Batch Effects are Causal Effects: Applications in Human Connectomics

Eric W. Bridgeford[1,†], Michael Powell[1], Gregory Kiar[2], Stephanie Noble[3], Jaewon Chung[1], Sambit Panda[1], Ross Lawrence[1], Ting Xu[2], Michael Milham[2], Brian Caffo[1], 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 them as associational effects, rather than as causal effects, despite the fact that the sources of variance that comprise the batch—potentially including experimental design and population demographics—causally impact downstream inferences. Our key insight is that batch effects can be modeled as causal, rather than associational effects. We develop a simple strategy for augmenting existing techniques for batch effect correction to demonstrate that modeling batches as causal, rather than associational, effects leads to disparate downstream conclusions across a range of simulated and real data experiments. Our work therefore introduces a conceptual framing for thinking about how and whether to combine data across batches, and the potential limitations of doing so.

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 a source of undesirable variability known as a batch effect. Several strategies have been proposed to define a batch effect. Lazar et al. [22] 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.

Broadly, many approaches model the batch collection or measurement process as a nuisance variable which is to be removed [19, 23, 24, 34, 35, 52, 57], which implies batch effect is a particular component of an associational or conditional model. These approaches typically make strong, potentially unjustified, and often inappropriate parametric assumptions about the data. Two of the most prominent examples of these techniques which can be generalized to many of these other cases are the `ComBat` and Conditional `ComBat` batch effect correction procedures [19]. These approaches have demonstrated empirical utility in certain genomics and neuroimaging contexts [35, 58]; 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 [38, 39, 44]. Methodologically, we introduce a simple strategy appended to existing techniques for associational or conditional batch effect detection and removal to reflect this new perspective.

[1] Johns Hopkins University [2] Child Mind Institute, [3] Yale University, [†] Corresponding author: Eric W. Bridgeford (ebridge2@jhu.edu).

We demonstrated the utility of this perspective by developing *Causal* DCORR [46]—building on modern, non-parametric statistical methods—to estimate and detect the presence of batch effects. Second, we utilized this strategy to develop *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 simulations and a large neuroimaging mega-study assembled by the Consortium for Reliability and Reproducibility (CoRR) [59], consisting of more than 1700 connectome measurements across 24 disparate studies. Our investigations reveal that previous strategies fail to differentiate between veridical signal and sources of variability due to other experimental design factors. Further, our simulations demonstrate that existing strategies can, under many realistic use-cases, remove more apparent 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 Appendix B.

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 independently from some true but unknown distribution (which implies *exchangeability*). 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 [26, 33, 34, 37]:

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

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 backdoor formula (1).

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

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

Colloquially, if we seek to detect a batch effect, we seek to be able to characterize when the relationship described above is true. The key limitation of this definition of a batch effect is that, as elucidated in Equation (1), the estimands of interest $f_{x,z}(y|t)$ require an integration over the *unconditional* distribution of both the known *and* unknown covariates, $f(x, z)$, which we will not have in practice due to the fact that the covariates Z are unknown. Therefore, if one were to construct a hypothesis test on the basis of Definition 1, we would be unable to characterize principled estimators of the quantities $f_{x,z}(y|t)$ directly. To this end, we define a sequence of related effects below, and we elucidate the specific sets of progressively less restrictive assumptions about (Y, T, X, Z) under which the existence of the effect of interest implies the existence of a batch effect. 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:*

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

Practically, techniques which aim to address these associational site effects are known as associational site effect correction techniques.

Definition 3 (Associational Site Effect Correction). *Associational site effect correction is the process of estimating and applying a function $\hat{g}$ which is learned from the data (y_i, t_i), where:*

$$v_i = \hat{g}(y_i, t_i).$$

The associational site effect is removed between batches t and t' if $f(v|t) = f(v|t')$. In practice, associational techniques involve approaches for associational site effect detection and correction which *ignore* measured covariates, such as `DCORR` for batch effect detection and `ComBat` for batch effect removal.

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

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

Techniques which aim to address these conditional site effects are known as conditional site effect correction techniques.

Definition 5 (Conditional Site Effect Correction). *Conditional site effect correction is the process of estimating and applying a function $\hat{g}$ which is learned from the data (y_i, x_i, t_i), where:*

$$v_i = \hat{g}(y_i, x_i, t_i).$$

The conditional site effect is removed between batches t and t' if $f(v|t,x) = f(v|t',x)$. Conditional techniques involve approaches for batch effect detection and correction that account for measured covariates, such as Conditional DCORR for batch effect detection and Conditional ComBat for batch effect removal.

When would a conditional site effect be equivalent to a causal batch effect? To answer this question, we first define the propensity score $e(t|x,z)$ [38, 39]; 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: $T \perp\!\!\!\perp Y | X, Z$ (ignorability).
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 [38]. 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 6 (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.

Techniques which aim to address adjusted site effects are known as adjusted site effect correction techniques.

Definition 7 (Adjusted Site Effect Correction). *Reweight the observations such that $\tilde{f}(x|t) \approx \tilde{f}(x|t')$ via weights w_i for each observation i.*

Adjusted site effect correction is the process of estimating and applying a function $\hat{g}$ which is learned from the reweighted data (y_i, x_i, w_i, t_i), where:

$$v_i = \hat{g}(y_i, x_i, w_i, t_i).$$

The adjusted site effect is removed between batches t and t' if $\tilde{f}(v|t, x) = \tilde{f}(v|t', x)$. In practice, *adjusted* procedures amount to augmenting conditional procedures by first reweighting the measurements across batches on the basis of known confounders. This can be done, for instance, via propensity trimming and matching [44] subjects on the basis of observed covariates. To this end, we develop Causal DCORR to estimate adjusted effects and Causal ComBat to remove adjusted effects, both of which are discussed below.

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.

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 8 (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? Given that 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). Detection and correction for crossover effects amount to applying techniques for site effect detection or correction to datasets where each subject is measured in *every* batch.

2.2 Detecting and mitigating causal batch effects

At present, literature approaches for detecting and mitigating batch effects tend to focus attention heavily on effects which could be classified as associational or conditional, which for the above listed reasons, fail to adequately account for confounding biases. To this end, we propose a simple technique to augment associational and conditional strategies and derive stronger causal claims. We propose the following strategy:

1. Use classical causal procedures to re-weight the measurements from the batches so that the demographic distributions across all batches to be considered are approximately equal. For the purposes of elucidating the benefits of this perspective, we choose the simple strategy of propensity trimming the datasets to remove outliers, followed by nearest-neighbor matching to the dataset with the fewest individuals [36, 44].
2. Apply classical procedures for detection or removal of batch effects. For the purposes of elucidating the benefits of this perspective, we choose to apply our techniques to conditional DCORR for batch effect detection [45, 46], and conditional ComBat for batch effect removal [19, 23, 24, 34, 35, 52, 57].

For a more technical description of the causal procedures employed in this manuscript, see Appendix A. It is important to note that we do *not* make the claim that the specific choices for re-weighting measurements (propensity trimming followed by nearest-neighbor matching) are necessarily the best

possible approach to balancing the demographic distributions across the batches. Indeed, there are almost two decades of research dedicated towards making these design choices for observational studies. Rather, we chose simple, principled, and straightforward approaches for illustrative purposes. Our simulations elucidate that even these possibly simplistic balancing approaches are sufficient to highlight substantial shortcomings in the existing literature for batch effects and the substantial benefits of a causal perspective.

2.3 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 to other batches, and younger children may have distinct patterns of brain connectivity compared to 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 Traditional approaches fail to adequately remove batch effects when covariate overlap is poor We constructed simulations to help us understand the differences between associational or conditional reasoning and causal reasoning. Figure 2(A).I. shows a simulation where the outcome of interest (e.g., a feature of a connectome, y-axis) is associated with a covariate of interest (x-axis) through a non-linear relationship (the sigmoid). If samples are from the blue batch, they are offset 1-unit higher than samples from the orange batch. $N = 400$ samples are drawn with equal probability from each batch. In the high overlap regime, the orange and blue batches have the same covariate distribution. Figure 2(B).I. shows the expected signal (solid line) for each batch, along with the difference between the expected signal for each batch (gray box). These curves are shown along with the group-specific linear model fit (dashed lines) estimated from the sample points, which is the model that is later used by conditional ComBat for batch effect adjustment. Figure 2(C).I. and Figure 2(D).I. show the impact

that conditional ComBat and causal ComBat have on the expected signal at a given covariate level. Intuitively, one would anticipate that a batch effect correction strategy should place the expected signals virtually overlapping after correction, since the batch effect is characterized by the expected signals in each group differing by an offset factor. For both strategies, the expected signal after correction is virtually the same for each of the two batches, indicating that the batch effect has been successfully removed.

Figure 2(A).II. shows similar plots; however, the covariate distributions have been shifted such that they no longer overlap perfectly. The orange batch has a left-shifted covariate distribution, and the blue batch has a right-shifted covariate distribution (the covariate is *confounding* the relationship between the outcome and the batch). Figure 2(B).II. shows that while the expected signal in each batch at a given covariate level is the same as in Figure 2(B).II., the linear model fits for each group estimated from the sample now differ (the lines are rotated slightly in a clockwise direction, and are now offset further apart). Figure 2(C).II. shows that after correction, the expected signal at each covariate level do not overlap; in fact, they differ more after correction than before in magnitude (the height of the residual signal disparity, red box, is about 50% larger than the actual batch effect, grey box). This is because the model being employed by conditional ComBat is biased under the given true data distribution. Figure 2(D).II. shows that, unlike conditional ComBat, causal ComBat restricts inference to the region of shared covariate overlap between the two batches, and does not perform inference that is unsupported by the data. This is indicated by the fact that the expected signal after correction has been restricted to no longer include the left-most and right-most portions of the covariates. Within this region, the expected signal for each batch after correction are virtually overlapping, indicating that the batch effect has been removed by causal ComBat.

Figure 2(A).III. has furthered this trend such that there are almost no samples in a region of shared covariate overlap between the two batches. In Figure 2(B).III. the expected signal in each batch at a given covariate level is the same as in Figure 2(B).I., but the per-group linear model fits are now even further offset. This has the impact that in Figure 2(C).III., the expected signal after correction are almost four times more offset than they were before any batch effect correction was applied. On the other hand, 2(D).III. shows that causal ComBat would indicate to not perform inference at all on the batches due to the fact that there are no portions of the covariate range in which the two batches have samples which could be compared.

In Appendix D.1 and Appendix D.2, we explore further implications of using causal techniques for batch effect detection and removal respectively. Conceptually, researchers may feel as though they are left with a tradeoff:

1. When datasets do not have high overlap demographically, the (implied or explicit) limitations of a model leveraged by a batch effect detection or correction technique can impart substantial bias on the conclusions of the investigation. While the limitations or shortcomings may *seem* inconsequential, the scale of the residual signal disparities imparted can dwarf the original batch effect, and yield measurements which are *more* dissimilar than prior to batch effect adjustment, as we saw in Figure 2(C).II. and (C).III.
2. Enforcing demographic overlap mechanistically through propensity trimming and/or matching may *seem* like it yields fewer samples for analysis, or narrows the scope of conclusions.

In practice, the reality is that with biological data, there are *no* ground truths. As a result, there is no way to evaluate the introduction of imparted residual bias, and no way to know whether downstream conclusions are a product of imparted residual biases. This means that any statistical inference we derive may be arbitrarily true or false, and we have no way of qualifying in which pool our inference stands.

When we mechanistically re-weight the data to impose demographic overlap, proper statistical inference tempers conclusions as a direct consequence of the smaller demographic region under consideration. Stated another way, this means that when we balance *prior to* detecting or correcting for

batch effects, we have tools which allow us to make internally valid conclusions supported by the data in regions in which the data *are able*. This means that, although our conclusions may not apply to the *entire* region of covariate support in the observed data (such as, in Figure 2(D).II., where we were only able to perform analysis in the region of shared support), they will be *statistically and methodologically valid* with respect to the re-weighted population that we *did* analyze.

3.2 The CoRR Studies have disparate demographic characteristics Figure 3A 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 [16, 41, 53], age [14, 40, 50], and continent [27] are variables that have been associated with brain connectivity, they serve as conditional covariates used in our investigation.

Figure 3B investigates the level of demographic overlap in the CoRR mega-study, using the distribution-free overlapping index [32], as discussed in Appendix E. Many of the datasets do not overlap at all, and inference on the presence or absence of batch effects would be entirely model-based, as-per Figure 2.III. Further, while some of the datasets overlap, the level of balance varies widely. In the case where demographic imbalances are substantial, the detection or removal of batch effects would be extremely subject to model-based assumptions, as-per Figure 2II. From a causal perspective, this is troublesome, as the covariate records for the CoRR mega-study only include three covariates: age, sex, and continent of measurement. The addition of successive covariates can only reduce the estimated overlap between the pairs of datasets, so having poor overlap on such a coarse set of covariates indicates that the actual demographic overlap may be even lower. This would be problematic if, for instance, unmeasured covariates (such as participant mental health) are confounders.

3.3 Detecting Batch Effects in the CoRR mega-study Figure 4A 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 4B. Many pairs of studies could not have an adjusted effect estimated due to poor demographic alignment as shown in Figure 3B, which resulted in no suitable matches existing between the studies, indicated by a red X (279 of 406 pairs of studies). Colloquially, when covariate overlap is moderate or limited as in Figure 2, we cannot tell if effects are misrepresented (or, comparisons are unreasonable entirely) due to demographic imbalance. In this way, causal strategies prepended to batch effect detection and mitigation strategies are *more conservative*, in that they prevent the experimenter from potentially deriving spurious or erroneous conclusions due to confounding biases. 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.

3.4 Traditional Approaches Produce Disparate Inference from Techniques which Leverage Causality At this point, we have demonstrated, through simulation and real examples, if a covariate is a *confounder*, failure to properly account for demographic balancing can, potentially, yield improper detection

and mitigation of batch effects and may, in turn, fundamentally alter true signals in the dataset. We saw that, for many of the CoRR [59] datasets, demographic balance was extremely poorly aligned, which allows only a fraction of the original possible comparisons between datasets to be analyzed in a theoretically, and empirically, reasonable manner. The final question we have is: does it matter? Even if we disrupt true signal in the dataset, can we still detect it irrespective of whether the signal is maintained in its unadulterated form?

To this end, 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). We also compared these techniques to NeuroComBat (Conditional and Causal NeuroH) [11, 35], an implementation of `ComBat` for neuroimaging data. 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, dark purple). Causal approaches produce vastly different statistical inference compared to other correction techniques on the American Clique. Causal approaches were 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`, Conditional NeuroH, 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. On the other hand, Causal NeuroH has a high overlap with Causal `ComBat`.

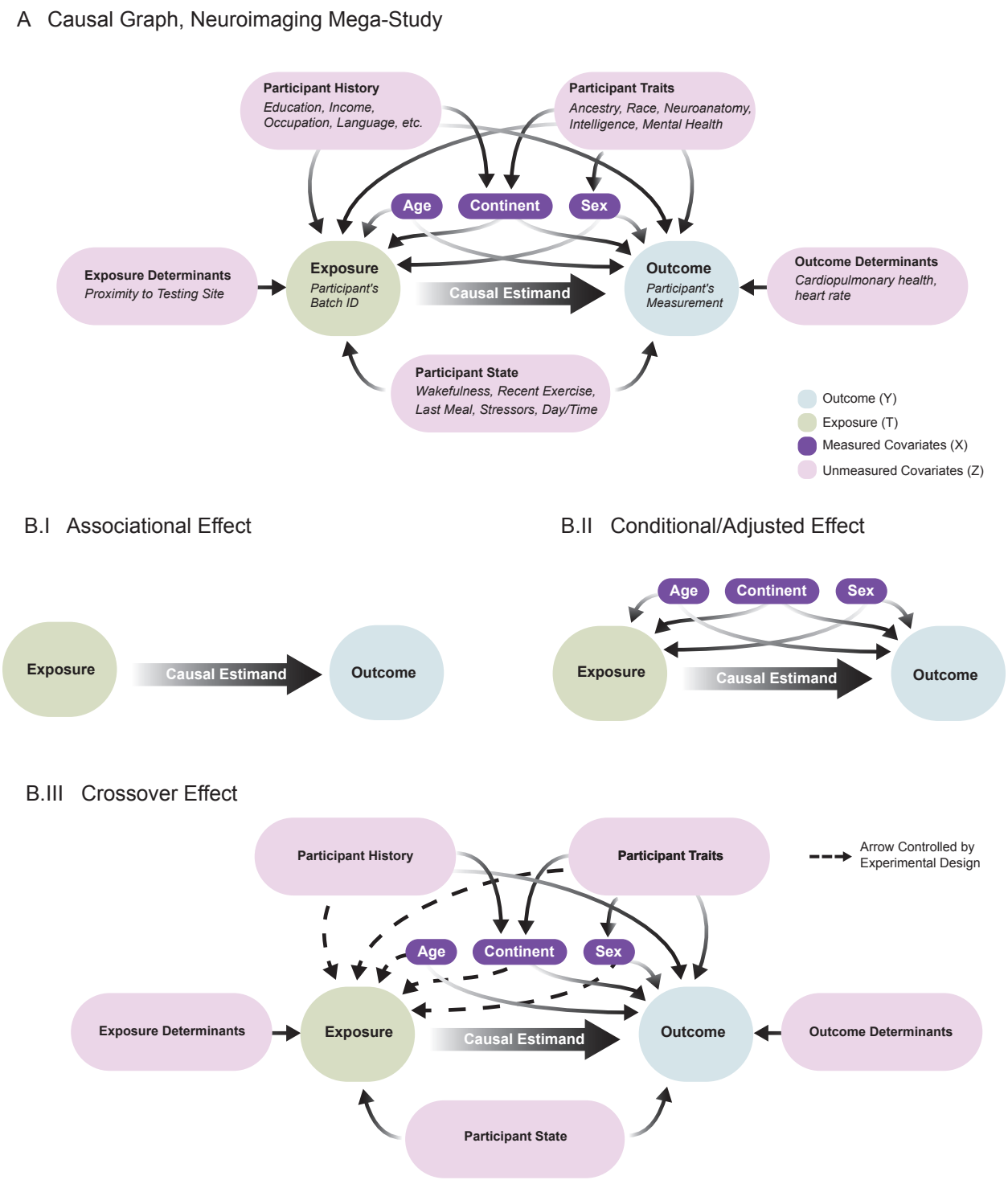

Figure 1: **Causal Graph of Study Covariates**. **(A)** A causal directed acyclic graph (DAG) illustrating potential relationships which can reasonably be assumed to underlie a mega-study. Descriptions reflect possible attributes that might fall into the specific exposures, outcomes, and covariates for a neuroimaging study, for illustration purposes. **(B)** DAGs which illustrate the assumptions which underlie 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.

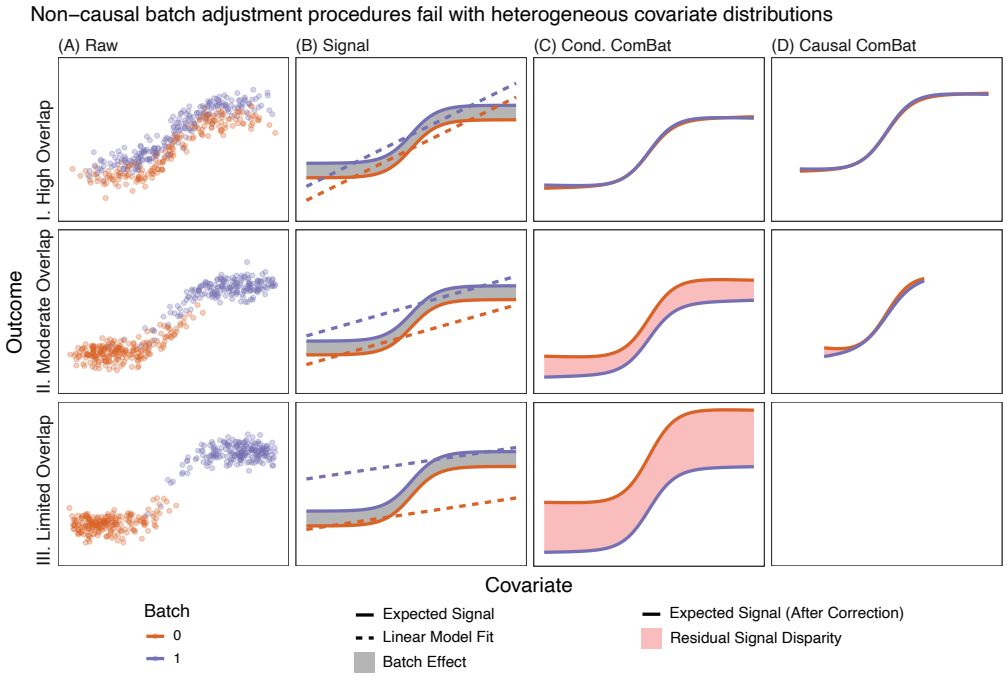

Figure 2: **Non-causal procedures are subject to strong biases without high covariate overlap**. $N = 400$ points are simulated from either batch 0 or batch 1 in (A). In (B), the expected signal at a given covariate level is indicated by the solid lines. The batch effect is the difference between the two solid lines (gray box). The estimated linear model fit employed by conditional `ComBat` is indicated with the dashed line. After adjustment (C. and D.), the original expected signal relationship is passed through the trained model, and any residual difference between the true signals of the two batches is termed a residual signal disparity (red box). When coverlap overlap is high in I, both Conditional `ComBat` and Causal `ComBat` remove the batch effect successfully, and the expected signal after correction accurately overlaps. With moderate covariate overlap in II, the limitations of the Conditional `ComBat` model introduce residual bias that results in the expected signals for each batch after correction differing in magnitude by more than the original batch effect by a factor of $> 50\%$. While Causal `ComBat` reduces the covariate space for inference to remove samples with covariates unlike the other batch (the covariate region to the far left and far right), the samples within the region of covariate overlap are successfully adjusted for a batch effect without corrupting the signal relationship. When covariate overlap is low in III, Causal strategies report that no inference can be performed, which in our opinion, is the correct procedure. On the other hand, conditional strategies may impart substantial residual signal disparities (here, about 400% the scale of the original batch effect).

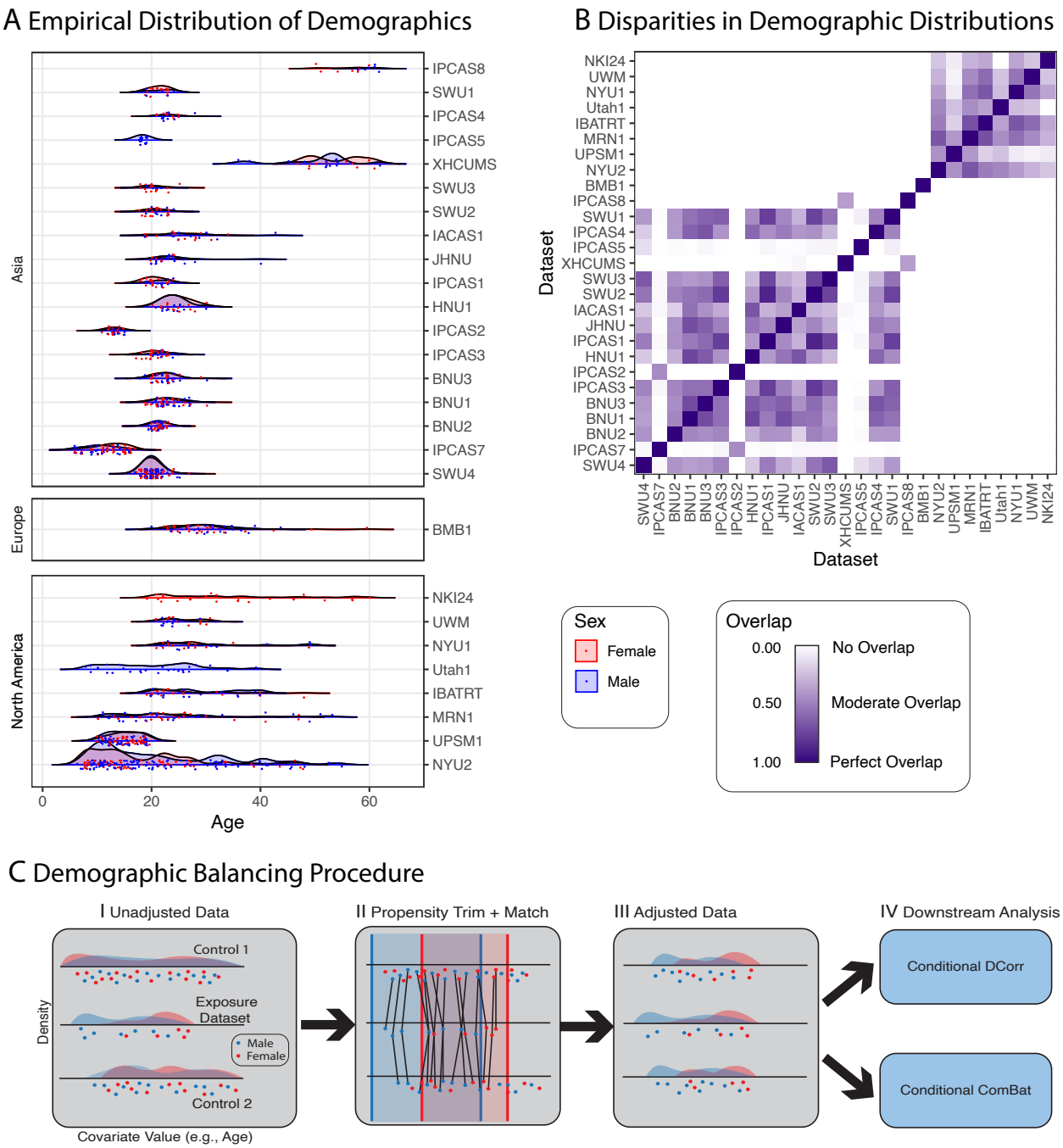

Figure 3: **Demographic data for each of the $N = 24$ studies from the CoRR Mega-Study**. **(A)** Each point represents the age of a participant corresponding to a single measurement, rows are studies, boxes are continents, and color indicates sex. **(B)** Even with only 3 covariates made available (sex, age, continent of measurement) the CoRR studies generally show extremely poor covariate overlap [32]. This makes inference regarding batch effects difficult. **(C)** The demographic balancing procedure, to demographically align poorly balanced datasets using Causal DCORR or Causal ComBat. (C.I) The unadjusted datasets are imbalanced in covariate distributions. The smallest dataset, the *reference*, has a narrower age distribution than the controls. (C.II) Propensity trimming (shaded boxes) and matching (lines) are conducted to identify reasonable matches between the reference and the control datasets. (C.III) The adjusted datasets have far improved demographic overlap after propensity trimming, matching, and removing reference subjects who do not have suitable matches in the control datasets. (C.IV) Downstream batch effect detection via DCORR or correction using ComBat are performed on the data using conditional strategies on the adjusted data.

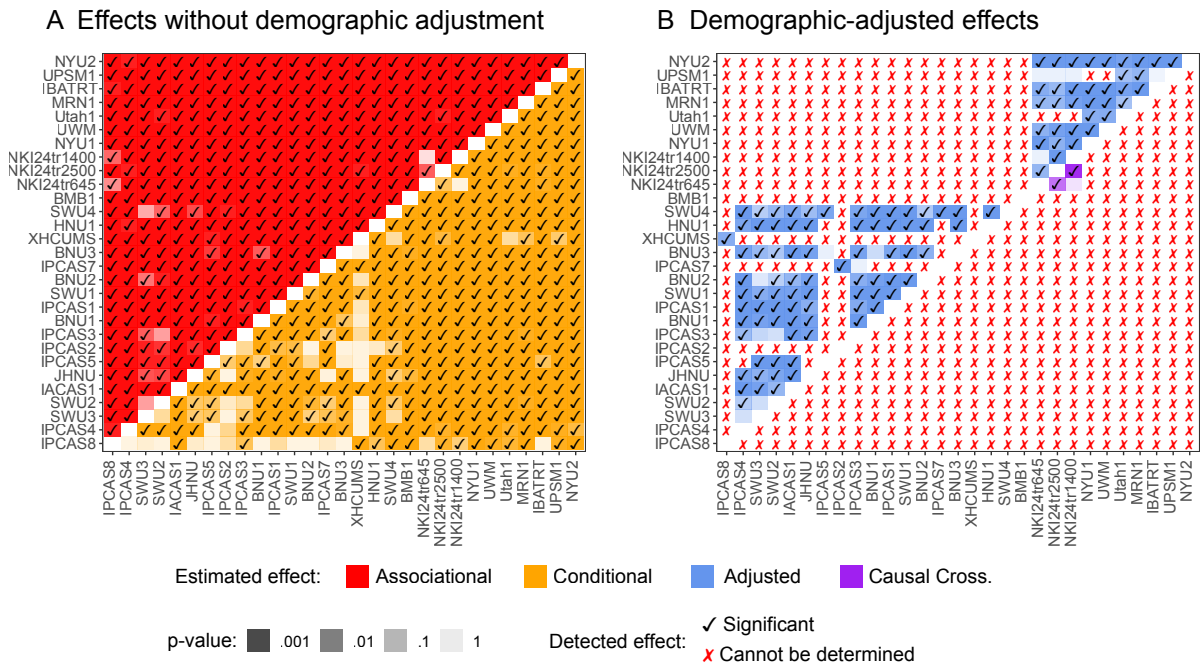

Figure 4: **Comparison of Estimated Effects for CoRR Mega-Study**. Heatmaps indicating the p-value of effect estimates per pair of studies. Studies are organized by continent, followed by the number of samples in the study. Comparisons which are significant at $\alpha = .05$ are indicated by a black 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. Effects which properly account for demographics require covariate overlap, which is indicated in Figure 3B.

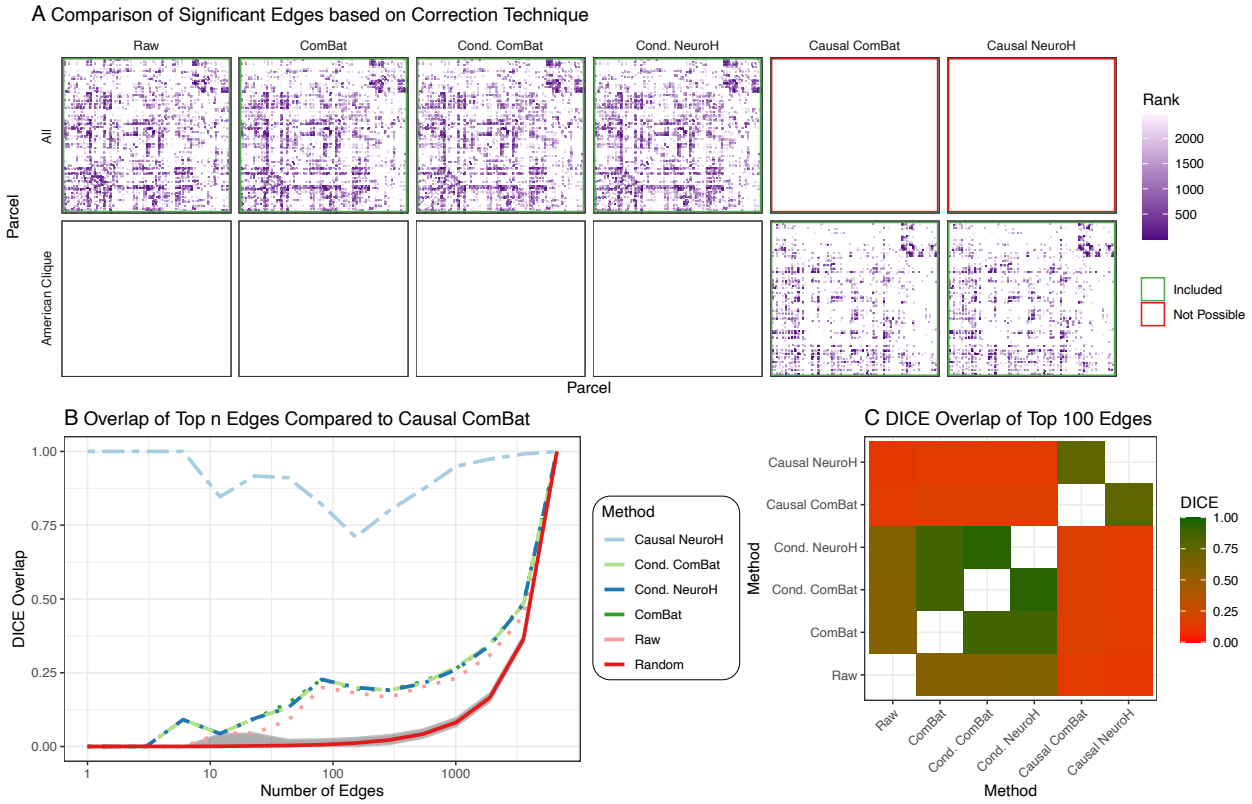

Figure 5: **Significant Edges Before and After Batch Effect Removal**. **(A)** The presence of a sex effect (conditional on individual age) is investigated for each edge in the connectome. The top 100 significant edges are shown in rank-order from largest (rank = 1) to smallest sex effect ($\alpha = .05$, Benjamini Hochberg [2] correction). 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` and Causal NeuroH can only be executed on the American Clique (red square denotes Causal approaches are 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`, Conditional `ComBat`, and Conditional NeuroH, but all techniques have low overlap with Causal `ComBat`and Causal NeuroH, indicating that Causal approaches yield measurements with 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 straightforward approach to augment existing strategies for proper batch effect correction and removal, by first prepending existing techniques using causal balancing procedures. We explore the advantages of these procedures by demonstrating that causal approaches (such as Causal `ComBat` and Causal `DCORR`) yield empirically valid approaches to handling batch effects, or report that no inference can be made. This represents a much more principled approach to alternative strategies which simply introduce or presume potentially disastrous artifactual 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 [7, 10]. Together, our work rigorously defines and studies measurement bias decoupled from demographic confounders in complex, non-Euclidean data.

The first space for improvement in our analysis is the choice of causal procedures for analysis. We took one of the simplest possible approaches (propensity score trimming followed by nearest-neighbor matching) because, in our opinions, the best place to start is typically with the most straightforward procedures to illustrate a conclusion. These strategies were sufficient to conclude that non-causal tools cannot properly isolate batch effects, and that causal approaches can trivially subvert a major limitation of the existing tools. Indeed, there are numerous other strategies one could choose to balance a dataset. Many strategies exist for robust propensity estimation, such as via Neural Networks, Support Vector Machines, or Random Forests [42, 54]. Further, numerous strategies exist for matching individuals across multiple disparate cohorts on the basis of covariates [25]. Finally, continuous re-weighting solutions can subvert the observation that matching techniques may harshly throw away data points. Data points can simply be score on the basis of how much (or little) their covariates resemble the covariates of data points in other cohorts, and can be up (or downweighted) in the test statistics or model fitting process to play a bigger (or smaller) role in the estimated[44, 47].

Together, the conclusions of our work bring into focus the pervasive question of how to properly interpret neuroimaging—or more generally, multi-site 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 to which we wish to apply the conclusions of our 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.

Through simulation, we demonstrated that minor misspecifications of the statistical model employed for batch effect removal can be disastrous in the presence of confounding biases. Specifically, a minor change of the relationship between the confounder (the covariate) and the measurement from a straight line to a (still monotonic) nonlinearity, the sigmoid, not only inhibited Conditional `ComBat` from removing a batch effect, but also introduced an artifact that dwarfed the magnitude of the original batch effect itself. As discussed in the Appendix D.2, the resulting estimate of the batch effect must, by construction, be considered neither statistically unbiased nor consistent, thus failing two preconditions for internal validity [55]. Colloquially, the statistical model is *always* wrong, and therefore we *cannot possibly* know whether models employed for batch effect correction will be reasonable to apply to out-of-sample data (e.g., areas of the covariate distribution for which there is no covariate overlap). That is, we cannot tell whether the data we have will fall into Appendix Figure 7.III and the assumptions of the batch effect correction technique will be reasonable, or 2.III where the assumptions of the batch effect correction

technique will be entirely unreasonable. On the other hand, causal strategies can help us to apply far more conservative logic, indicating to us when batch effects are unreasonable to estimate. Alternatively, they can aid us to first balance the covariate distributions so that we can avoid the pitfalls of Figure 2.B.II or 2.B.III.

Given the major shortcomings theoretically and empirically of traditional approaches for detecting and mitigating batch effects, we therefore chose to prioritize the development of strategies which sub-sample individuals into demographically-aligned subsets using causal techniques. In so doing, we prioritized internal over external validity, as internal validity is a prerequisite for external validity. Colloquially, one cannot generalize results of an experiment to a target population if the results are, at best, dubiously applicable to the study population. The trade-off of subsampling individuals is that sample size, sample diversity, and external validity are exchanged for internal validity. We saw this tradeoff materialize dramatically in the CoRR dataset, where over half of all possible dataset pairs could not be assessed for batch effects due to a lack of demographic alignment.

The final hurdle of our analysis, therefore, is to address the question: does the internal validity of batch effect removal techniques actually impact downstream statistical inference? A failure to remove batch effects in an internally valid manner, we believe, presents a substantial hurdle for producing internally valid statistical inference downstream, since downstream statistical inference is a product of the corrected measurements. 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 entire analysis only on those individuals (as causal approaches, like Causal `ComBat` or Causal NeuroH, do4). Second, one can run an algorithm for batch effect correction (and potentially benefit from a richer number of samples), and then filter to ensure demographic balancing only when measuring downstream demographic-specific effect sizes (as we did by looking for a sex effect only considering individuals from the Raw, `ComBat`, Conditional `ComBat`, and Conditional NeuroH corrected connectomes in the American Clique). If internally valid, causal batch effect correction techniques may substantially restrict the sample size. Perhaps, even if the batch effect correction technique is not internally valid, measurements corrected using it may still have utility for downstream analysis. We anticipated that the most significant features for the inference task would be, at least, ordered similarly between the two analyses. However, our finding that the analyses produce radically different downstream inferences, even when we restrict the non-causal techniques to the same datasets, suggests that there is a link between downstream conclusions and the internal validity of batch effect correction techniques. Stated another way, when we perform batch adjustment using as few assumptions as possible about the study population through causal techniques, we get extremely disparate downstream scientific conclusions from state-of-the-art conditional techniques. This suggests to us that 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 through careful consideration of the study populations of analyses.

Biomedical data science, including neuroimaging and connectomics, are plagued by a lack of generalizability and replicability [3, 48]. While many investigations posit that this may be a consequence of small sample sizes or overly simplistic datasets [29, 48] 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 [59], in which potentially demographically-disparate populations are analyzed and only loosely harmonized on the basis of the measurement protocols. We saw that post-hoc batch effect detection and removal (and consequently, downstream statistical inference) presents both theoretical and conceptual inconsistencies with regard to internal validity when not first passed through the lens of causality. This problem is highlighted by looking at the poor demographic overlap for popular neuroimaging studies such as CoRR in Figure 3, and in the ABIDE and SRPBS studies detailed in Ap-

pendix E. Future work will focus on applying the methods developed here to harmonize mega-studies such as the Adolescent Brain Cognitive Development (ABCD) [20], 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. [30] 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. [30], Yamashita et al. [56], which advocate for the aggregation of studies featuring individuals measured at a wide array of sites (so called *travelling subject* datasets), so that the nature of batch effects can be explored with minimal assumptions needed due to the crossover property. In such datasets, scientists can appreciate internally valid demographic-specific effects on the measurement, and learn more reasonable modeling assumptions, so that perhaps extrapolatory pre- or post-hoc batch effect removal techniques (such as in Appendix Figure 7.III) can be employed to mitigate the batch effects in other non-demographic-aligned datasets. 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 of existing results in scientific meta-analyses.

Code Availability The code used for analysis within this manuscript is available at github.com/ebridge2/batch_effects, and a docker container with all software versions is available at hub.docker.com/r/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. We also wanted to thank Betsy Ogburn for numerous insightful conversations regarding the presentation and framing of this work.

References

1. Yashar Behzadi, Khaled Restom, Joy Liau, and Thomas T. Liu. A Component Based Noise Correction Method (CompCor) for BOLD and Perfusion Based fMRI. Neuroimage, 37(1):90, Aug 2007. doi: 10.1016/j.neuroimage.2007.04.042.
2. Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society, 1995.
3. Eric W. Bridgeford, Shangsi Wang, Zhi Yang, Zeyi Wang, Ting Xu, Cameron Craddock, Jayanta Dey, Gregory Kiar, William Gray-Roncal, Carlo Colantuoni, Christopher Douville, Stephanie Noble, Carey E. Priebe, Brian Caffo, Michael Milham, Xi-Nian Zuo, , and Joshua T. Vogelstein. Eliminating accidental deviations to minimize generalization error and maximize reliability: applications in connectomics and genomics. bioRxiv, page 802629, Dec 2020. URL https://doi.org/10.1101/802629.
4. Cameron Craddock, Sikka Sharad, Cheung Brian, Khanuja Ranjeet, Ghosh Satrajit, Yan Chaogan, Li Qingyang, Lurie Daniel, Joshua Vogelstein, Randal Burns, Colcombe Stanley, Maarten Mennes, Kelly Clare, Di Adriana, Francisco Castellanos, and Milham Michael. Towards Automated Analysis of Connectomes: The Configurable Pipeline for the Analysis of Connectomes (C-PAC). Front. Neuroinf., 7, Jan 2013. doi: 10.3389/conf.fninf.2013.09.00042.
5. Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular psychiatry, 19(6):659–667, 2014.
6. Adriana Di Martino, David O'connor, Bosi Chen, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Joshua H Balsters, Leslie Baxter, Anita Beggiato, Sylvie Bernaerts, et al. Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii. Scientific data, 4(1):1–15, 2017.
7. Peng Ding and Luke W. Miratrix. To Adjust or Not to Adjust? Sensitivity Analysis of M-Bias and Butterfly-Bias. Journal of Causal Inference, 3(1):41–57, Mar 2015. ISSN 2193-3685. doi: 10.1515/jci-2013-0021.
8. Bradley Efron. Large-Scale Simultaneous Hypothesis Testing. J. Am. Stat. Assoc., 99(465):96–104, Mar 2004. ISSN 0162-1459. doi: 10.1198/016214504000000089.
9. V. S. Fonov, A. C. Evans, R. C. McKinstry, C. R. Almli, and D. L. Collins. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. Neuroimage, 47:S102, Jul 2009. ISSN 1053-8119. doi: 10.1016/S1053-8119(09)70884-5.
10. Wolfgang Forstmeier, Eric-Jan Wagenmakers, and Timothy H. Parker. Detecting and avoiding likely false-positive findings – a practical guide. Biol. Rev., 92(4):1941–1968, Nov 2017. ISSN 1464-7931. doi: 10.1111/brv.12315.
11. Jean-Philippe Fortin, Nicholas Cullen, Yvette I. Sheline, Warren D. Taylor, Irem Aselcioglu, Philip A. Cook, Phil Adams, Crystal Cooper, Maurizio Fava, Patrick J. McGrath, Melvin McInnis, Mary L.

Phillips, Madhukar H. Trivedi, Myrna M. Weissman, and Russell T. Shinohara. Harmonization of cortical thickness measurements across scanners and sites. Neuroimage, 167:104, February 2018. doi: 10.1016/j.neuroimage.2017.11.024.

12. K. J. Friston, S. Williams, R. Howard, R. S. Frackowiak, and R. Turner. Movement-related effects in fMRI time-series. Magn. Reson. Med., 35(3):346–355, Mar 1996. ISSN 0740-3194. doi: 10.1002/mrm.1910350312.
13. Douglas N. Greve and Bruce Fischl. Accurate and robust brain image alignment using boundary-based registration. Neuroimage, 48(1):63–72, Oct 2009. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2009.06.060.
14. Michelle Hampson, Fuyuze Tokoglu, Xilin Shen, Dustin Scheinost, Xenophon Papademetris, and R. Todd Constable. Intrinsic Brain Connectivity Related to Age in Young and Middle Aged Adults. PLoS One, 7(9):e44067, Sep 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0044067.
15. Daniel E Ho, Kosuke Imai, Gary King, and Elizabeth A Stuart. MatchIt: Nonparametric preprocessing for parametric causal inference. Journal of Statistical Software, 42(8):1–28, 2011. URL https://www.jstatsoft.org/v42/i08/.
16. Madhura Ingalhalikar, Alex Smith, Drew Parker, Theodore D. Satterthwaite, Mark A. Elliott, Kosha Ruparel, Hakon Hakonarson, Raquel E. Gur, Ruben C. Gur, and Ragini Verma. Sex differences in the structural connectome of the human brain. Proc. Natl. Acad. Sci. U.S.A., 111(2):823–828, Jan 2014. ISSN 0027-8424. doi: 10.1073/pnas.1316909110.
17. Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage, 17 (2):825–841, Oct 2002. ISSN 1053-8119. doi: 10.1016/s1053-8119(02)91132-8.
18. Mark Jenkinson, Christian F. Beckmann, Timothy E. J. Behrens, Mark W. Woolrich, and Stephen M. Smith. FSL. Neuroimage, 62(2):782–790, Aug 2012. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2011.09.015.
19. W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 8(1):118–127, Jan 2007. ISSN 1465-4644. doi: 10.1093/biostatistics/kxj037.
20. Nicole R. Karcher and Deanna M. Barch. The ABCD study: understanding the development of risk for mental and physical health outcomes - Neuropsychopharmacology. Neuropsychopharmacology, 46(1):131–142, Jan 2021. ISSN 1740-634X. doi: 10.1038/s41386-020-0736-6.
21. Gregory Kiar, Eric W. Bridgeford, William R. Gray Roncal, , Vikram Chandrashekhar, Disa Mhembere, Sephira Ryman, Xi-Nian Zuo, Daniel S. Margulies, R. Cameron Craddock, Carey E. Priebe, Rex Jung, Vince D. Calhoun, Brian Caffo, Randal Burns, Michael P. Milham, and Joshua T. Vogelstein. A High-Throughput Pipeline Identifies Robust Connectomes But Troublesome Variability. bioRxiv, page 188706, Apr 2018. URL https://doi.org/10.1101/188706.
22. Cosmin Lazar, Stijn Meganck, Jonatan Taminau, David Steenhoff, Alain Coletta, Colin Molter, David Y. Weiss-Solís, Robin Duque, Hugues Bersini, and Ann Nowé. Batch effect removal methods for microarray gene expression data integration: a survey. Briefings Bioinf., 14(4):469–490, Jul 2013. ISSN 1477-4054. doi: 10.1093/bib/bbs037.
23. Jeffrey T Leek and Roger 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, April 2015. Accessed: 2020-10-12.
24. Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet., 11(10):733–739, October 2010.
25. Michael J. Lopez and Roee Gutman. Estimation of Causal Effects with Multiple Treatments: A

Review and New Ideas. Statist. Sci., 32(3):432–454, August 2017. ISSN 0883-4237. doi: 10.1214/17-STS612.

26. Wang Miao, Wenjie Hu, Elizabeth L. Ogburn, and Xiaohua Zhou. Identifying effects of multiple treatments in the presence of unmeasured confounding. arXiv, Nov 2020. URL https://arxiv.org/abs/2011.04504v3.
27. Maria B. Misiura, J. Christina Howell, Junjie Wu, Deqiang Qiu, Monica W. Parker, Jessica A. Turner, and William T. Hu. Race modifies default mode connectivity in Alzheimer's disease. Transl. Neurodegener., 9, 2020. doi: 10.1186/s40035-020-0186-4.
28. 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. doi: 10.1006/nimg.2001.0978.
29. Derek Evan Nee. fMRI replicability depends upon sufficient individual-level data. Commun. Biol., 2 (130):1–4, Apr 2019. ISSN 2399-3642. doi: 10.1038/s42003-019-0378-6.
30. Stephanie Noble, Dustin Scheinost, Emily S. Finn, Xilin Shen, Xenophon Papademetris, Sarah C. McEwen, Carrie E. Bearden, Jean Addington, Bradley Goodyear, Kristin S. Cadenhead, Heline Mirzakhanian, Barbara A. Cornblatt, Doreen M. Olvet, Daniel H. Mathalon, Thomas H. McGlashan, Diana O. Perkins, Aysenil Belger, Larry J. Seidman, Heidi Thermenos, Ming T. Tsuang, Theo G. M. van Erp, Elaine F. Walker, Stephan Hamann, Scott W. Woods, Tyrone D. Cannon, and R. Todd Constable. Multisite reliability of MR-based functional connectivity. Neuroimage, 146:959–970, Feb 2017. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2016.10.020.
31. Kate Brody Nooner, Stanley Colcombe, Russell Tobe, Maarten Mennes, Melissa Benedict, Alexis Moreno, Laura Panek, Shaquanna Brown, Stephen Zavitz, Qingyang Li, Sharad Sikka, David Gutman, Saroja Bangaru, Rochelle Tziona Schlachter, Stephanie Kamiel, Ayesha Anwar, Caitlin Hinz, Michelle Kaplan, Anna Rachlin, Samantha Adelsberg, Brian Cheung, Ranjit Khanuja, Chaogan Yan, Cameron Craddock, Vincent Calhoun, William Courtney, Margaret King, Dylan Wood, Christine Cox, Clare Kelly, Adriana DiMartino, Eva Petkova, Philip Reiss, Nancy Duan, Dawn Thompsen, Bharat Biswal, Barbara Coffey, Matthew Hoptman, Daniel C. Javitt, Nunzio Pomara, John Sidtis, Harold Koplewicz, Francisco Xavier Castellanos, Bennett Leventhal, and Michael Milham. The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry. Front. Neurosci., 6, Oct 2012. ISSN 1662-453X. doi: 10.3389/fnins.2012.00152.
32. Massimiliano Pastore and Antonio Calcagnì. Measuring Distribution Similarities Between Samples: A Distribution-Free Overlapping Index. Front. Psychol., 0, 2019. ISSN 1664-1078. doi: 10.3389/fpsyg.2019.01089.
33. Judea Pearl. Causal inference in statistics: An overview. ssu, 3(none):96–146, Jan 2009. ISSN 1935-7516. doi: 10.1214/09-SS057.
34. Judea Pearl. On measurement bias in causal inference. AUAI Press, Arlington, TX, USA, Jul 2010. ISBN 978-0-97490396-5. doi: 10.5555/3023549.3023599.
35. Raymond Pomponio, Guray Erus, Mohamad Habes, Jimit Doshi, Dhivya Srinivasan, Elizabeth Mamourian, Vishnu Bashyam, Ilya M. Nasrallah, Theodore D. Satterthwaite, Yong Fan, Lenore J. Launer, Colin L. Masters, Paul Maruff, Chuanjun Zhuo, Henry Völzke, Sterling C. Johnson, Jurgen Fripp, Nikolaos Koutsouleris, Daniel H. Wolf, Raquel Gur, Ruben Gur, John Morris, Marilyn S. Albert, Hans J. Grabe, Susan M. Resnick, R. Nick Bryan, David A. Wolk, Russell T. Shinohara, Haochang Shou, and Christos Davatzikos. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. Neuroimage, 208:116450, Mar 2020. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2019.116450.
36. Michael Powell, Allison Koenecke, James Brian Byrd, Akihiko Nishimura, Maximilian F. Konig, Ruoxuan Xiong, Sadiqa Mahmood, Vera Mucaj, Chetan Bettegowda, Liam Rose, Suzanne Tamang, Adam Sacarny, Brian Caffo, Susan Athey, Elizabeth A. Stuart, and Joshua T. Vogelstein. A how-to

guide for conducting retrospective analyses: example COVID-19 study. OSF Preprints, Sep 2020. doi: 10.31219/osf.io/3drch.
37. James 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, 7(9): 1393–1512, Jan 1986. ISSN 0270-0255. doi: 10.1016/0270-0255(86)90088-6.
38. Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, Apr 1983. ISSN 0006-3444. doi: 10.1093/biomet/70.1.41.
39. Paul R. Rosenbaum and Donald B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician, 39(1):33–38, 1985. ISSN 00031305. URL http://www.jstor.org/stable/2683903.
40. Roser Sala-Llonch, David Bartrés-Faz, and Carme Junqué. Reorganization of brain networks in aging: a review of functional connectivity studies. Front. Psychol., 0, 2015. ISSN 1664-1078. doi: 10.3389/fpsyg.2015.00663.
41. Theodore D. Satterthwaite, Daniel H. Wolf, David R. Roalf, Kosha Ruparel, Guray Erus, Simon Vandekar, Efstathios D. Gennatas, Mark A. Elliott, Alex Smith, Hakon Hakonarson, Ragini Verma, Christos Davatzikos, Raquel E. Gur, and Ruben C. Gur. Linked Sex Differences in Cognition and Functional Connectivity in Youth. Cerebral Cortex (New York, NY), 25(9):2383, Sep 2015. doi: 10.1093/cercor/bhu036.
42. Soko Setoguchi, Sebastian Schneeweiss, M. Alan Brookhart, Robert J. Glynn, and E. Francis Cook. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiol. Drug Saf., 17(6):546, June 2008. doi: 10.1002/pds.1555.
43. Cencheng Shen, Carey E Priebe, and Joshua T Vogelstein. From Distance Correlation to Multiscale Generalized Correlation. Journal of American Statistical Association, October 2017. URL http://arxiv.org/abs/1710.09768.
44. Elizabeth A. Stuart. Matching Methods for Causal Inference: A Review and a Look Forward. Statist. Sci., 25(1):1–21, Feb 2010. ISSN 0883-4237. doi: 10.1214/09-STS313.
45. Gábor J. Székely and Maria L. Rizzo. Partial distance correlation with methods for dissimilarities. Ann. Stat., 42(6):2382–2412, Dec 2014. ISSN 0090-5364. doi: 10.1214/14-AOS1255.
46. Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. Ann. Stat., 35(6):2769–2794, Dec 2007. ISSN 0090-5364. doi: 10.1214/009053607000000505.
47. Zhiqiang Tan. Regression and Weighting Methods for Causal Inference Using Instrumental Variables. J. Am. Stat. Assoc., 101(476):1607–1618, December 2006. ISSN 0162-1459. doi: 10.1198/016214505000001366.
48. Benjamin O. Turner, Erick J. Paul, Michael B. Miller, and Aron K. Barbey. Small sample sizes reduce the replicability of task-based fMRI studies. Commun. Biol., 1(62):1–10, Jun 2018. ISSN 2399-3642. doi: 10.1038/s42003-018-0073-z.
49. A. W. van der Vaart. Asymptotic Statistics. Cambridge University Press, Cambridge, England, UK, October 1998. ISBN 978-0-51180225-6. doi: 10.1017/CBO9780511802256.
50. Eleanna Varangis, Christian G. Habeck, Qolamreza R. Razlighi, and Yaakov Stern. The Effect of Aging on Resting State Connectivity of Predefined Networks in the Brain. Front. Aging Neurosci., 0, 2019. ISSN 1663-4365. doi: 10.3389/fnagi.2019.00234.
51. Joshua T Vogelstein, Eric W Bridgeford, Qing Wang, Carey E Priebe, Mauro Maggioni, and Cencheng Shen. Discovering and deciphering relationships across disparate data modalities. Elife, 8, January 2019. URL http://dx.doi.org/10.7554/eLife.41690.
52. Christian Wachinger, Anna Rieckmann, and Sebastian Pölsterl. Detect and Correct Bias in Multi-Site Neuroimaging Datasets. arXiv, Feb 2020. doi: 10.1016/j.media.2020.101879.
53. Susanne Weis, Kaustubh R. Patil, Felix Hoffstaedter, Alessandra Nostro, B. T. Thomas Yeo, and

Simon B. Eickhoff. Sex Classification by Resting State Brain Connectivity. Cereb. Cortex, 30(2): 824–835, Mar 2020. ISSN 1047-3211. doi: 10.1093/cercor/bhz129.

54. Daniel Westreich, Justin Lessler, and Michele Jonsson Funk. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. J. Clin. Epidemiol., 63(8):826–833, August 2010. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2009.11.020.
55. Daniel Westreich, Jessie K. Edwards, Catherine R. Lesko, Stephen R. Cole, and Elizabeth A. Stuart. Target Validity and the Hierarchy of Study Designs. Am. J. Epidemiol., 188(2):438, February 2019. doi: 10.1093/aje/kwy228.
56. Ayumu Yamashita, Noriaki Yahata, Takashi Itahashi, Giuseppe Lisi, Takashi Yamada, Naho Ichikawa, Masahiro Takamura, Yujiro Yoshihara, Akira Kunimatsu, Naohiro Okada, Hirotaka Yamagata, Koji Matsuo, Ryuichiro Hashimoto, Go Okada, Yuki Sakai, Jun Morimoto, Jin Narumoto, Yasuhiro Shimada, Kiyoto Kasai, Nobumasa Kato, Hidehiko Takahashi, Yasumasa Okamoto, Saori C. Tanaka, Mitsuo Kawato, Okito Yamashita, and Hiroshi 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., 17(4):e3000042, Apr 2019. ISSN 1545-7885. doi: 10.1371/journal.pbio.3000042.
57. Meichen Yu, Kristin A. Linn, Philip A. Cook, Mary L. Phillips, Melvin McInnis, Maurizio Fava, Madhukar H. Trivedi, Myrna M. Weissman, Russell T. Shinohara, and Yvette I. Sheline. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. Hum. Brain Mapp., 39(11):4213–4227, Nov 2018. ISSN 1065-9471. doi: 10.1002/hbm.24241.
58. Yuqing Zhang, Giovanni Parmigiani, and W. Evan Johnson. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics Bioinf., 2(3), Sep 2020. ISSN 2631-9268. doi: 10.1093/nargab/lqaa078.
59. Xi-Nian Zuo, Jeffrey S. Anderson, Pierre Bellec, Rasmus M. Birn, Bharat B. Biswal, Janusch Blautzik, John C. S. Breitner, Randy L. Buckner, Vince D. Calhoun, F. Xavier Castellanos, Antao Chen, Bing Chen, Jiangtao Chen, Xu Chen, Stanley J. Colcombe, William Courtney, R. Cameron Craddock, Adriana Di Martino, Hao-Ming Dong, Xiaolan Fu, Qiyong Gong, Krzysztof J. Gorgolewski, Ying Han, Ye He, Yong He, Erica Ho, Avram Holmes, Xiao-Hui Hou, Jeremy Huckins, Tianzi Jiang, Yi Jiang, William Kelley, Clare Kelly, Margaret King, Stephen M. LaConte, Janet E. Lainhart, Xu Lei, Hui-Jie Li, Kaiming Li, Kuncheng Li, Qixiang Lin, Dongqiang Liu, Jia Liu, Xun Liu, Yijun Liu, Guangming Lu, Jie Lu, Beatriz Luna, Jing Luo, Daniel Lurie, Ying Mao, Daniel S. Margulies, Andrew R. Mayer, Thomas Meindl, Mary E. Meyerand, Weizhi Nan, Jared A. Nielsen, David O'Connor, David Paulsen, Vivek Prabhakaran, Zhigang Qi, Jiang Qiu, Chunhong Shao, Zarrar Shehzad, Weijun Tang, Arno Villringer, Huiling Wang, Kai Wang, Dongtao Wei, Gao-Xia Wei, Xu-Chu Weng, Xuehai Wu, Ting Xu, Ning Yang, Zhi Yang, Yu-Feng Zang, Lei Zhang, Qinglin Zhang, Zhe Zhang, Zhiqiang Zhang, Ke Zhao, Zonglei Zhen, Yuan Zhou, Xing-Ting Zhu, and Michael P. Milham. An open science resource for establishing reliability and reproducibility in functional connectomics. Sci. Data, 1(140049):1–13, Dec 2014. ISSN 2052-4463. doi: 10.1038/sdata.2014.49.

Appendix A. Procedures for detecting and mitigating batch effects.

A.1 Detecting batch effects with Causal DCORR 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 of Causal DCORR, in which a conditional two-sample test is performed on pairs of observational studies which are balanced against measured demographic 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. Propensity trim and then match the largest integer less than $n_k/n_{k'}$ control individuals using nearest-neighbor matching to individuals in the exposure dataset. Methods B 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 [45, 46] 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.

Figure 3C illustrates visually the causal procedures employed for demographic balancing. In the event that the conditioning set closes backdoor paths [33, 34], the adjusted effect detected by Causal DCORR is a causal batch effect, and does not require extrapolation assumptions like Conditional DCORR [15, 38, 39, 44]. If the conditioning set does not close backdoor paths, the effect detected is a conditional effect and may yield the detection of demographic effects.

A.2 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 propose 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. Propensity trim and then match the greatest integer less than $n_k/n_{k'}$ control individuals using nearest-neighbor matching to individuals in the exposure dataset. Methods B 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 [19] 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 [33, 34], Causal ComBat yields the removal of an internally valid causal effect and does not require extrapolation assumptions, unlike Conditional ComBat [15, 38, 39, 44]. 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, as we saw in Figure 6. Appendix E depicts the impact on the empirical covariate distribution of the adjustment procedure.

Appendix B. Methods.

B.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 [15]. 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 [36].

B.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 [46], which is a non-parametric test for testing whether two variables are correlated. A simple augmentation of the distance correlation procedure [43, 51] 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 [45], 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 [8]. 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 [2].

B.3 Site Effect Correction To correct for site effects, we compare both z-scoring per-feature within each batch [22] and traditional `ComBat` to Causal `ComBat`.

B.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

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 [28]. 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 $\texttt{pdcorr}(\vec{y}(k, l), \vec{s}|\vec{a})$.

B.5 Datasets The CoRR Mega-Study [59] is an aggregate dataset consisting of 24 studies collected with a similar goal: assessing the reliability and reproducibility of neuroimaging data. The mega-study consists of 1313 individuals, most of whom are measured numerous times, for a total of $N = 3597$ connectomes. All connectomes are estimated using the `m2g` (MRI to Graphs) pipeline [21], which provides a wrapper for the `CPAC` Pipeline [4]. fMRI scans for each individual are first processed to remove motion artifacts using `mcflirt` [17]. The fMRI scans are then registered to the corresponding individual's anatomical scan using FSL's boundary-based registration (BBR) via `epireg` [13]. A non-linear transformation from the individual's anatomical scan to the MNI152 [9] template is estimtaed using `FNIRT` [18]. Nuisance artifacts are removed by fitting the voxelwise timeseries to a regression model incorporating regressors for the Friston 24-parameter model [12], the top five principal components of the voxelwise timeseries in cerebrospinal fluid `aCompCor` [1], and a quadratic drift term. The adjusted voxelwise timeseries is downsampled to the regions of interest (ROIs) of the Automated Anatomical Labelling (AAL) parcellation [28] by taking the spatial mean signal for each timepoint across voxels within the region of interest. Functional connectivity is estimated using the pairwise correlation between all pairs of ROI timeseries within the AAL parcellation. 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. For each study, 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. The 833 connectomes comprising the studies of the American clique are reduced to the $N = 321$ connectomes with maximal demographic overlap identified through covariate adjustment B.1.

The NKI Rockland Sample [31] 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.

Appendix C. Correction for Batch-Related Effects.

Appendix D. Simulations.

D.1 Batch Effect Detection Simulations $R = 200$ iterations of each simulation are produced and the following question is asked: do the measurements, conditional on the covariates, differ between batch 0 and batch 1? For each iteration, for each simulation, we sample $n = 100$ points from Batch 0 or Batch 1 with probability 0.5; e.g., $T_i \overset{iid}{\sim} \text{Bern}(0.5)$. The simulation settings are:

Limited Overlap, Signal The covariate distribution is:

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. `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. `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

$$X_i | T_i = t = \overset{d}{\sim} \begin{cases} \text{Beta}(2, 8) & t = 0 \\ \text{Beta}(8, 2) & t = 1 \end{cases}$$

With $Z_i \overset{iid}{\sim} \mathcal{N}(0,1)$, the measurement is:

$$Y_i = -2(X_i - .5) + \frac{1}{2}Z_i + T_i.$$

Limited Overlap, Null The covariate distribution is:

$$X_i | T_i = t \overset{d}{\sim} \begin{cases} \text{Beta}(2,8) & t = 0 \\ \text{Beta}(8,2) & t = 1 \end{cases}$$

With $Z_i \overset{iid}{\sim} \mathcal{N}(0,1)$, the measurement is:

$$Y_i = -2(X_i - .5) + \frac{1}{2}Z_i.$$

Overlap, Signal The covariate distribution is $X_i \overset{iid}{\sim} \text{Unif}(0,1)$. With $Z_i \overset{iid}{\sim} \mathcal{N}(0,1)$, the measurement is:

$$Y_i = -2(X_i - .5) + \frac{1}{2}Z_i + T_i.$$

Overlap, Null The covariate distribution is $X_i \overset{iid}{\sim} \text{Unif}(0,1)$. With $Z_i \overset{iid}{\sim} \mathcal{N}(0,1)$, the measurement is:

$$Y_i = -2(X_i - .5) + \frac{1}{2}Z_i.$$

Conditional DCORR and Causal DCORR are used to test for conditional and adjusted effects respectively in Figure 6. Across all simulations, simulated measurements are collected from a gaussian additive model, where the measurement tends to be lower when the covariate is higher, and the measurement tends to be higher when the covariate is lower. Simulated measurements are collected from one of the two batches with equal probability (0.5). In the bottom two simulations, the covariate distribution is equal for the two indicated batches, and both techniques are able to nearly always correctly identify whether or not a batch effect is present. In the top two simulations, however, the covariate distribution in batch 0 tends to favor smaller covariate values, whereas for batch 1 the covariate distribution is higher. This makes isolating the batch effect from the covariate-specific effect difficult without heavy assumptions regarding the measurement distribution, which are often not feasible or well enough established for high dimensional measurements. Conditional procedures struggle with this setting, and nearly always report the incorrect answer (detecting a batch effect when none is present, or vice-versa). On the other hand, causal strategies tend to be far more conservative in these types of difficult settings, and tend to report that any further analysis of the dataset will be inconclusive. These simulations highlight the empirical value of causal techniques, as the ability to conclude that the data do not support conclusions you are trying to make is far more principled than proceeding and doing the *wrong* thing.

D.2 Batch Effect Removal Simulations $n = 400$ points are sampled from Batch 0 or Batch 1 with probability 0.5; e.g. $T_i \overset{iid}{\sim} \text{Bern}(0.5)$. We have the following 4 simulations, in which the covariates and measurements are normalized to fall on the same scale for visualization purposes:

Assumption Correct, Overlap The covariate distribution is $X_i \overset{iid}{\sim} \text{Unif}(0,1)$. With $Z_i \overset{iid}{\sim} \mathcal{N}(0,1)$, the measurement is:

$$Y_i = -2(X_i - 0.5) + Z_i + T_i.$$

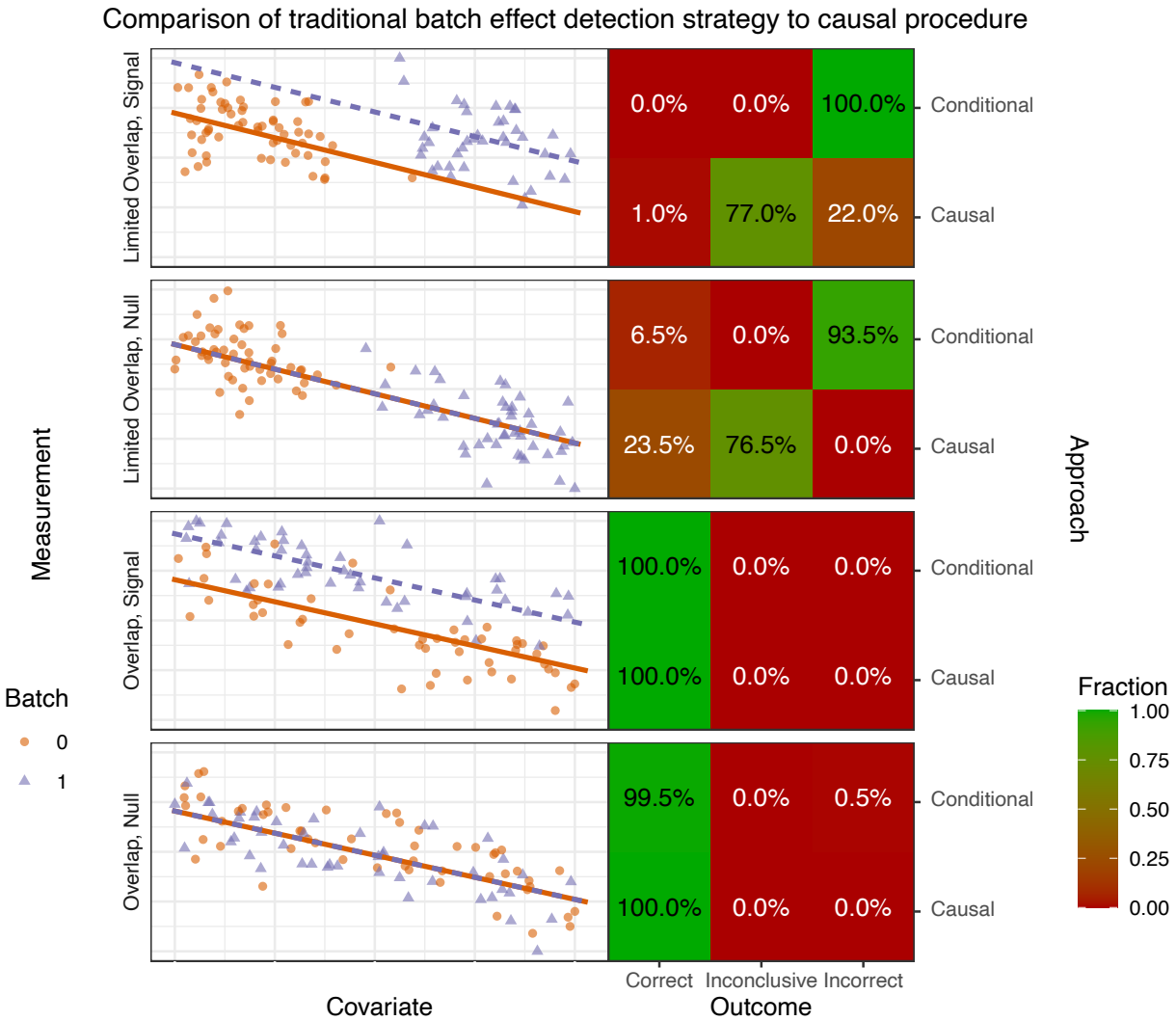

Figure 6: **Empirical simulations demonstrate the utility of causal approaches for batch effect detection**. $n = 400$ simulated univariate measurements are collected from a gaussian additive model (points) with equal probability of being from batch 0 or batch 1 (color). The expected measurement for a given batch at an indicated covariated value is indicated by the lines. The lines overlap when no batch effect is present (null simulations, rows two and four) and do not overlap when a batch effect is present (signal simulations, rows one and three). Tests are performed to determine whether a batch effect is present (or not) using conditional and causal strategies for each of $R = 200$ repeated simulations for each setting. For the top two rows, the covariate overlap between the two batches is limited (batch 0 tends to have lower covariate values, and batch 1 tends to have high covariate values). Consequently, isolating the covariate effect on the measurement from the batch effect on the measurement is difficult. Causal DCORR tends to be sensitive to this type of disparity and reasonably reports that the covariates are inconclusive to support inference for this particular question. On the other hand, Conditional DCORR tends to almost always yield the wrong answer (a batch effect is reported when none is present, and vice versa). Both Conditional and Causal DCORR tend to correctly identify batch effects when covariate overlap is high (bottom two rows). The details of these simulations are found in Appendix D.1.

Assumption Correct, Moderate Overlap The covariate distribution is:

$$X_i | T_i = t \overset{d}{\sim} \begin{cases} \text{Beta}(2, 6) & t = 0 \\ \text{Beta}(6, 2) & t = 1 \end{cases},$$

The measurement is:

$$Y_i = -2(X_i - 0.5) + Z_i + T_i.$$

Assumption Correct, Limited Overlap The covariate distribution is:

$$X_i | T_i = t \overset{d}{\sim} \begin{cases} \text{Beta}(2, 8) & t = 0 \\ \text{Beta}(8, 2) & t = 1 \end{cases},$$

The measurement is:

$$Y_i = -2(X_i - 0.5) + Z_i + T_i.$$

Assumption Incorrect, Overlap The covariate distribution is $X_i \overset{iid}{\sim} \text{Unif}(-1, 1)$. With $Z_i \overset{iid}{\sim} \mathcal{N}(0, 1)$, the measurement is:

$$Y_i = 4\text{sigmoid}(8X_i) + Z_i + T_i,$$

where sigmoid(x) is the non-linear sigmoid function; e.g.:

$$\text{sigmoid}(x) = \frac{1}{1 + \exp(-x)}.$$

We chose the sigmoid function because, under the limited overlap covariate regimes employed, where X_i is conditionally (on the batch) beta distributed, the non-linear sigmoid appears, visually, to be locally linear. Therefore, if one were to attempt to argue that exploratory analysis could be employed to post-hoc tune the statistical model employed by `ComBat`, the point of inflection of the sigmoid would not be prominent in either of the batches.

Assumption Incorrect, Moderate Overlap The covariate distribution is:

$$X_i | T_i = t \overset{d}{\sim} \begin{cases} \text{Beta}(2, 6) & t = 0 \\ \text{Beta}(6, 2) & t = 1 \end{cases}.$$

With $Z_i \overset{iid}{\sim} \mathcal{N}(0, 1)$, the measurement is:

$$Y_i = 4\text{sigmoid}(8X_i) + Z_i + T_i.$$

Assumption Incorrect, Limited Overlap The covariate distribution is:

$$X_i | T_i = t \overset{d}{\sim} \begin{cases} \text{Beta}(2, 8) & t = 0 \\ \text{Beta}(8, 2) & t = 1 \end{cases}.$$

With $Z_i \overset{iid}{\sim} \mathcal{N}(0, 1)$, the measurement is:

$$Y_i = 4\text{sigmoid}(8X_i) + Z_i + T_i.$$

Notation Across all simulations, we term the *raw true signal* at a particular covariate level x to be the quantity:

$$\mathbb{E}_{f_{x,z}(y|t)}[Y_i]$$

where x is the particular covariate level, z are (potential) unknown covariates (which are not present for the purposes of this simulation), and t is the batch.

$n = 100$ triplets (t_i, x_i, y_i) are sampled for each of the four simulations, and are passed to the `ComBat` algorithm to hypothetically detect and remove the batch effect. For a given covariate value x_i and a particular batch t_i, the *raw assumed signal* is the term:

$$\hat{y}_i = \hat{\beta}_x \hat{x}_i + \hat{\beta}_1 \mathbb{1}_{\{t_i=1\}} + \hat{\beta}_t \mathbb{1}_{\{t_i=0\}} + \bar{\mu}$$

where $\hat{\beta}_x$ is the regression coefficient for the covariate x, β_1 is an intercept offset term for batch 1, β_0 is an intercept offset term for batch 0, and $\bar{\mu}$ is the grand mean of the data.

To adjust points for the batch effect, `ComBat` simply removes the estimated batch intercepts; e.g.:

$$y_i^{(a)} = y_i - \left(\hat{\beta}_1 \mathbb{1}_{\{t_i=1\}} + \hat{\beta}_0 \mathbb{1}_{\{t_i=0\}}\right)$$

For a particular batch t, the true signal after adjustment is the term:

$$\mathbb{E}_{f_{x,z}(y|t)}[Y_i] - \left(\hat{\beta}_1 \mathbb{1}_{\{t_i=1\}} + \hat{\beta}_t \mathbb{1}_{\{t_i=0\}}\right)$$

And the assumed signal after adjustment is the term:

$$\hat{y}_i - \left(\hat{\beta}_1 \mathbb{1}_{\{t_i=1\}} + \hat{\beta}_0 \mathbb{1}_{\{t_i=0\}}\right) = \hat{\beta}_x \hat{x}_i + \bar{\mu}$$

It is obvious that after adjustment, the assumed signals for batches 0 and 1 will be equal, as $\hat{\beta}_x \hat{x}_i + \bar{\mu}$ does not depend on t_i.

D.3 Causal `ComBat` outperforms Conditional `ComBat` on simulated data Across all simulations, simulated measurements are collected from an additive gaussian model with either a linear (Figure 7.I to 7.III) or non-linear (Figure 7.IV to 7.VI) covariate-specific effect as well as a batch-specific offset in 7(A). We performed these simulations for cases where the covariate distributions overlap (Figure 7.I and Figure 7.IV), have a moderate degree of overlap (Figure 7.II and Figure 7V), and have limited overlap (Figure 7.III and Figure 7.VI). When the assumptions between the covariate and the measurements are correct, covariate overlap is irrelevant, as the covariate-specific signal can be extrapolated successfully across the entire range of the covariate. When the assumptions between the covariate and the measurements are incorrect, the batch effect can still be removed, as shown in Figure 7.III, when further causal adjustment conditions are satisfied (such as the covariate distributions overlapping). This can be conceptualized as the inaccurate assumptions of the model, in effect, being uniformly *misrepresented* by the model in each of the two batches, which *still* enables a consistent estimate of the batch effect to be estimated and removed.

As shown in the simulation, these artifacts can be arbitrarily large depending on the nature of the incorrect assumption. In this example, the batch effect artifact introduced by conditional `ComBat` exceeds the magnitude of the batch effect in the raw data for V and VI. On the other hand, as shown in Figure 7.C, Causal `ComBat` first pre-aligns the covariate distributions, and even with model misspecifications, is able to successfully remove a batch effect for the subset of the distribution with sufficient covariate overlap. Significantly, causal strategies *explicitly prohibit* learning about data which falls outside of the region of demographic overlap. When overlap conditions are not met for a portion of the sample space, causal strategies do not do anything, as in C.III. In this sense, causal techniques are therefore more conservative for batch effect detection and mitigation. We believe this to be favorable for batch effect detection and mitigation as, in practice, it is *impossible* to tell whether the implicit or explicit assumptions of the model for batch effect correction are sufficient or not. This presents an enormous point of caution when applying batch effect removal techniques to high-dimensional datasets with confounding, as it is infeasible to explore whether extrapolatory modelling assumptions are within reason on a dimension-by-dimension basis.

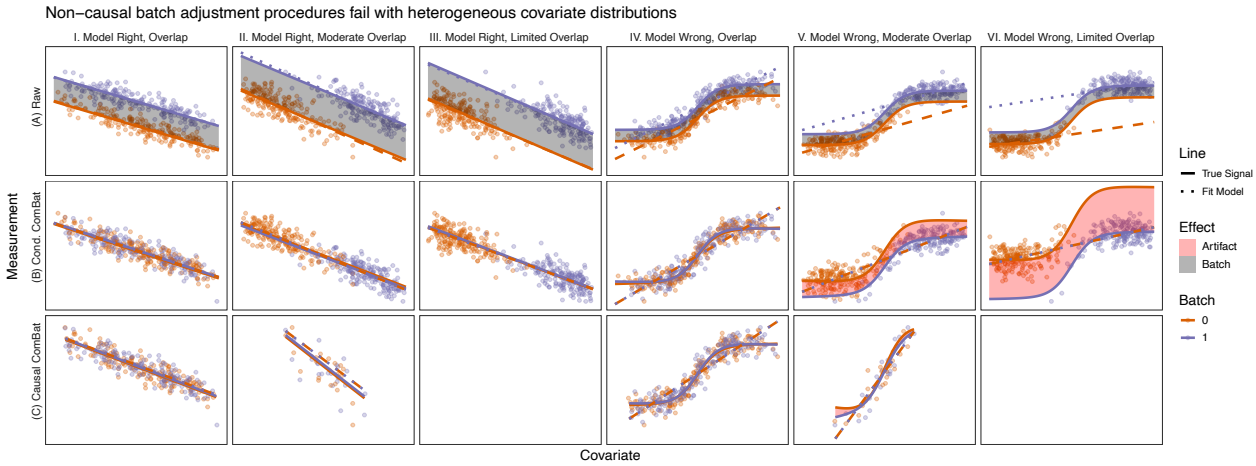

Figure 7: **Empirical simulations exhibit the shortcomings of Conditional** `ComBat`. $n = 400$ simulated univariate measurements are collected from a gaussian additive model (points) with equal probability of being from batch 0 or batch 1 (color). The batch effect (the difference between the two batches) is indicated by the translucent gray box. The expected measurement for a given batch at an indicated covariate value is indicated by the solid lines (the "true signal"). The signal assumed by the model employed is indicated by the dotted line (the "assumed signal"). Columns I to III show examples where the assumptions of `ComBat` are correct, and the fit model closely resembles the true signal in each batch across the covariate range (row A). Consequently, when the batch effect is removed (row B), the signal is consistent between the two batches. On the other hand, columns IV to VI show examples where the assumptions of `ComBat` are not correct. The fit model, a line for each batch, does not resemble the true signal, which is sigmoidal. When the conditions for covariate overlap are met as in IV, the batch effect can still be successfully removed. When the conditions for covariate overlap are not met as in VI, such as by Conditional `ComBat`, attempts to adjust for batch effects introduce a batch-specific artifact that dwarfs the initial batch effect by a factor of approximately 3 (red box). On the other hand, causal approaches facilitate robustness to covariate misspecifications, as in row (C). When demographic overlap is high, the batch effect is mitigated, and when demographic overlap is only moderate, Causal `ComBat` limits inference to the region of covariate overlap between the two batches, yielding a consistent estimate of the batch effect, whereas Conditional `ComBat` introduces a batch artifact of similar magnitude to the original batch effect (B.V). When there is limited demographic overlap (C.III and C.VI), Causal `ComBat` conservatively reports that there is insufficient data for inference. For further technical details of these simulations, see Appendix D.2.

Addressing the batch effect One way to characterize the batch effect would be analogous to the average treatment effect [38] at a particular covariate level (x, z); e.g.:

$$\text{ATE} = \text{ATE}(x, z) = \mathbb{E}_{f_{x,z}(y|1)}[Y_i] - \mathbb{E}_{f_{x,z}(y|0)}[Y_i]$$

For all simulations, it is clear that the ATE is 1 and does not depend on the covariates.

By definition, a difference in the expectations implies that $f_{x,z}(y|1) \neq f_{x,z}(y|0)$, and a batch effect is present by Definition 1.

It is well understood in the statistical literature that for models of the form $Y_i = g(X_i) + \beta T_i + Z_i$ where $Z_i \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$, that as long as either the strong ignorability condition [38] or f is known, the adjustment factor $\hat{\beta}_1 - \hat{\beta}_0$ will be a consistent and unbiased estimate of the ATE [38, 49]. For a non-technical explanation in words, we can obtain reasonable estimates of the ATE in one of two ways (and consequently, account for the batch effect):

1. Extrapolation: if we know $g(x)$, then we can "guess" what $\mathbb{E}_{f_{x,z}(y|t)}[Y_i]$ looks like for values of x that do not exist in our batch, by appropriate estimation of g via the elements of our dataset that we *do* have. In this case, where the model is linear, extrapolation can be performed by linear regression. Once we have identified $g(x)$, we can extrapolate what $g(x)$ looks like for values of x which fall outside the range of our dataset, as in Simulation II.

2. Use asymptotics: if we do not know $g(x)$ but we can reasonably assume that the strong ignorability condition holds, we can leverage asymptotic statistics to conclude that we can still adequately estimate the ATE. Proofs of this result can be found in [38, 39]. In this particular case, $X_i \perp\!\!\!\perp T_i$,

which implies the strong ignorability condition [38]. For more trivial reading specific to the case where $X_i \perp\!\!\!\perp T_i$, see [49]. This result is exhibited in Simulation III.
By a similar argument, the estimated batch offset terms, $\hat{\beta}_1$ for batch 1 and $\hat{\beta}_0$ for batch 0, can also be consistently estimated in either of these contexts. In these cases, removal of the estimated offset associated with the particular batch of a measurement is a reasonable approach to removal of the estimated batch effect.

Simulation IV, in which the assumption about $g(x)$ is *incorrect* and X_i is distributed conditionally on T_i (and consequently, the strong ignorability condition will not hold), will not necessarily yield a reasonable estimate of the ATE. In this case, due to the sigmoidal construction and the manner in which Batches 0 and 1 were balanced, we obtain a biased estimate of the ATE, which yields the batch artifact shown in red.

On a technical note, with $a_{x,y}(y|t)$ the distribution of the measurement Y_i after batch adjustment, we say a *residual artifact* is present if a batch effect *still* exists; e.g., if for two batches t and t':

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

Appendix E. Overlap of Empirical Covariates.

The empirical overlap of the covariate distributions is difficult to compute in the case of data without making heavy parametric assumptions. For this reason, we turn to the distribution-free overlapping index [32]. The distribution free overlapping index, $\hat{v}_{kl}$, 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:

$$\hat{v}_{kl} = \sum_{c \in \mathcal{C}} \min\Big(\hat{\mathbb{P}}(C = c|D = k), \hat{\mathbb{P}}(C = c|D = l)\Big) \sum_{s \in \{\text{M,F}\}} \min\Big(\hat{\mathbb{P}}(S = s|C = c, D = k), \hat{\mathbb{P}}(S = s|C = c, D = l)\Big) \Bigg[\int_{\mathcal{A}} \min\Big(\hat{f}(a|S = s, C = c, D = k), \hat{f}(a|S = s, C = c, D = l)\Big)\, \mathrm{d}a \Bigg]$$

which is obtained via numerical quadrature.

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.

Batch-Related Effects are reduced with Causal `ComBat` Here we investigate batch-related effects before and after correction. Figure 10 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 F.

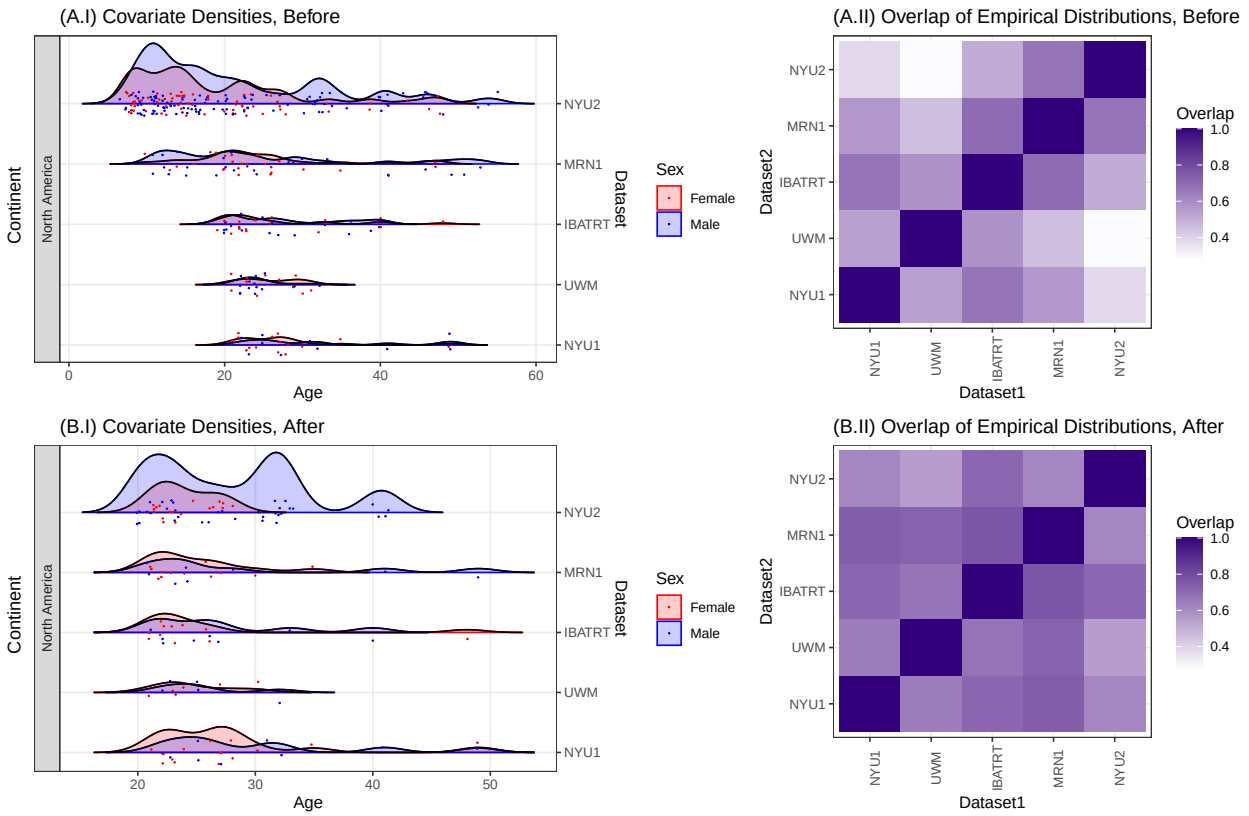

Figure 8: **The overlap of the empirical covariate distributions for the American Clique, before and after adjustment**. **(A)** The empirical distributions of the covariates before adjustment. **(B)** The empirical overlap of the covariate distributions after the adjustment procedure is applied, as discussed in Appendix C.

For this reason, we do not consider Z-scoring going forward, and focus instead on variations of the ComBat procedure.

Appendix F. 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 11. The left column of Figure 11A 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 11A 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 11B, these two properties are examined statistically for the individuals from the American Clique (Mann-Whitney U Statistic, see Methods B.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 12 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

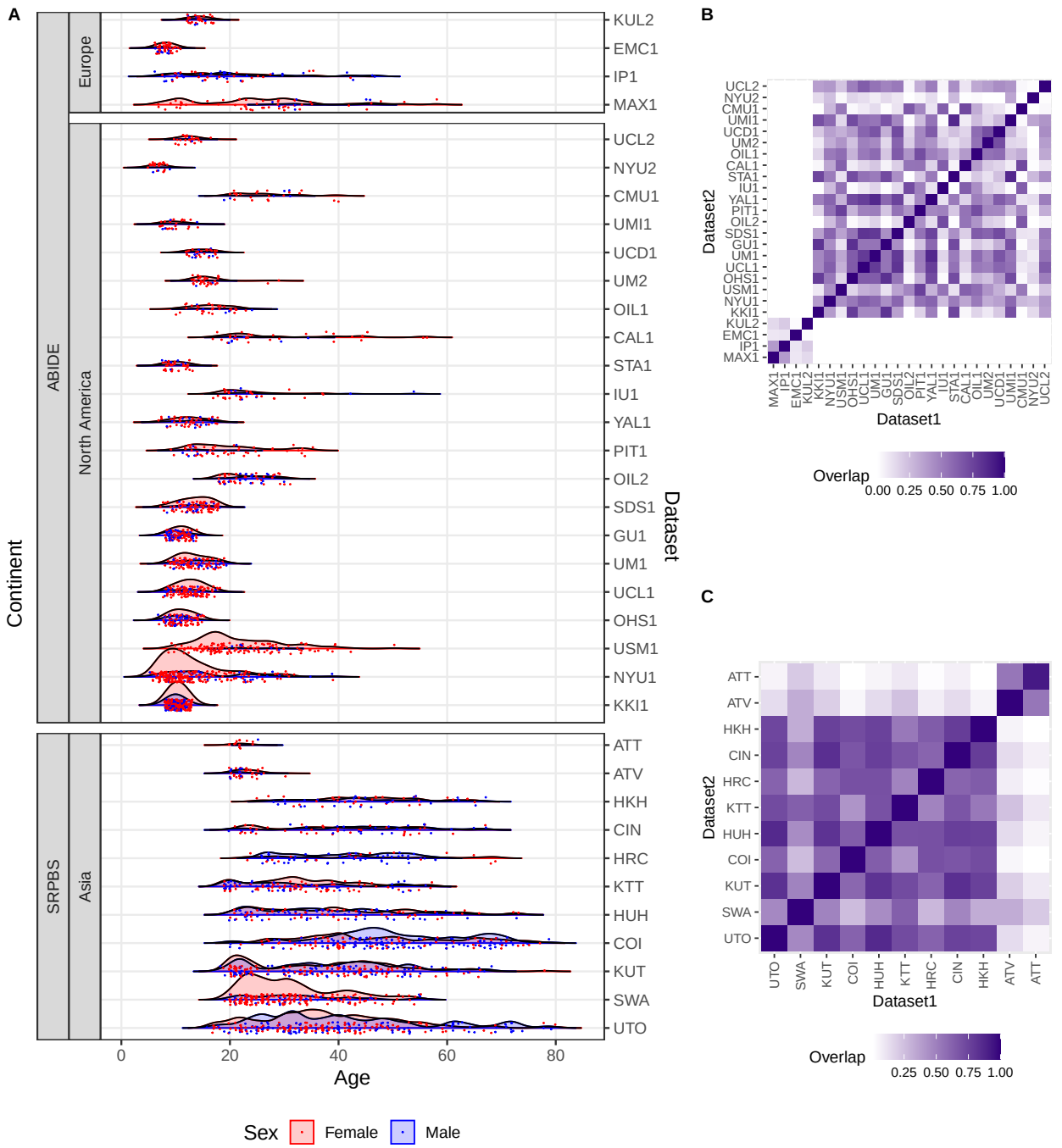

Figure 9: **The overlap of the empirical covariate distributions for two mega-studies. (A)** The empirical distribution of covariates. **(B)** The overlap of covariate distributions given by the distribution-free overlapping index for ABIDE mega-study [5, 6]. **(C)** The overlap of covariate distributions given by the distribution-free overlapping index for the SRPBS mega-study. [56]. Like for the CoRR mega-study, while several pairs of sites have overlapping demographic distributions, many of the sites have extremely poor overlap in both mega-studies. In these cases, attempts to normalize for batch effects using model-based approaches like Conditional `ComBat` would be subject to the pitfalls of Figure 7V or VI if modelling assumptions are not reasonable.

property that the homotopic edges have far higher connectivity than non-homotopic edges.

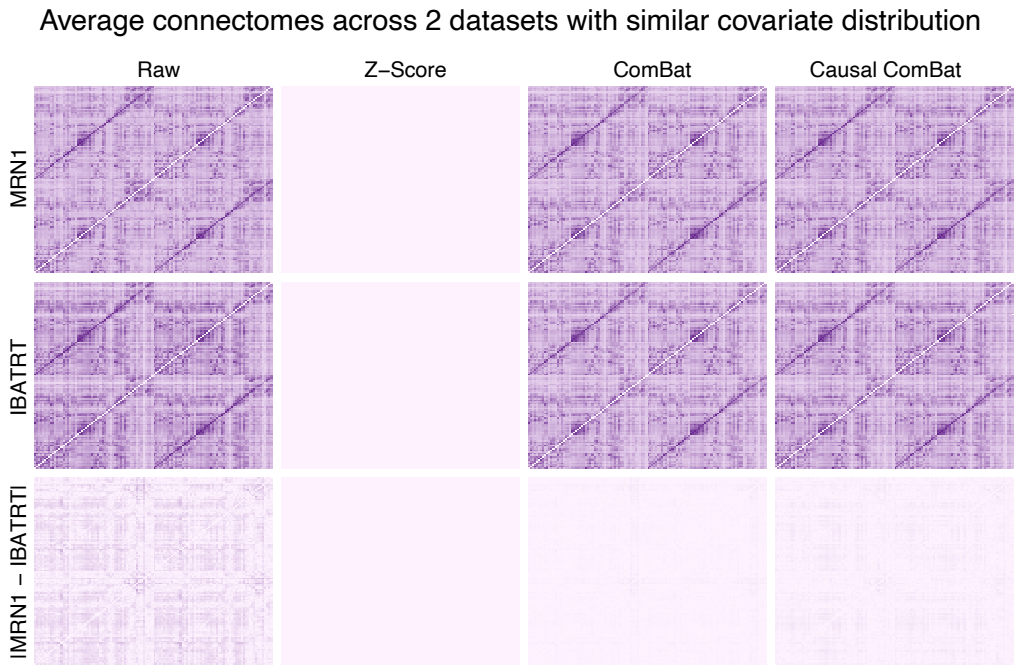

Figure 10: **Distribution of Detected Effects Before and After Correction**. A comparison of the average connectomes from two studies from the same continent, with similar sex and age distributions conditional on sex. While the *Raw* average connectomes appear similar, the absolute difference between the two is appreciable. All of the three adjustment strategies reduce the absolute difference between the average connectomes for the two batches.

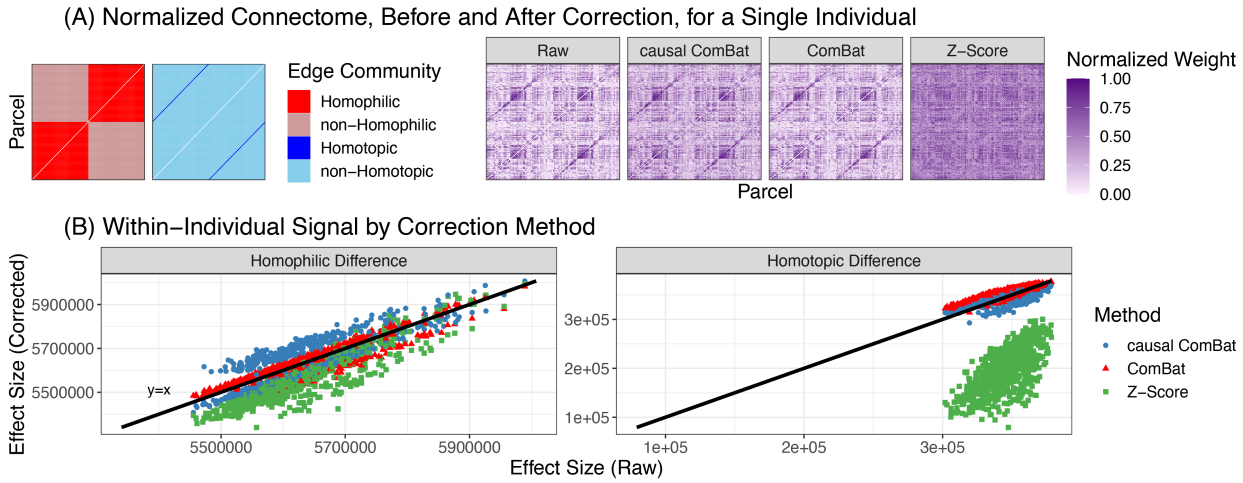

Figure 11: **Preservation of Topological Properties of Connectomes after Batch Correction**. The presence of topological effects is explored in the context of the homophilic and homotopic differences in connectivity in the American Clique. **(A)** (*Left*) A map indicating which edges are associated 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.

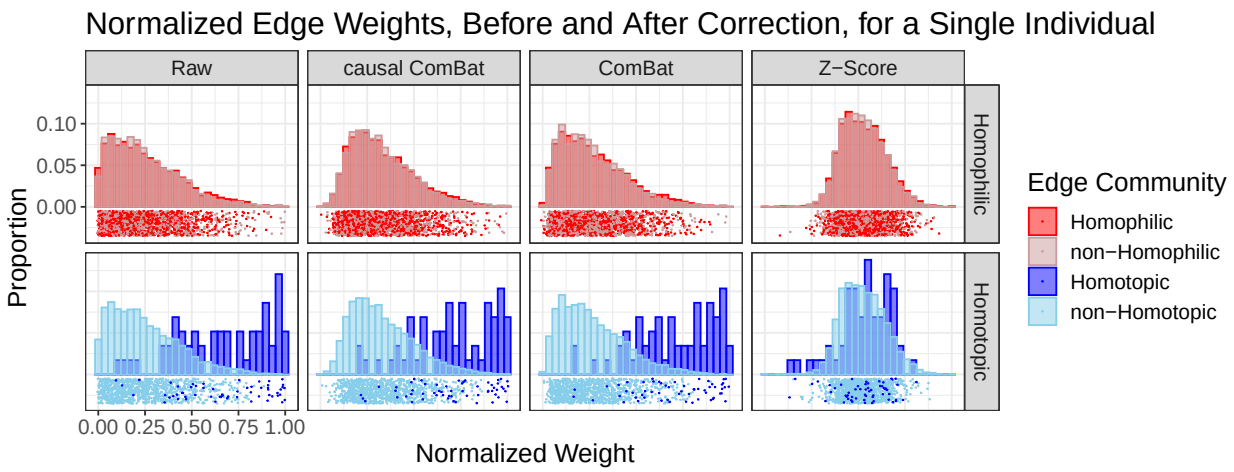

Figure 12: Histogram and rug plot indicating the empirical distribution of the normalized edge weights associated with each of the four edge communities for a single individual's connectome.