlation two-sample test, BH correction, $\alpha = 0.05$). The NKI Rockland Sample takes this a step further, allowing a crossover effect to be determined in a few pairs of batches (purple squares). This study was, in fact, collected on an identical set of participants, in the same location, with the same research technicians, with data varying only in the repetition time for the magnet used in the scanner. Indeed, for two of three comparisons, crossover effects remain prominent. This provides evidence that the parameters under which a study are conducted is a causal predictor of measured connectomes, assuming no participant state confounders are present.

3.3 Mitigating batch effects using Causal ComBat

Unfortunately, many existing techniques for the removal of batch effects fail to adequately account for confounding biases that may be present in the data. We instead propose the use of Causal ComBat, in which Conditional ComBat is performed on a subset of observational studies in which all pairs of studies are balanced against measured demographic covariates. Causal ComBat is performed as follows. Given measurements and covariates from n individuals across K batches, each of which has n_k individuals:

1. Identify the smallest dataset k' as the exposure dataset and all $k \neq k'$ datasets as control datasets.
2. For each control dataset k, match the greatest integer less than $n_k/n_{k'}$ control individuals using nearest-neighbor matching to individuals in the exposure dataset. Methods 5 discusses the adjustment procedure in more detail.
3. Discard individuals from the exposure dataset who do not have matches in all K control datasets, yielding the matched exposure cohort.
4. Discard individuals from the control datasets who are not matched to the matched exposure cohort, yielding the matched control cohorts.
5. Perform Conditional ComBat [11] on the measurements of the matched exposure and matched control individuals across the K batches, conditioned on the measured covariates X.

In the event that the conditioning set closes backdoor paths [23, 24], Causal ComBat yields the removal of an internally valid causal effect and does not require extrapolation assumptions, unlike Conditional ComBat [9, 28, 29, 33]. If the conditioning set does not close backdoor paths, the effect removed is a conditional effect and may potentially yield the removal of demographic effects. Appendix B depicts the impact on the empirical covariate distribution of the adjustment procedure.

Batch-Related Effects are Reduced with Causal ComBat Here we investigate batch-related effects before and after correction. Figure 4 shows average connectomes from the MRN1 and IBATRT studies. With no correction, while the average connectomes appear similar between the two studies, a difference

between the average connectomes is present (bottom row). Three different batch effect mitigation strategies (Z-scoring across features within a single batch, `ComBat`, and Causal `ComBat`) all appear to eliminate the differences between the average connectomes. Z-scoring has disrupted fundamental topological properties of the human connectome, as explored in detail in Appendix C. For this reason, we do not consider Z-scoring going forward, and focus instead on variations of the `ComBat` procedure.

Causal `ComBat` Produces Disparate Inference from Competing Correction Approaches For each edge in the connectome, we investigate the strength of the sex effect, conditional on age, in the connectomes before and after batch effect correction (Figure 5A). `ComBat` and Conditional `ComBat` are performed on the entire CoRR study (top row), and subsequently all techniques are restricted to the subset of connectomes upon which Causal `ComBat` was executed (the *American Clique*, bottom row). The edges which show a significant, conditional sex effect (partial distance correlation, $\alpha = .05$) are colored from smallest test statistic to largest test statistic (rank = 1). Causal `ComBat` produces vastly different statistical inference compared to other correction techniques on the American Clique. Causal `ComBat` was unable to be executed on the set of all studies, because many of the studies in the CoRR dataset do not overlap in terms of their covariate distributions (the red square indicates the analysis is not possible). Next, we seek to answer whether statistical inference for the entire CoRR study (top row) is similar to inference that would be obtained on the American Clique (Figure 5B) using only Causal `ComBat`. We compare the DICE overlap of the top n edges (by effect size) from each of the three non-causal approaches applied to all the data with the top n edges of Causal `ComBat` on the American Clique. A high DICE overlap indicates that the subsets of edges are similar. None of the techniques show high overlap with the edges that are indicated to have a large effect after correction with Causal `ComBat`. Finally, using the raw connectomes, `ComBat`, and Conditional `ComBat` tend to produce similar inferences, as all have high DICE overlaps near one. In contrast, they all have low overlap with Causal `ComBat` (Figure 5C), with DICE overlap with Causal `ComBat` nearly zero.

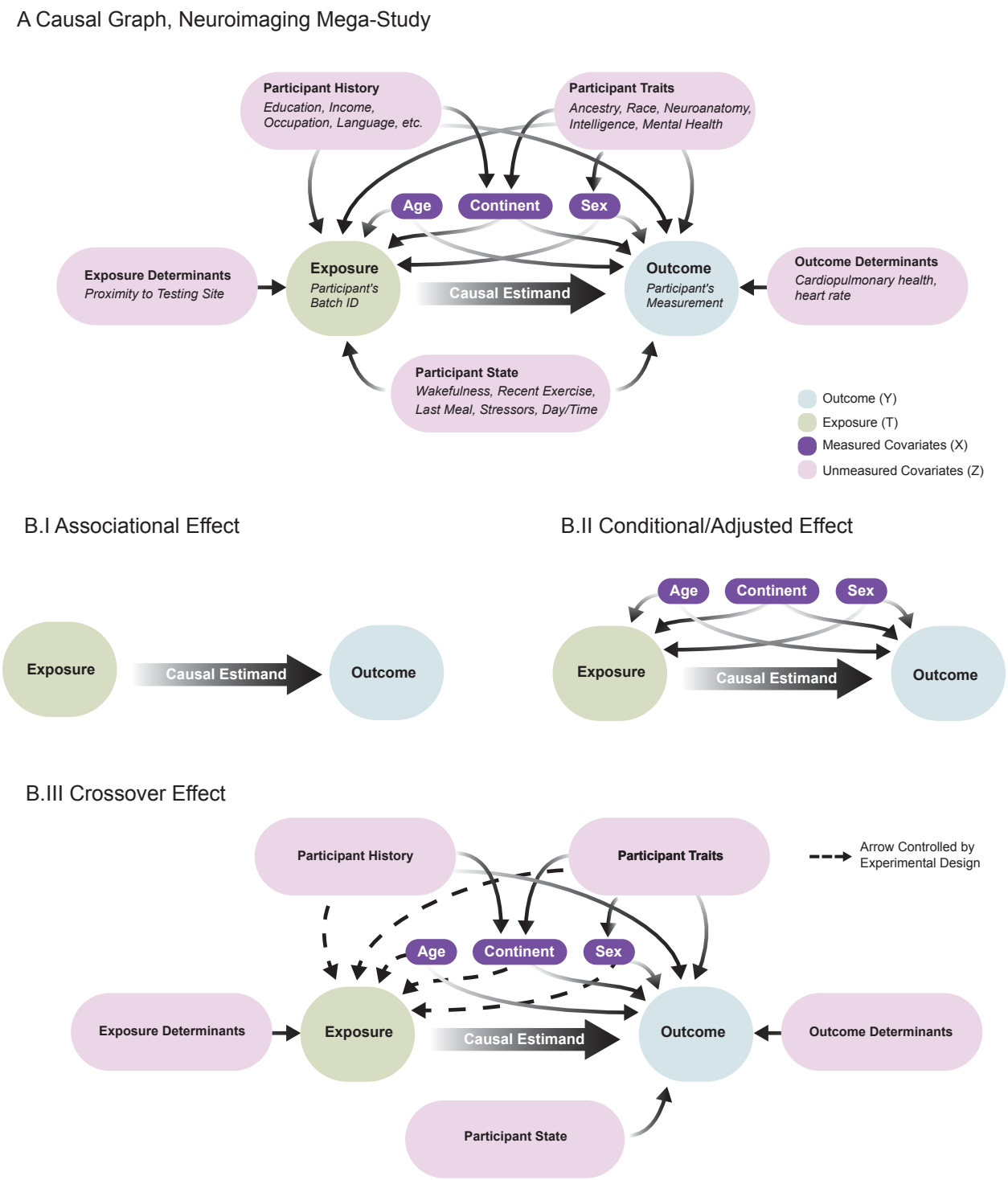

Figure 1: **Causal Graph of Study Covariates**. **(A)** A causal directed acyclic graph (DAG) illustrating potential relationships which can reasonably be assumed to underlie a mega-study. Descriptions reflect possible attributes that might fall into the specific exposures, outcomes, and covariates for a neuroimaging study, for illustration purposes. **(B)** DAGs which illustrate the assumptions which underly assorted effects in order for those effects to be causal. Associational effects **(B.I)** are causal when all external covariates are independent the exposure and outcome. Conditional effects and adjusted effects **(B.II)** are causal when the strong ignorability condition holds, and in particular when the measured covariates close backdoor paths. Crossover Effects **(B.III)** are causal when participant states are non-confounding between the exposure and the outcome. The experimental design of a crossover study ensures that other forms of measured and unmeasured covariates are non-confounding.

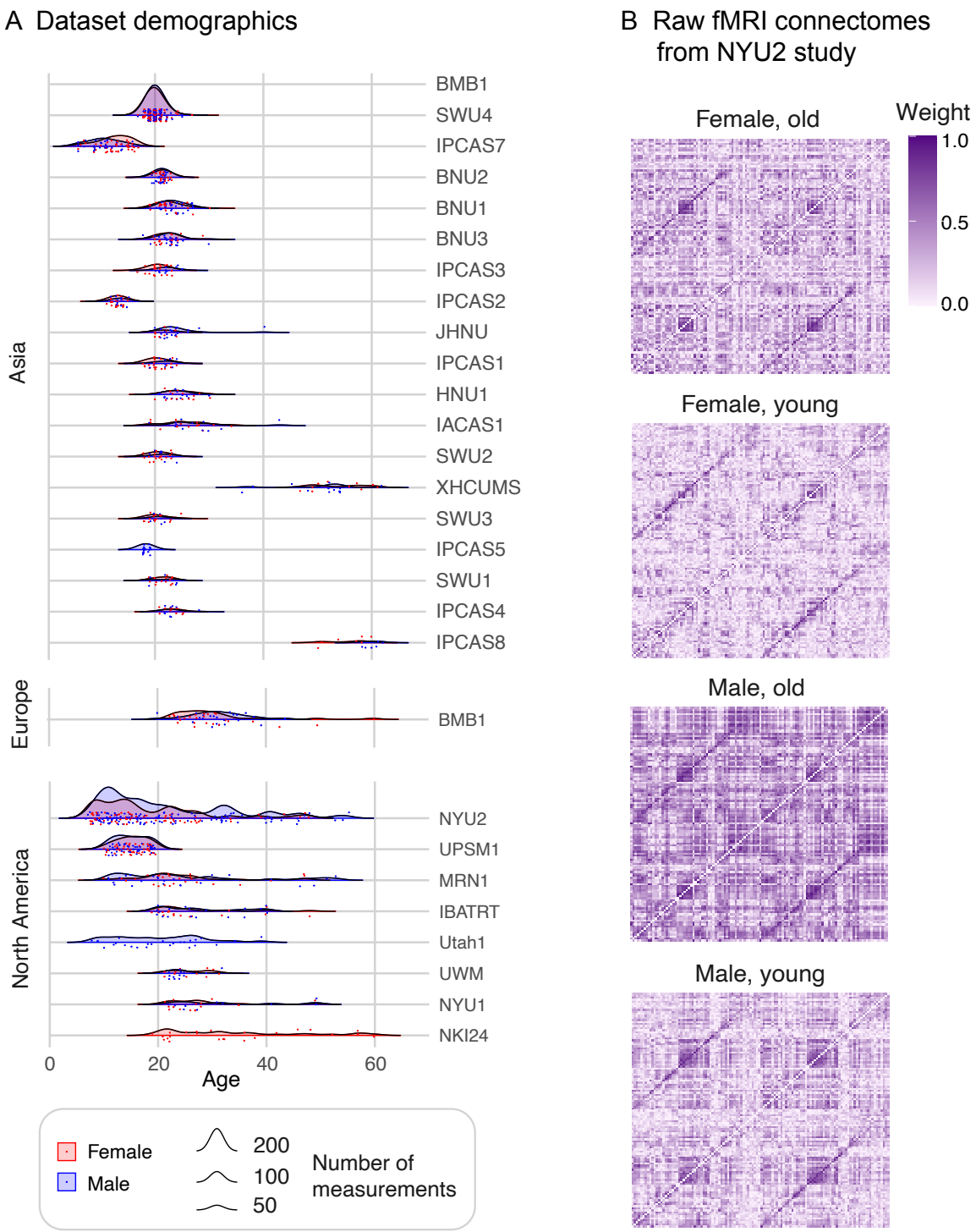

Figure 2: Demographic data for each of the $N = 24$ studies from the CoRR Mega-Study. **(A)** Each point represents the age of a participant corresponding to a single measurement, rows are studies, boxes are continents, and color indicates sex. **(B)** Comparison of connectomes across disparate demographics. Each panel shows an example connectome from the NYU2 study from a disparate demographic group, indicating how the groups share certain similarities (e.g., they each exhibit homotopy), while also being different.

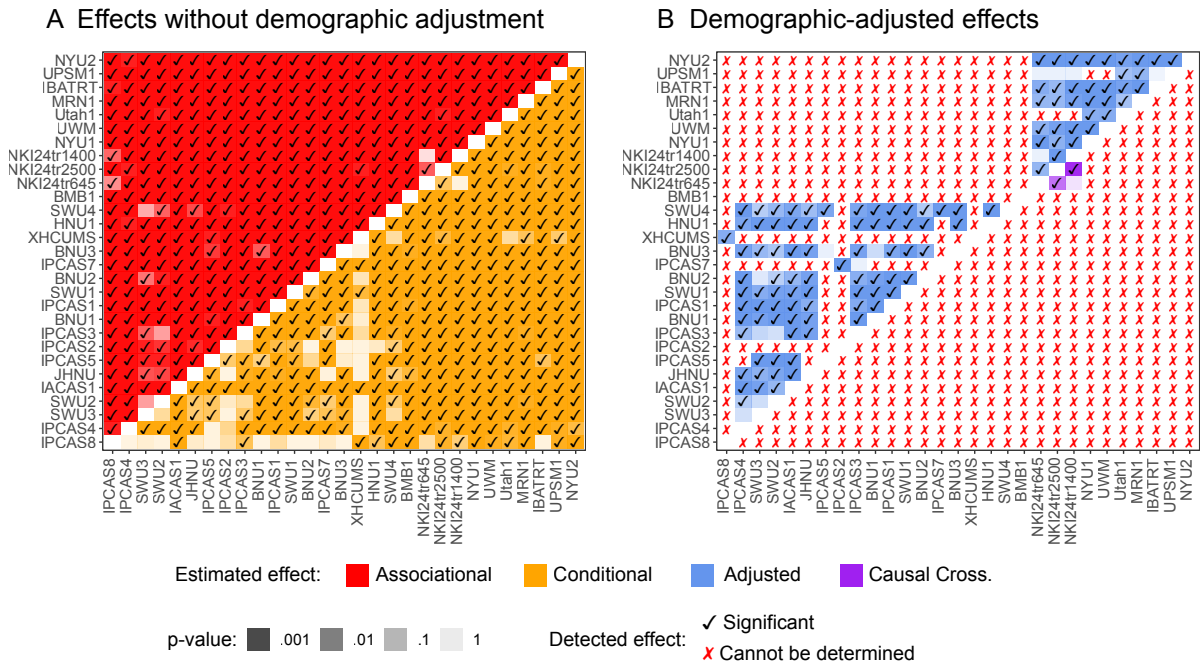

Figure 3: **Comparison of Estimated Effects for CoRR Mega-Study**. Heatmaps indicating the p-value of effect estimates per pair of studies. Studies are organized by continent, followed by the number of samples in the study. Comparisons which are significant at $\alpha = .05$ are indicated by a black check. Since effects are symmetric (an effect between batch t and t' is equivalent to an effect between batch t' and t), we collapse symmetric heatmaps into a single triangle for each effect type. **(A)** Effects which can be estimated but do not make adjustments for demographics. Associational and conditional effects can always be estimated between pairs of studies. (B) Effects which account for demographics through covariate adjustment. Squares in which a causal effect was not estimable are indicated by a red X.

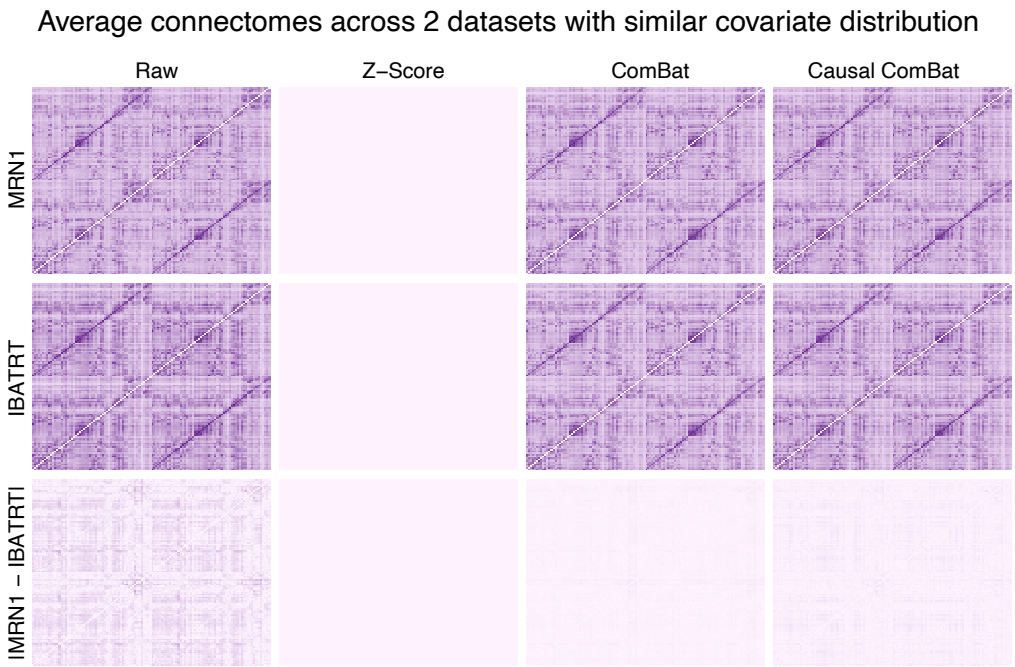

Figure 4: **Distribution of Detected Effects Before and After Correction**. A comparison of the average connectomes from two studies from the same continent, with similar sex and age distributions conditional on sex. While the *Raw* average connectomes appear similar, the absolute difference between the two is appreciable. All of the three adjustment strategies reduce the absolute difference between the average connectomes for the two batches.

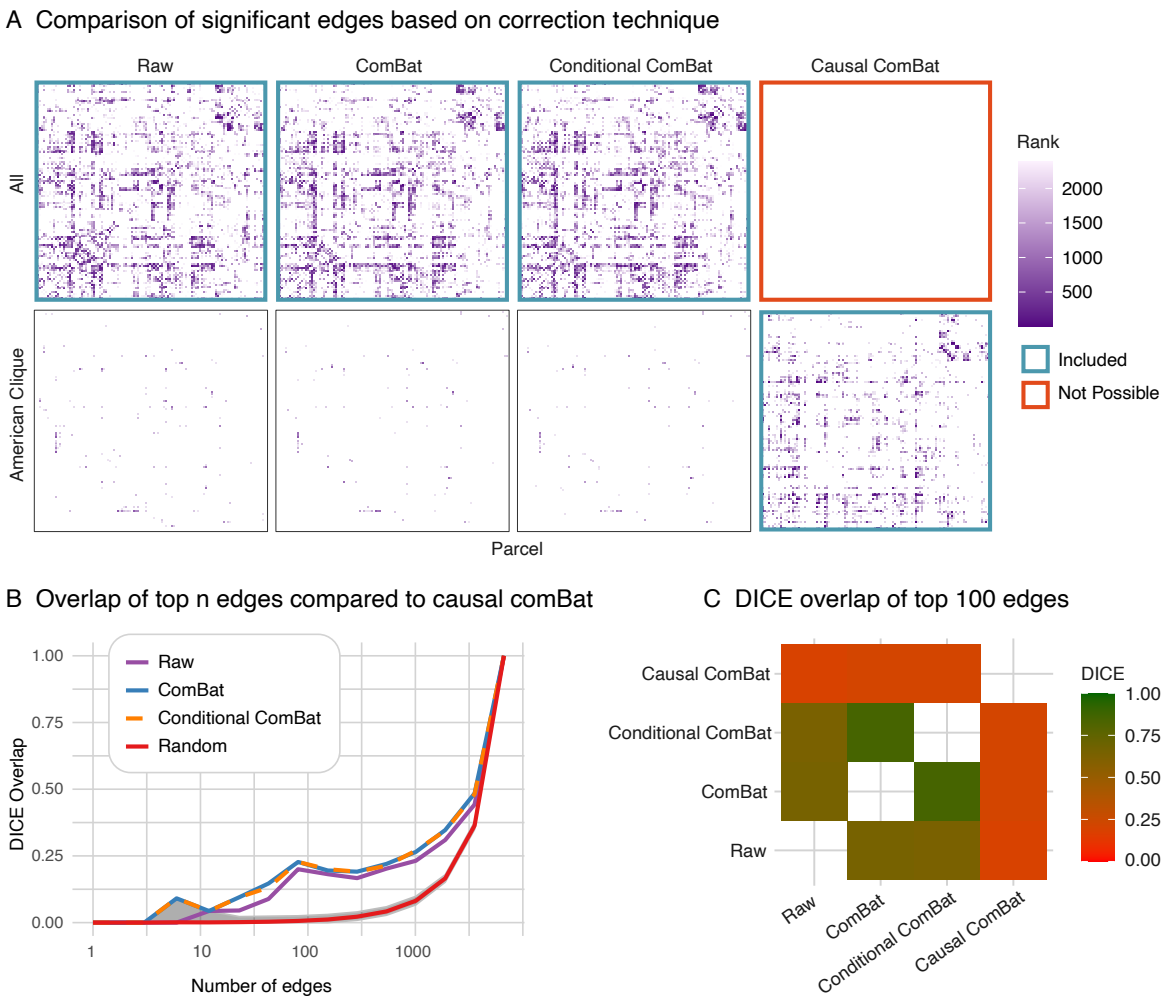

Figure 5: **Significant Edges before and after Batch Effect Removal**. **(A)** The presence of a sex effect (conditional on individual age) is investigated for each edge in the connectome. The top 100 edges are shown in rank-order from largest (rank = 1) to smallest sex effect. Analysis is either performed over *all* individuals (top row) or only upon the individuals for which sufficient demographic overlap could be obtained, the *American clique* (bottom row). Causal `ComBat` can only be executed on the American Clique (red square denotes Causal `ComBat` is not possible with *all* studies). **(B)** The DICE overlap of the top n edges of each batch effect removal technique on all the data (green squares, top row) as compared to those from Causal `ComBat` (green square, bottom row). The red line (random) indicates the overlap to be obtained for a particular choice of n to be expected as a result of random chance, and the gray ribbon indicates the 99% confidence interval of the estimate (1000 repetitions). The DICE overlap increases as a function of n, regardless of whether any signal is actually present. **(C)** The DICE overlap of the top 100 edges for all pairs of correction techniques indicated with a green square from **(A)**. Overlap of the top signal edges is high between no correction (raw), `ComBat`, and Conditional `ComBat`, but all techniques have low overlap with Causal `ComBat`, indicating that Causal `ComBat`detects signal that the associational approaches fail to identify.

4 Discussion Our primary conceptual contribution is establishing that batch effects can be formalized as causal effects. Given this realization, we propose a framework for quantifying and removing batch effects in high-dimensional datasets featuring non-Euclidean measurements. We propose a non-parametric test for quantifying batch effects, Causal DCORR, and develop a strategy for the removal of batch effects, Causal ComBat, to limit confounding biases. Finally, our neuroimaging use-case shows the utility of the methods developed herein for estimating and removing batch-related effects from large biological datasets. Specifically, our methods identify different causal effects as compared to prior approaches, each of which identifies features similar to a completely unadjusted approach, which is well established in the causal literature to permit numerous false positives and false negatives [4, 6]. Together, our work rigorously defines and studies measurement bias decoupled from demographic confounders in complex, non-Euclidean data.

We find that Z-scoring, ComBat, and Conditional ComBat have substantial limitations empirically and theoretically in the context of removing batch effects. While straightforward, Z-scoring tends to eliminate topological signal from connectomes, which reduces the interpretability of the connectome in downstream analyses. Both ComBat and Conditional ComBat also exhibit shortcomings in the removal of batch effects. Specifically, one can endeavor to mitigate batch effects in one of two complementary ways. First, one can filter individuals by ensuring demographic balancing, and then run the analysis only on those individuals (as Causal ComBat does). Second, one can run an analysis on all individuals, and then filter to ensure demographic balancing only when estimating the causal effect sizes (as we did by running ComBat on the American clique). These two approaches yield radically different answers, as evidenced by Figure 5.

Together, these conclusions bring into focus the pervasive question of how to properly interpret neuroimaging—or more generally, scientific—studies.This concern is at the root of any statistical inference task: a conflict between the internal and external validity. Internal validity refers to how well the conclusions generalize from the *sample* population to the *study* population; that is, from the observed samples to other samples that are distributed the same as the observed samples. External validity refers to the degree to which the conclusions generalize to the *target* population; that is, the population we would like to draw inferences. The target population is typically much broader than the study population, to include, for example, all members of the species, or possibly even multiple species. The trade-off of subsampling individuals is that sample size, sample diversity, and external validity are exchanged for internal validity. In this study, we chose to prioritize internal over external validity, as internal validity is a prerequisite for external validity. Further work is needed to better understand how the removal of batch effects can be achieved while mitigating the deleterious effects on either internal or external validity.

Biomedical data science, including neuroimaging and connectomics, are plagued by a lack of generalizability and replicability [2, 36]. While many investigations posit that this may be a consequence of small sample sizes or overly simplistic datasets [20, 36] lacking sufficient subject-specific measurements, we argued here that a lack of understanding regarding batch effects, and more fundamentally *causality*, further exacerbates these issues. This work suggests that harmonized analyses should be conducted with both harmonized measurement parameters and demographic distributions. This represents a departure from the manner in which many mega-studies, including so-called harmonized effects, are currently generated [44], in which potentially demographically-disparate populations are analyzed and only loosely harmonized on the basis of the measurement protocols. Future work will focus on applying the methods developed here to harmonize mega-studies such as the Adolescent Brain Cognitive Development (ABCD) [12], which includes $N > 12,000$ demographically similar individuals from across the United States, and measurements were produced with a consistent, harmonized measurement protocol. Further, future studies may seek to build upon the NKI-RS and the recent work by Noble et al. [21] by collecting measurements from the same expansive and diverse group of people across multiple neuroimaging sites or scanning protocols. Our work provides a theoretical foundation

for evaluating existing neuroimaging studies, such as Noble et al. [21], Yamashita et al. [41], which advocate for the aggregation of studies featuring individuals measured at a wide array of sites, so that the nature of batch effects can be explored with minimal assumptions needed due to the crossover property. The existence of such datasets would allow a far better understanding of the prevalence and manifestation of batch effects. Failure to attain a comprehensive understanding of batch effects and the extent to which they pervade statistical inferences, we believe, calls into question the validity and reproducability of existing results in neuroimaging.

Code Availability The code used for analysis within this manuscript is available at github.com/neurodata/batch_effects.

Data Availability The raw data analyzed in this manuscript can be obtained at CoRR Mega-Study. The pre-processed data analyzed in this manuscript is available at neurodata.io/mri.

Acknowledgements The authors are grateful for the support from the National Science Foundation (NSF) administered through NSF Career Award NSF 17-537, the National Institute of Health (NIH) through National Institute of Mental Health (NIMH) Research Project 1R01MH120482-01, and the NIH through Research Project RO1AG066184-01.

5 Methods

5.1 Covariate Adjustment The exposed group t is selected to be the smaller of the two groups, and the unexposed group t' is selected to be the larger of the two groups, where n_t is the number of individuals in the exposed group and $n_{t'}$ is the number of individuals in the unexposed group. The "covariate overlap" procedure attempts to ensure positivity of the propensity distribution for the unexposed group (i.e., $e(t'|x) > 0$). Intuitively, we exclude individuals from the unexposed group who do not appear "similar" to individuals in the exposed group. In this work, covariate overlap is established through logistic regression by excluding individuals in the control group with a propensity for exposure lower than 0.01. Similarly, the "covariate balancing" procedure attempts to re-weight observations in the per-batch covariate distributions such that the covariate distributions are approximately equal (i.e., $f(x|t) \approx f(x|t')$). In this work, we perform $k:1$ nearest-neighbor matching using the `MatchIt` package [9]. The number of matches $k = \lfloor \frac{n_{t'}}{n_t} \rfloor$ is chosen to be the largest number of unexposed matches possible. Individuals are balanced on the basis of individual sex, individual age, and individual continent of study. As there are likely many other categories with which brain connectivity may be confounded, we do not believe this covariate set is sufficient to identify a Causal Batch Effect 1, as we would need to be confident that these covariates exhibited the covariate sufficiency property. We perform exact matching on the basis of individual sex and individual continent of measurement, and use a 0.1-standard deviation caliper on the propensity score to obtain at most k matched participants for each treated individual [26].

5.2 Hypothesis Testing Recall that statements of the form $f(y) = g(y)$ against $f(y) \neq g(y)$ are equivalent to $P_f = P_g$ against $P_f \neq P_g$, as probability densities uniquely define distribution functions. Therefore, hypotheses for the effects described in Section 2.1 are natural two-sample and conditional two-sample testing procedures. A natural test statistic for the two-sample testing procedure is the Distance Correlation [35], which is a non-parametric test for testing whether two variables are correlated. A simple augmentation of the distance correlation procedure [32, 38] shows that `dcorr` can be used for the two-sample test, or a test of whether two samples are drawn from different distributions. `dcorr` is exactly equivalent in this context to the Maximum Mean Discrepency (MMD), which embeds points in a reproducing kernel Hilbert Space (RKHS) and looks for functions over the unit ball in the RKHS that maximize the difference of the means of the embedded points. When we instead consider the conditional two-sample test (i.e., a test of $f(y|x) = g(y|x)$ against $f(y|x) \neq g(y|x)$), we instead use the partial distance correlation [34], a kernel-based approach in which the points are embedded in a new, non-linear Hilbert Space, which augments the traditional linear Hilbert Space used in distance correlation to allow the definition of the squared distance covariance. Below, we let $\mathbf{Y} = (y_i) \in \mathcal{Y}^n$ denote realizations across both samples and $\vec{t} = (t_i) \in \{t, t'\}^n$ indicate from which sample each realization is drawn. $\mathbf{X} = (x_i) \in \mathcal{X}^n$ denotes covariates which are known about the objects of interest.

Associational Effect We have the following null and alternative hypotheses:

$$H_0 : f(y|t) = f(y|t') \text{ against } H_A : f(y|t) \neq f(y|t')$$

A test of the preceding hypotheses is performed using the distance correlation, and the natural test statistic is $\texttt{dcorr}(\mathbf{Y}, \vec{t})$.

Conditional Effect We have the following null and alternative hypotheses:

$$H_0 : f(y|t, x) = f(y|t', x) \text{ against } H_A : f(y|t, x) \neq f(y|t', x)$$

A test of the preceding hypotheses is performed using the partial distance correlation, and the natural test statistic is $\texttt{pdcorr}(\mathbf{Y}, \vec{t}|\mathbf{X})$.

Adjusted Conditional Effect We have the following null and alternative hypotheses:

$$H_0 : \tilde{f}(y|t, x) = \tilde{f}(y|t', x) \text{ against } H_A : \tilde{f}(y|t, x) \neq \tilde{f}(y|t', x)$$

Unlike the preceding tests, we instead consider the data $(\tilde{\mathbf{Y}}, \vec{\tilde{t}}, \tilde{\mathbf{X}})$, which are the measurements, sample indicators, and covariates of the n realizations after covariate adjustment. A test of the preceding hypotheses is performed using the partial distance correlation, and the natural test statistic is $\texttt{pdcorr}(\tilde{\mathbf{Y}}, \vec{\tilde{t}}|\tilde{\mathbf{X}})$.

Causal Crossover Effect We have the following null and alternative hypotheses:

$$H_0 : f\left(y^{(t)}|t, x^{(t)}\right) = f\left(y^{(t')}|t', x^{(t')}\right) \text{ against } H_A : H_0 : f\left(y^{(t)}|t, x^{(t)}\right) \neq f\left(y^{(t')}|t', x^{(t')}\right)$$

If the known covariates are identical between batches t and t', we test the preceding hypotheses using the distance correlation, and the natural test statistic is $\texttt{dcorr}(\mathbf{Y}, \vec{t})$. If the known covariates are not identical between batches t and t', we test the preceding hypotheses using the partial distance correlation, and the natural test statistic is $\texttt{pdcorr}(\mathbf{Y}, \vec{t}|\mathbf{X})$.

More Than Two Batches The above approaches generalize sufficiently to K batches using K-sample testing approaches. With f_k for $k \in [K]$ denoting the densities associated with K sites or batches, this motivates hypotheses of the form:

$$H_0 : f_k = f_l \text{ for all } k, l \text{ against } H_A : f_k \neq f_l \text{ for some } k \neq l,$$

which can be tested using the distance correlation or the partial distance correlation as above, with the caveat that $\vec{t}$ becomes the matrix $\mathbf{T} = (t_{i,k}) \in \{0, 1\}^{n \times K}$. Each entry $t_{ik} = 1$ if sample i is in batch k, and 0 otherwise. For the purposes of this manuscript, we focus on the two-batch case.

p-values and Multiple Hypothesis Correction p-values in this manuscript are estimated using permutation testing, which is an approach to obtain the distribution of the test statistic under the null with minimal assumptions and approximations [5]. All p-values are estimated using $N = 10{,}000$ permutations. Across all figures associated with this work, we are concerned with obtaining a proper estimate of the rate at which we detect effects (*discoveries*). Therefore, we control the false discovery rate (FDR) with the Benjamini-Hochberg Correction [1].

5.3 Site Effect Correction

To correct for site effects, we compare both z-scoring per-feature within each batch [14] and traditional `ComBat` to Causal `ComBat`.

5.4 Control Numerical Experiments

Control experiments are performed to ensure that after batch effect correction, the resulting data maintains interpretability and utility for scientific inquiry. Even if the data is devoid of batch effects, it must still be useful for downstream inference. For our connectome data, we identify two key control scenarios, the preservation of demographic effects and the preservation topological effects, which are desirable to preserve after batch effect correction.

Testing Topological Properties After applying a batch effect correction technique, it is valuable to know whether known topological properties of the connectome have been altered. Particularly, if one was to attempt to learn properties about individual-level connectomes (that is, a property that exists *in* a single brain, rather than *between* brains of different groups of people) it would be valuable to know whether that property is, or is not, altered by the correction technique. Two commonly used properties of connectomes tend to be the homotopic effect and the homophilic effect [3]. The homotopic effect describes the tendency of bilateral regions of the brain (that is, the same region, in opposing hemispheres) to be more strongly connected than other regions of the brain. The homophilic effect is the tendency of connectivity to be stronger between regions in the same hemisphere than different hemispheres. To formalize this relationship, we leverage the Structured Independent Edge Model [3]. We use the Mann-Whitney U-statistic [8] to investigate the magnitude of the homotopic and homophilic effects.

Demographic Effect Demographic effects are investigated across both the subset of connectomes upon which Causal `ComBat` is executed (the American Clique) and the full dataset. We observe the tuple $(y_i(k,l), s_i, a_i, t_i)$ for $i \in [n]$, and $k, l \in [V]$, where $V = 116$ denotes the number of parcels in the Automated Anatomical Labelling (AAL) parcellation [19]. We suppose that $Y(k,l)$ is the $[0,1]$-valued random variable denoting the weight of edge (k,l), S is the binary-valued random variable denoting the biological sex, A is the positive real-valued random variable denoting age, and T is the $[K]$-valued random variable denoting the batch. We let $\vec{y}(k,l) = (y_i(k,l)) \in [0,1]^n$ denote the realized edge weights, $\vec{s} = (s_i) \in \{0,1\}^n$ denote the realized biological sexes, $\vec{a} = (a_i) \in \mathbb{R}^n$ denote the realized biological ages, and $\vec{t} = (t_i) \in [K]^n$ denote the realized batches. We say that a **demographic sex effect** exists when:

$$f(y|a,s) \neq f(y|a,s')$$

To test for a demographic sex effect, we have the following null and alternative hypotheses:

$$H_0 : f(y|a,s) = f(y|a,s') \text{ against } f(y|a,s) \neq f(y|a,s')$$

We are able to test the preceding hypothesis using the partial distance correlation, and the test statistic is $\texttt{pdcorr}(\vec{y}(k,l), \vec{s}|\vec{a})$.

5.5 Datasets The CoRR Mega-Study [44] is an aggregate dataset consisting of 26 studies collected with a similar goal: assessing the reliability and reproducibility of neuroimaging data. The mega-study consists of $N > 1700$ individuals, most of whom are measured numerous times. We estimate resting-state functional connectivity using the AAL cortical parcellation [19]. Parcels are sorted throughout the manuscript according to hemispheric order, in which the parcels are aligned with left hemisphere parcels followed by right hemisphere parcels. Within hemisphere, parcels are sorted by AAL parcel number. All connectomes are estimated using the `m2g` (MRI to Graphs) pipeline [13]. For each protocol, we have baseline covariates for the continent, sex, and the age of participants.

The American Clique The "American Clique" describes a subset of the CoRR Mega-Study in which the sample populations share similarities in sample demographic characteristics. These studies share a demographic focusing on males and females (in roughly equal proportions) of individuals across a wide age range, and include the "NYU2", "IBATRT", "MRN1", "UWM", and "NYU1" studies. As Section 3.1 explores, adjusted effects can be estimated across the American Clique.

The NKI Rockland Sample [22] is a single study from the CoRR Mega-Study consisting of 24 individuals, each of whom is measured two times across three functional MRI acquisition protocols, which vary in the repetition time for each slice of the sequence (TR). The data was collected with the intention of investigating the impact of the different MRI protocols in a crossover-randomized approach. Due to the crossover property, evidence in favor of an effect provides strong evidence of a causal batch effect. Images with a TR of 645 millisecond (ms), 1400 ms, and 2500 ms are measured, with the prompt for each subject remaining identical.

References

1. Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," Journal of the Royal Statistical Society, 1995.
2. E. W. Bridgeford, S. Wang, Z. Yang, Z. Wang, T. Xu, C. Craddock, J. Dey, G. Kiar, W. Gray-Roncal, C. Colantuoni, C. Douville, S. Noble, C. E. Priebe, B. Caffo, M. Milham, X.-N. Zuo, , and J. T. Vogelstein, "Eliminating accidental deviations to minimize generalization error and maximize reliability: applications in connectomics and genomics," bioRxiv, p. 802629, Dec 2020. [Online]. Available: https://doi.org/10.1101/802629
3. J. Chung, E. Bridgeford, J. Arroyo, B. D. Pedigo, A. Saad-Eldin, V. Gopalakrishnan, L. Xiang, C. E. Priebe, and J. T. Vogelstein, "Statistical Connectomics," OSF Preprints, Aug 2020.
4. P. Ding and L. W. Miratrix, "To Adjust or Not to Adjust? Sensitivity Analysis of M-Bias and Butterfly-Bias," Journal of Causal Inference, vol. 3, no. 1, pp. 41–57, Mar 2015.
5. B. Efron, "Large-Scale Simultaneous Hypothesis Testing," J. Am. Stat. Assoc., vol. 99, no. 465, pp. 96–104, Mar 2004.
6. W. Forstmeier, E.-J. Wagenmakers, and T. H. Parker, "Detecting and avoiding likely false-positive findings – a practical guide," Biol. Rev., vol. 92, no. 4, pp. 1941–1968, Nov 2017.
7. M. Hampson, F. Tokoglu, X. Shen, D. Scheinost, X. Papademetris, and R. T. Constable, "Intrinsic Brain Connectivity Related to Age in Young and Middle Aged Adults," PLoS One, vol. 7, no. 9, p. e44067, Sep 2012.
8. T. P. Hettmansperger and J. W. McKean, Robust Nonparametric Statistical Methods. Andover, England, UK: Taylor & Francis, Dec 2010.
9. D. E. Ho, K. Imai, G. King, and E. A. Stuart, "MatchIt: Nonparametric preprocessing for parametric causal inference," Journal of Statistical Software, vol. 42, no. 8, pp. 1–28, 2011. [Online]. Available: https://www.jstatsoft.org/v42/i08/
10. M. Ingalhalikar, A. Smith, D. Parker, T. D. Satterthwaite, M. A. Elliott, K. Ruparel, H. Hakonarson, R. E. Gur, R. C. Gur, and R. Verma, "Sex differences in the structural connectome of the human brain," Proc. Natl. Acad. Sci. U.S.A., vol. 111, no. 2, pp. 823–828, Jan 2014.
11. W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," Biostatistics, vol. 8, no. 1, pp. 118–127, Jan 2007.
12. N. R. Karcher and D. M. Barch, "The ABCD study: understanding the development of risk for mental and physical health outcomes - Neuropsychopharmacology," Neuropsychopharmacology, vol. 46, no. 1, pp. 131–142, Jan 2021.
13. G. Kiar, E. W. Bridgeford, W. R. Gray Roncal, , V. Chandrashekhar, D. Mhembere, S. Ryman, X.-N. Zuo, D. S. Margulies, R. C. Craddock, C. E. Priebe, R. Jung, V. D. Calhoun, B. Caffo, R. Burns, M. P. Milham, and J. T. Vogelstein, "A High-Throughput Pipeline Identifies Robust Connectomes But Troublesome Variability," bioRxiv, p. 188706, Apr 2018. [Online]. Available: https://doi.org/10.1101/188706
14. C. Lazar, S. Meganck, J. Taminau, D. Steenhoff, A. Coletta, C. Molter, D. Y. Weiss-Solís, R. Duque, H. Bersini, and A. Nowé, "Batch effect removal methods for microarray gene expression data integration: a survey," Briefings Bioinf., vol. 14, no. 4, pp. 469–490, Jul 2013.
15. J. T. Leek and R. D. Peng, "Statistics: *P* values are just the tip of the iceberg," http://www.nature.com/news/statistics-p-values-are-just-the-tip-of-the-iceberg-1.17412, Apr. 2015, accessed: 2020-10-12.
16. J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry, "Tackling the widespread and critical impact of batch effects in high-throughput data," Nat. Rev. Genet., vol. 11, no. 10, pp. 733–739, Oct. 2010.
17. W. Miao, W. Hu, E. L. Ogburn, and X. Zhou, "Identifying effects of multiple treatments in the presence of unmeasured confounding," arXiv, Nov 2020. [Online]. Available: https://arxiv.org/abs/2011.04504v3

18. M. B. Misiura, J. C. Howell, J. Wu, D. Qiu, M. W. Parker, J. A. Turner, and W. T. Hu, "Race modifies default mode connectivity in Alzheimer's disease," Transl. Neurodegener., vol. 9, 2020.
19. N. N Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," Neuroimage, Jan 2002.
20. D. E. Nee, "fMRI replicability depends upon sufficient individual-level data," Commun. Biol., vol. 2, no. 130, pp. 1–4, Apr 2019.
21. S. Noble, D. Scheinost, E. S. Finn, X. Shen, X. Papademetris, S. C. McEwen, C. E. Bearden, J. Addington, B. Goodyear, K. S. Cadenhead, H. Mirzakhanian, B. A. Cornblatt, D. M. Olvet, D. H. Mathalon, T. H. McGlashan, D. O. Perkins, A. Belger, L. J. Seidman, H. Thermenos, M. T. Tsuang, T. G. M. van Erp, E. F. Walker, S. Hamann, S. W. Woods, T. D. Cannon, and R. T. Constable, "Multisite reliability of MR-based functional connectivity," Neuroimage, vol. 146, pp. 959–970, Feb 2017.
22. K. B. Nooner, S. Colcombe, R. Tobe, M. Mennes, M. Benedict, A. Moreno, L. Panek, S. Brown, S. Zavitz, Q. Li, S. Sikka, D. Gutman, S. Bangaru, R. T. Schlachter, S. Kamiel, A. Anwar, C. Hinz, M. Kaplan, A. Rachlin, S. Adelsberg, B. Cheung, R. Khanuja, C. Yan, C. Craddock, V. Calhoun, W. Courtney, M. King, D. Wood, C. Cox, C. Kelly, A. DiMartino, E. Petkova, P. Reiss, N. Duan, D. Thompsen, B. Biswal, B. Coffey, M. Hoptman, D. C. Javitt, N. Pomara, J. Sidtis, H. Koplewicz, F. X. Castellanos, B. Leventhal, and M. Milham, "The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry," Front. Neurosci., vol. 6, Oct 2012.
23. J. Pearl, "Causal inference in statistics: An overview," ssu, vol. 3, no. none, pp. 96–146, Jan 2009.
24. ——, On measurement bias in causal inference. Arlington, TX, USA: AUAI Press, Jul 2010.
25. R. Pomponio, G. Erus, M. Habes, J. Doshi, D. Srinivasan, E. Mamourian, V. Bashyam, I. M. Nasrallah, T. D. Satterthwaite, Y. Fan, L. J. Launer, C. L. Masters, P. Maruff, C. Zhuo, H. Völzke, S. C. Johnson, J. Fripp, N. Koutsouleris, D. H. Wolf, R. Gur, R. Gur, J. Morris, M. S. Albert, H. J. Grabe, S. M. Resnick, R. N. Bryan, D. A. Wolk, R. T. Shinohara, H. Shou, and C. Davatzikos, "Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan," Neuroimage, vol. 208, p. 116450, Mar 2020.
26. M. Powell, A. Koenecke, J. B. Byrd, A. Nishimura, M. F. Konig, R. Xiong, S. Mahmood, V. Mucaj, C. Bettegowda, L. Rose, S. Tamang, A. Sacarny, B. Caffo, S. Athey, E. A. Stuart, and J. T. Vogelstein, "A how-to guide for conducting retrospective analyses: example COVID-19 study," OSF Preprints, Sep 2020.
27. J. Robins, "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect," Mathematical Modelling, vol. 7, no. 9, pp. 1393–1512, Jan 1986.
28. P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," Biometrika, vol. 70, no. 1, pp. 41–55, Apr 1983.
29. ——, "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score," The American Statistician, vol. 39, no. 1, pp. 33–38, 1985. [Online]. Available: http://www.jstor.org/stable/2683903
30. R. Sala-Llonch, D. Bartrés-Faz, and C. Junqué, "Reorganization of brain networks in aging: a review of functional connectivity studies," Front. Psychol., vol. 0, 2015.
31. T. D. Satterthwaite, D. H. Wolf, D. R. Roalf, K. Ruparel, G. Erus, S. Vandekar, E. D. Gennatas, M. A. Elliott, A. Smith, H. Hakonarson, R. Verma, C. Davatzikos, R. E. Gur, and R. C. Gur, "Linked Sex Differences in Cognition and Functional Connectivity in Youth," Cerebral Cortex (New York, NY), vol. 25, no. 9, p. 2383, Sep 2015.
32. C. Shen, C. E. Priebe, and J. T. Vogelstein, "From Distance Correlation to Multiscale Generalized Correlation," Journal of American Statistical Association, Oct. 2017. [Online]. Available: http://arxiv.org/abs/1710.09768

33. E. A. Stuart, “Matching Methods for Causal Inference: A Review and a Look Forward,” Statist. Sci., vol. 25, no. 1, pp. 1–21, Feb 2010.
34. G. J. Székely and M. L. Rizzo, “Partial distance correlation with methods for dissimilarities,” Ann. Stat., vol. 42, no. 6, pp. 2382–2412, Dec 2014.
35. G. J. Székely, M. L. Rizzo, and N. K. Bakirov, “Measuring and testing dependence by correlation of distances,” Ann. Stat., vol. 35, no. 6, pp. 2769–2794, Dec 2007.
36. B. O. Turner, E. J. Paul, M. B. Miller, and A. K. Barbey, “Small sample sizes reduce the replicability of task-based fMRI studies,” Commun. Biol., vol. 1, no. 62, pp. 1–10, Jun 2018.
37. E. Varangis, C. G. Habeck, Q. R. Razlighi, and Y. Stern, “The Effect of Aging on Resting State Connectivity of Predefined Networks in the Brain,” Front. Aging Neurosci., vol. 0, 2019.
38. J. T. Vogelstein, E. W. Bridgeford, Q. Wang, C. E. Priebe, M. Maggioni, and C. Shen, “Discovering and deciphering relationships across disparate data modalities,” Elife, vol. 8, Jan. 2019. [Online]. Available: http://dx.doi.org/10.7554/eLife.41690
39. C. Wachinger, A. Rieckmann, and S. Pölsterl, “Detect and Correct Bias in Multi-Site Neuroimaging Datasets,” arXiv, Feb 2020.
40. S. Weis, K. R. Patil, F. Hoffstaedter, A. Nostro, B. T. T. Yeo, and S. B. Eickhoff, “Sex Classification by Resting State Brain Connectivity,” Cereb. Cortex, vol. 30, no. 2, pp. 824–835, Mar 2020.
41. A. Yamashita, N. Yahata, T. Itahashi, G. Lisi, T. Yamada, N. Ichikawa, M. Takamura, Y. Yoshihara, A. Kunimatsu, N. Okada, H. Yamagata, K. Matsuo, R. Hashimoto, G. Okada, Y. Sakai, J. Morimoto, J. Narumoto, Y. Shimada, K. Kasai, N. Kato, H. Takahashi, Y. Okamoto, S. C. Tanaka, M. Kawato, O. Yamashita, and H. Imamizu, “Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias,” PLoS Biol., vol. 17, no. 4, p. e3000042, Apr 2019.
42. M. Yu, K. A. Linn, P. A. Cook, M. L. Phillips, M. McInnis, M. Fava, M. H. Trivedi, M. M. Weissman, R. T. Shinohara, and Y. I. Sheline, “Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data,” Hum. Brain Mapp., vol. 39, no. 11, pp. 4213–4227, Nov 2018.
43. Y. Zhang, G. Parmigiani, and W. E. Johnson, “ComBat-seq: batch effect adjustment for RNA-seq count data,” NAR Genomics Bioinf., vol. 2, no. 3, Sep 2020.
44. X.-N. Zuo, J. S. Anderson, P. Bellec, R. M. Birn, B. B. Biswal, J. Blautzik, J. C. S. Breitner, R. L. Buckner, V. D. Calhoun, F. X. Castellanos, A. Chen, B. Chen, J. Chen, X. Chen, S. J. Colcombe, W. Courtney, R. C. Craddock, A. Di Martino, H.-M. Dong, X. Fu, Q. Gong, K. J. Gorgolewski, Y. Han, Y. He, Y. He, E. Ho, A. Holmes, X.-H. Hou, J. Huckins, T. Jiang, Y. Jiang, W. Kelley, C. Kelly, M. King, S. M. LaConte, J. E. Lainhart, X. Lei, H.-J. Li, K. Li, K. Li, Q. Lin, D. Liu, J. Liu, X. Liu, Y. Liu, G. Lu, J. Lu, B. Luna, J. Luo, D. Lurie, Y. Mao, D. S. Margulies, A. R. Mayer, T. Meindl, M. E. Meyerand, W. Nan, J. A. Nielsen, D. O’Connor, D. Paulsen, V. Prabhakaran, Z. Qi, J. Qiu, C. Shao, Z. Shehzad, W. Tang, A. Villringer, H. Wang, K. Wang, D. Wei, G.-X. Wei, X.-C. Weng, X. Wu, T. Xu, N. Yang, Z. Yang, Y.-F. Zang, L. Zhang, Q. Zhang, Z. Zhang, Z. Zhang, K. Zhao, Z. Zhen, Y. Zhou, X.-T. Zhu, and M. P. Milham, “An open science resource for establishing reliability and reproducibility in functional connectomics,” Sci. Data, vol. 1, no. 140049, pp. 1–13, Dec 2014.

Appendix A. Correction for Batch-Related Effects.

Algorithm 1 Causal ComBat. The Causal ComBat approach for batch effect correction. Conditional ComBat is performed on demographically matched batches to eliminate confounding biases in observational studies. The smallest batch is arbitrarily chosen as the treated batch, and then successive batches are matched against the treated batch. Treated individuals (as well as their matched controls) with no suitable matches in one (or more) batches are eliminated from the analysis. $\texttt{CovariateMatch}()$ is a function which takes the treatment and control covariates and matches control individuals to treatment individuals. $\mathcal{I}^k_{mat}$ is a list of the treated individuals with at least one match, and $\mathcal{M}^k_{mat}$ is a list of the treated individual that each control individual was (potentially) matched to. $\texttt{BalanceCovariates}()$ uses these lists to identify the treatment individuals who have matches across all control batches and then excludes control individuals who are not matched to treatment individuals with matches across all batches.

Input: The tuples $(\mathbf{Y}^{(k)}, \mathbf{X}^{(k)})$ for $k = 1, \ldots, K$, where $\mathbf{Y}^{(k)} \in \mathbb{R}^{n_k \times p}$ denotes the measurements from batch id k, and $\mathbf{X}^{(k)} \in \mathbb{R}^{n_k \times d}$ denotes the covariates from batch id k, where $N = \sum_k n_k$ is the total number of measurements.

Output: The $N' \times p$ batch effect corrected data, where $N' \leq N$.

function CAUSALCOMBAT($\mathbf{Y}, \mathbf{X}, \vec{t}$)
  $\tilde{\mathbf{Y}}, \tilde{\mathbf{X}}, \tilde{n} = \texttt{BalanceCovariates}(\mathbf{Y}, \mathbf{X})$
  Let $\vec{t} = (1_{\tilde{n}_1}, 2_{\tilde{n}_2}, \ldots, K_{\tilde{n}_K})$ ▷ vector indicating the batch id of each individual
  $\tilde{\mathbf{Y}}_n = \texttt{ComBat}(\tilde{\mathbf{Y}}, \tilde{\mathbf{X}}, \vec{t})$ ▷ Remove the effect due to the batch $\vec{t}$, conditional on the covariates $\mathbf{X}$
  return $\tilde{\mathbf{Y}}_n$ ▷ Return the batch
end function
function BALANCECOVARIATES($\mathbf{Y}, \mathbf{X}$)
  $k' = \text{argmin}_k \, n_k$ ▷ Select treatment study to be the smallest study
  for $k \neq k'$ **do**
    $\mathcal{M}^k_{mat}, \mathcal{I}^k_{mat} = \texttt{CovariateMatch}(\mathbf{X}^{(k')}, \mathbf{X}^{(k)})$ ▷ Covariate match the two studies
  end for
  $\mathcal{I}_{mat} = \bigcap_{k \neq k'} \mathcal{I}^k_{mat}$ ▷ Identify treatment individuals with a control match across all control groups
  $\mathbf{Y}^{(k')} = \mathbf{Y}^{(k')}[\mathcal{I}_{mat}, :]$ ▷ Remove other treatment individuals
  $\mathbf{X}^{(k')} = \mathbf{X}^{(k')}[\mathcal{I}_{mat}, :]$
  $n_{k'} = |\mathcal{I}_{mat}|$ ▷ Number of treated individuals remaining
  for $k \neq k$ **do**
    $\mathcal{J}^k_{mat} = \texttt{which}(\mathcal{M}^k_{mat} \, \%in\% \, \mathcal{I}_{mat})$
      ▷ Control individuals matched to treatment individuals with matches across all groups
    $\mathbf{Y}^{(k)} = \mathbf{Y}^{(k)}[\mathcal{J}^k_{mat}, :]$ ▷ Remove other control individuals
    $\mathbf{X}^{(k)} = \mathbf{X}^{(k)}[\mathcal{J}^k_{mat}, :]$
    $n_k = \left|\mathcal{J}^k_{mat}\right|$ ▷ Number of control individuals remaining
  end for
  $\mathbf{Y} = [\mathbf{Y}^{(1)}; \ldots; \mathbf{Y}^{(K)}]$ ▷ raw measurements across batches
  $\mathbf{X} = [\mathbf{X}^{(1)}; \ldots; \mathbf{X}^{(K)}]$ ▷ covariates across batches
  return $\mathbf{Y}, \mathbf{X}, n$
end function

In Figure 6, we examine the prevalence of effects between different studies. It is important to note that whereas Causal ComBat can only be run on the American Clique, both ComBat and Z-scoring can be run on all studies. Recall that associational effects investigate whether *any* batch effects are present between pairs of studies (either measurement- or demographics-related). Both ComBat and Z-scoring eliminating nearly *all* variability between pairs of studies suggests that ComBat and Z-scoring

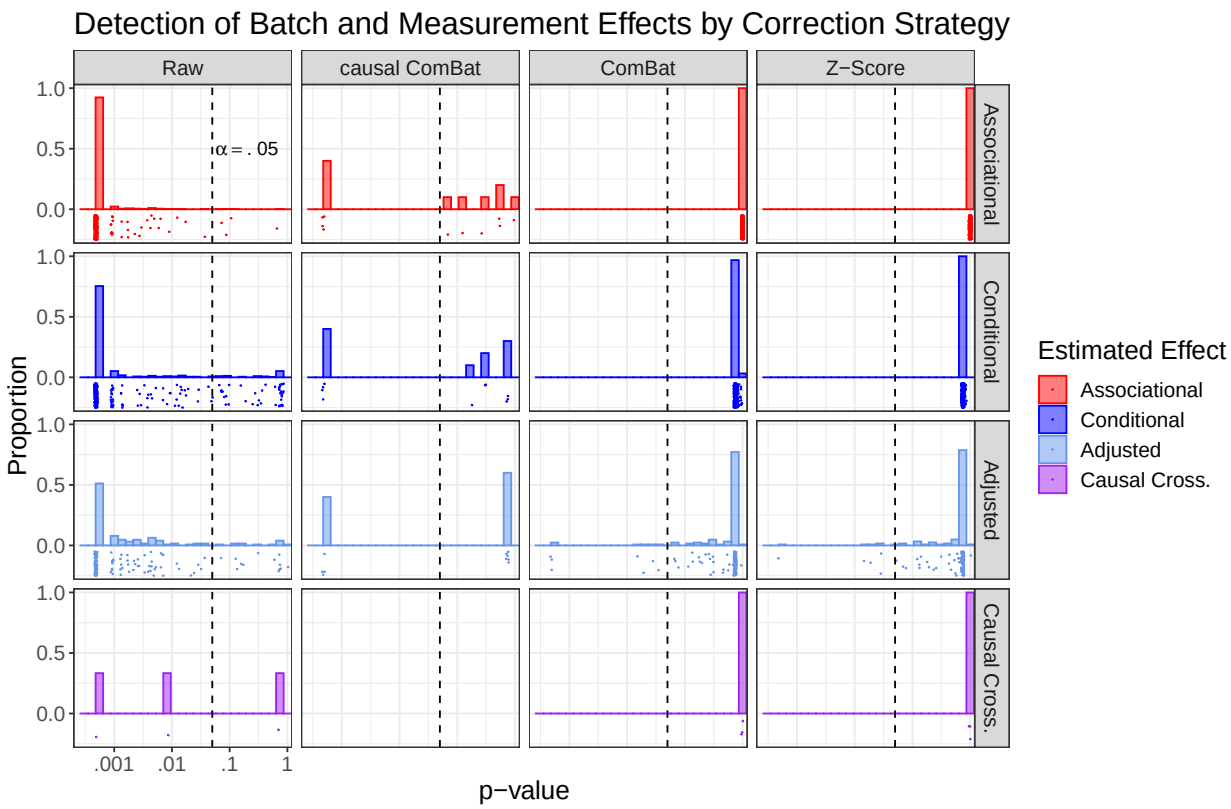

Figure 6: Empirical distribution of p-values for each estimated effect type by batch correction strategy.

are eliminating variability that is directly related to natural biological variability, which is not a favorable outcome. For both strategies, associational effects are not significant for 100% of comparisons. This is because the studies differ on the basis of demographic covariates possessing signal, and therefore observing no associational batch effects after correction implies we have removed demographic signal. On the other hand, limiting oneself to only pairs of studies in which adjusted effects are estimable with Causal `ComBat` may offer a solution to this undesirable property of Z-scoring and naive `ComBat`.

Appendix B. Overlap of Empirical Covariates.

The empirical overlap of the covariate distributions is computed by first approximating the density of the distribution of the measured covariates for each dataset D, $X = (A, S, C)$, where A is a random variable whose realizations $a \in \mathcal{A}$ are ages, S is a random variable whose realizations $s \in \mathcal{S}$ are sexes (M or F), and C is a random variable whose realizations $c \in \mathcal{C}$ denote continent, using the base R function `stats::density`. The random variable D has realizations $d \in \mathcal{D}$ whose realizations denote dataset. The density $f(a|S = s, C = c, D = d)$ is the conditional density of age, conditional on the individual's sex being s, continent of measurement being c, and dataset being d. The mass $\mathbb{P}(S = s|C = c, D = d)$ is the conditional mass of sex, conditional on the individual's continent of measurement being c, and dataset being d. Finally, the mass $\mathbb{P}(C = c|D = d)$ represents the conditional mass of an individual's continent of measurement being c, conditional on the dataset being d (0 or 1 for all d, since all individuals from dataset d are either measured on continent c or not). An estimate of the overlap between the two densities, $\hat{v}_{kl}$ between datasets k and l, is computed using the formula:

$$\hat{v}_{kl} = \sum_{c \in \mathcal{C}} \min\Big(\hat{\mathbb{P}}(C = c|D = k)\hat{\mathbb{P}}(C = c|D = l)\Big) \sum_{s \in \{\text{M,F}\}} \min\Big(\hat{\mathbb{P}}(S = s|C = c, D = k), \hat{\mathbb{P}}(S = s|C = c, D = l)\Big) \Big[$$

$$\int_{\mathcal{A}} \min\Big(\hat{f}(a|S=s, C=c, D=k), \hat{f}(a|S=s, C=c, D=l)\Big)\, \mathrm{d}a\Big]$$

Intuitively, this can be conceptualized as representing the mass of the "area under the curve" which is shared by the two densities for datasets k and l.

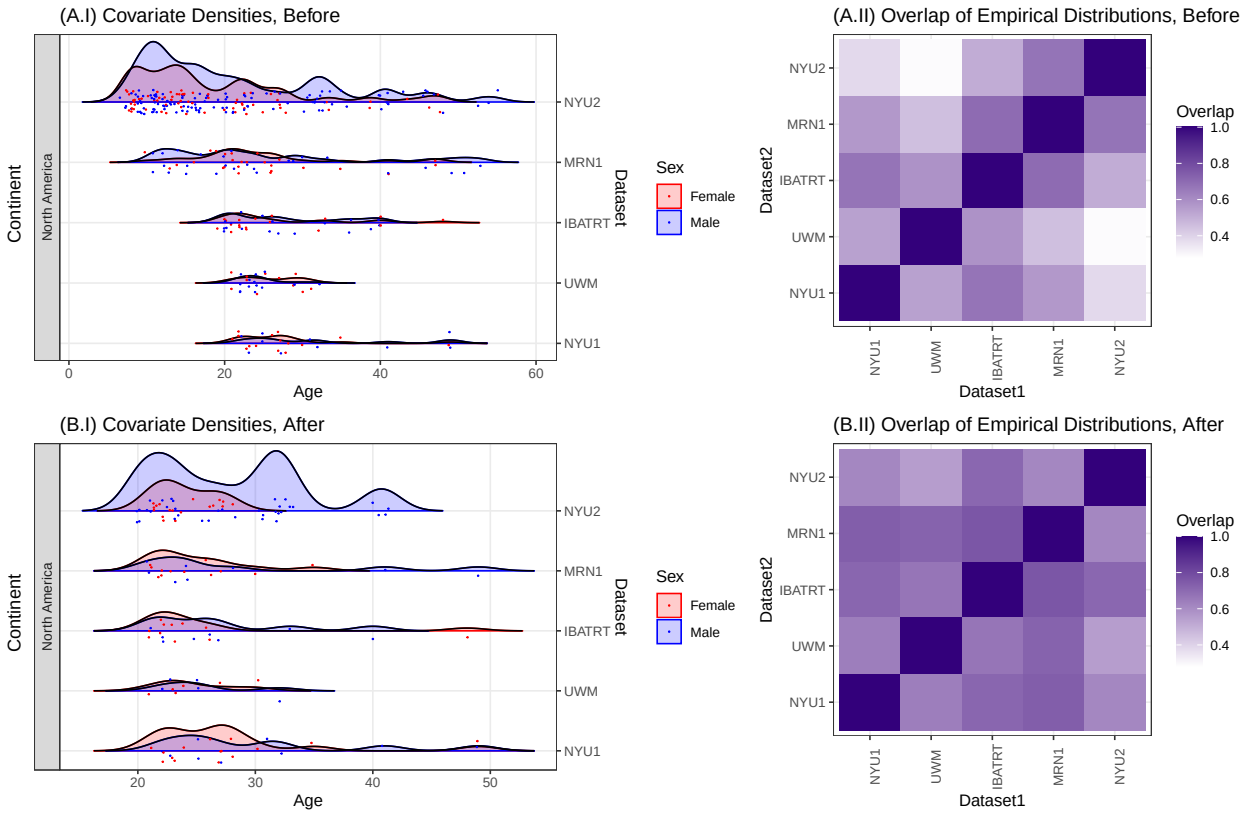

Figure 7: **The overlap of the empirical covariate distributions for the American Clique, before and after adjustment**. **(A)** The empirical distributions of the covariates before **(A.I)** and after **(A.II)** adjustment. **(B)** The empirical overlap of the covariate distributions before **(B.I)** and after **(B.II)** the adjustment procedure is applied, as discussed in Appendix A.

Appendix C. `ComBat` Preserves Topological Properties of fMRI Connectomes. The presence of two prominent topological *signal effects* in human connectomes, homophily and homotopy, are investigated in Figure 8. The left column of Figure 8A looks at the edges comprising these two disparate community structures. When edges are sorted by hemisphere, the homophilic effect is characterized by a modular two-block structure, with higher connectivity in the lower left and upper right blocks (dark red) than in the two off-diagonal blocks (pink). The homotopic effect is characterized by edges in the off-diagonal band (dark blue) showing higher connectivity than other edges within the connectome (light blue). The right column of Figure 8A looks at a single individual's connectome from the NYU2 study before and after correction. The homotopic property appears to be preserved in only the `ComBat` connectomes, whereas Z-scored connectomes do not appear to preserve this property. In Figure 8B, these two properties are examined statistically for the individuals from the American Clique (Mann-Whitney U Statistic, see Methods 5.4 for details). Points falling along the diagonal line $y = x$ tend to have a similar signal effect before and after batch correction. Whereas `ComBat` strategies tend to preserve the homophilic and homotopic effects, Z-scoring tends to decrease these two topological properties of functional connectomes. This is noted by the fact that the majority of the points for Z-scoring across both effects tend to fall below the $y = x$ line, indicating that the effect size after Z-scoring is lower than the effect size in the raw connectomes.

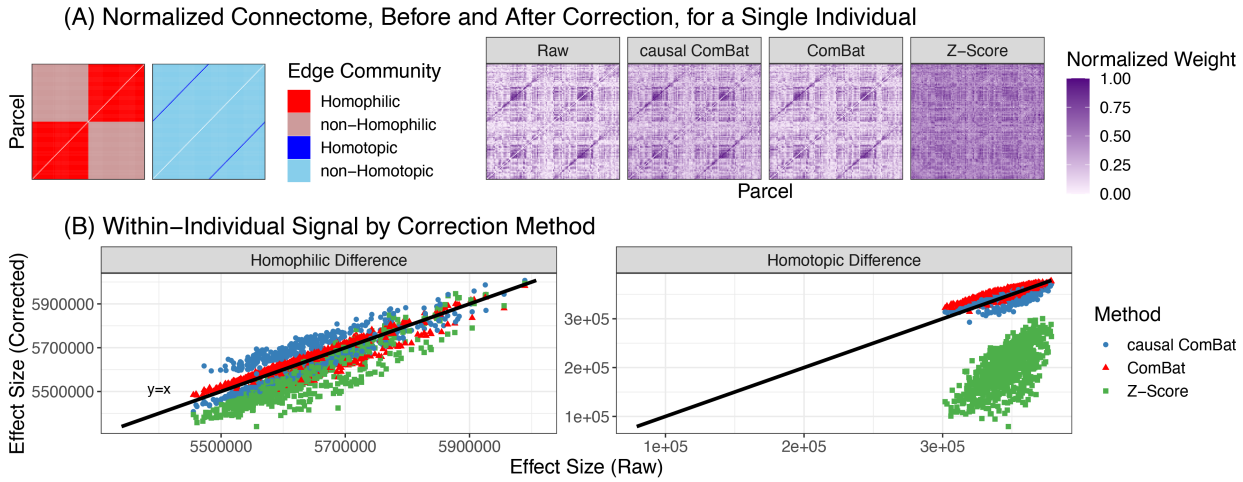

Figure 8: **Preservation of Topological Properties of Connectomes after Batch Correction**. The presence of topological effects is explored in the context of the homophilic and homotopic differences in connectivity in the American Clique. **(A)** (*Left*) A map indicating which edges are associated with homophilic/non-homophilic or homotopic/non-homotopic edge communities. (*Right*) The edge-weight normalized connectome associated with a single individual from the NYU2 study using each removal strategy. **(B)** A scatter plot, showing the effect size (Mann-Whitney U Statistic) before (x axis) and after (y axis) the indicated Batch Effect Correction Method.

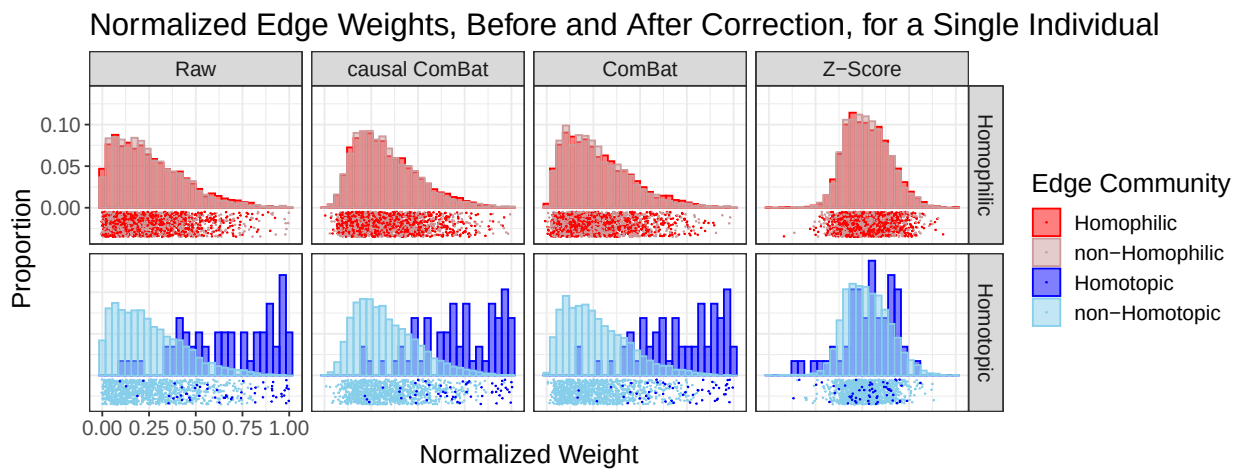

Figure 9: Histogram and rug plot indicating the empirical distribution of the normalized edge weights associated with each of the four edge communities for a single individual's connectome.

Figure 9 explores the distribution of edge weights within these two community structures before and after correction. The conclusions are similar in that Causal `ComBat` and `ComBat` preserve the property that the homotopic edges have far higher connectivity than non-homotopic edges.