A Causal Perspective for Batch Effects: When is no answer better than a wrong answer?

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 significant challenges for scientific and clinical discoveries. Specifically, batch effects can (i) produce spurious signals and/or (ii) obscure genuine signals, contributing to the ongoing reproducibility crisis. Typically, batch effects are modeled as classical, rather than causal, statistical effects. This model choice renders the methods unable to differentiate between biological or experimental sources of variability, leading to unnecessary false positive and negative effect detections and overconfidence. We formalize batch effects as causal effects to address these concerns, and augment existing batch effect detection and correction approaches with causal machinery. Simulations illustrate that our causal approaches mitigate spurious findings and reveal otherwise obscured signals as compared to non-causal approaches. Applying our causal methods to a large neuroimaging mega-study reveals instances where prior art confidently asserts that the data do not support the presence of batch effects when we expect to detect them. On the other hand, our causal methods correctly discern that there exists irreducible confounding in the data, so it is unclear whether differences are due to batches or not. This work therefore provides a framework for understanding the potential capabilities and limitations of analysis of multi-site data using causal machinery.

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 across different sites, requiring post-hoc data aggregation approaches. These "mega-studies" comprise numerous individual studies, rendering samples sizes substantially larger than any individual study and addressing the small-sample size woes associated with modern big biomedical data [1].

Unfortunately, aggregating data across diverse datasets introduces a source of undesirable variability known as a batch effect. Lazar et al. [2] provides a recent consensus definition of batch effect: "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 can be detected or mitigated.

Broadly, many approaches model the batch collection or measurement process as a nuisance variable [3–9]. The implicit model justifying these approaches assumes batch effects are associational or conditional, but not causal. Such assumptions are strong, potentially unjustified, and often inappropriate. Two of the most prominent examples of these techniques are `ComBat` and Conditional `ComBat` (`cComBat`) [4]. These approaches have demonstrated empirical utility in various genomics and neuroimaging contexts [9, 10]; however, it remains unclear when these approaches will be successful and when they will fail. Specifically, it is unclear when they remove biofidelic variability or fail to remove nuisance variability [11].

In this work, we develop a causal approach to define, detect, estimate, and mitigate batch effects. Our main conceptual advance is modeling batch effects as causal effects rather than associational or conditional effects. Given this structure, we introduce a formal definition of causal batch effects. This formal definition reveals the limitations of (typically inappropriate) assumptions implicit in existing approaches [12–14] and provides a theoretical explanations for many limitations of batch harmonization noted in Bayer et al. [11].

Methodologically, we introduce a simple pre-processing strategy that one can apply to existing techniques for batch effect detection and mitigation. First, to detect and estimate batch effects, we introduce `Causal cDcorr` [15, 16]—building on modern, non-parametric statistical methods—to es-

[1] Johns Hopkins University [2] Child Mind Institute, [3] Yale University, Northeastern University, [†] Corresponding author: Eric W. Bridgeford (ebridge2@jhu.edu), 3400 N. Charles Street Clark Hall, Suite 200. Baltimore, MD, USA 21218.

timate and detect the presence of batch effects. Second, to mitigate batch effects, we introduce `Causal cComBat`—an augmentation of the `ComBat` procedure [4]—to remove batch effects while limiting the removal of veridical biological variation.

We apply these methods to simulations and a large neuroimaging mega-study assembled by the Consortium for Reliability and Reproducibility (CoRR) [17], consisting of more than $1{,}700$ individuals across 27 disparate studies. Our simulations demonstrate that existing strategies can, under many realistic use-cases, either remove more apparent veridical biological signal, or fail to mitigate the batch effect entirely, than *Causal* `cComBat`. Our real-data investigations reveal that previous strategies fail to differentiate between veridical signal and sources of variability due to other experimental design factors.

2 A causal framework to model batch effects

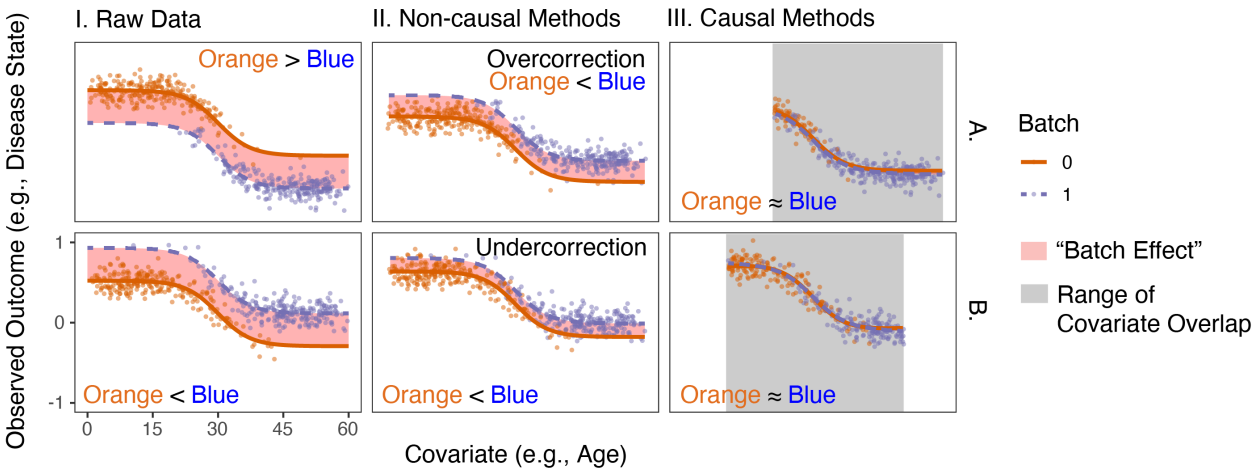

Figure 1: **Non-causal batch effect mitigation procedures are subject to both over and under correction and cannot rectify "confounding." Our causally enriched methods address these issues. I.** shows the observed data (points), where color indicates batch. The orange line and the blue line indicated the expected outcome per-batch, and the "batch effect" describes the observed difference (red band) between the expected outcomes. Ideally, after batch effect correction, the blue and orange lines should overlap. The orange batch tends to oversample people with younger ages, and the blue batch tends to oversample people with higher ages. **(A)** A scenario where the covariate distributions are moderately confounded and partially overlap, and the orange batch tends to see higher outcomes than the blue batch. **(B)** A scenario where the covariate distributions are moderately confounded and partially overlap, and the orange batch tends to see lower outcomes than the blue batch. **II.** and **III.** indicate the corrected data after correction, via non-causal and causal methods respectively. If the batch effect is removed, the orange and blue lines should be approximately equal. Non-causal methods attempt to adjust for the batch effect over the entire covariate range, and in so doing, are subject to confounding biases. Causal methods instead look to a reduced covariate range (gray box), finding points between the two datasets that are "similar," and are not subject to these issues.

A conceptual illustration of the value of causal modeling of batch effects Consider a simple example in Figure 1. We observe $n = 300$ measurements in two batches, where one batch (orange) tends to sample younger people, and the other batch (blue) tends to sample older people. The observed outcome of interest is disease state (y-axis), and there is a single potential confounder measured: age (x-axis). The solid lines indicate the true distribution governing the relationship between age and disease for both batches, which is unknown in practice. That the data-generating distributions differ indicates that a batch effect is present (red band). Techniques are desired to remove the batch effect given only the data measurements (outcome and covariate pairs, indicated as points). The two rows show two different setings: the top shows a case where the orange batch tends to be larger than the blue, and the bottom shows the reverse.

Figure 1A illustrates an example with a moderate amount of covariate overlap, meaning that data from the two batches have similar (but not identical) age ranges. Non-causal strategies learn from

each batch, and then *extrapolate* trends across covariates (in this case, age) to infer a relationship between the two batches. The problem is that this approach is strongly sensitive to the specifics of the extrapolation. Because the true data-generating distribution is unknown at the time of analysis, most non-causal approaches [4, 9, 18, 19] perform a linear extrapolation. When covariate overlap is imperfect, however, this misspecification can be disastrous (Figure 1A.II). While before correction the true data-generating distribution for the orange batch is higher than the blue batch, after correction the blue batch is higher the orange batch (the batch effect was *over-corrected*; i.e., too much correction was applied). Reversing the relationship between the blue and orange batches in Figure 1B, non-causal strategies are still unable to properly remove the batch effect and may still over- or under-correct for batch effects. As a result, "batch-effect-corrected data" from non-causal strategies may not actually be corrected, and in many situations may be more different after "correction" than before.

On the other hand in Figure 1III.A or Figure 1III.B, causal techniques focus instead on deriving conclusions within a range of covariate overlap in the data, where there is no confounding. The true data-generating distributions (and the points themselves) almost perfectly align after batch effect correction across the range of overlapping covariates, indicating that the batch effect has been removed, despite no foreknowledge of the true data-generating model.

In this manuscript, we posit that we desire that methods report confounding when it is present, something only causal methods do. Motivated by this perspective, we augment traditional approaches with causal machinery for batch effect detection and correction. Theory, simulations, and real data analysis demonstrate that traditional strategies to detect or correct batch effects from mega-studies lack the ability to identify confounding, and therefore, often add or remove batch effects inappropriately when covariate overlap is low. Therefore, such approaches cannot be trusted without further analysis. These issues highlight the primary challenges of performing valid statistical inference while pooling data across studies. This work therefore contributes to the ongoing effort to improve the validity and reliability of inferences in past and future mega-studies.

2.1 Models motivating different batch effect techniques Here we build up the implicit modeling assumption underlying various approaches to mitigating batch effects. Our working example is that the exposure is batch, and the outcome is connectome. We want to know whether the differences between the batches are due to variability in the participants themselves who are measured, or veridical differences in outcomes across the two batches. Figure 2 shows four different models and provides details of which kinds of variables are relevant in neuroimaging studies, though the concepts apply much more generally. Directionality of arrows (e.g., $T \to Y$) indicates that the variable T can influence the variable Y. Therefore, our goal is to identify a regime in which a strategy can provide evidence of a batch effect; that is, the degree to which the exposure influences the outcome (i.e., a *causal* effect).

We begin with the simplest approach, an "Associational" model (Figure 2A). Using an associational model enables us to obtain valid answers to: "are differences in outcome associated with differences in exposure?" The answer to this question is often important, for example, in biomarker studies, but may be insufficient for determining causality. `DCORR`[20] and `ComBat`[4] are methods designed to test and correct for associational effects. For example, if the batch and the outcome are dependent upon any covariates, which they often are in real studies, then an estimated associational effect is not a valid estimate of a batch effect (Figure 2**(A)**). For instance, if the two batches differ because they have sampled individuals in different age groups, and the outcome is a function of age, then using this approach will lead to falsely concluding causal batch effects. On the other hand, if a batch effect is truly present, but somehow another covariate cancels out this variability, then this approach could falsely conclude there does not exist a batch effect when it is present.

The "Conditional" model (Figure 2B) enables us to seek valid answers to a more nuanced question: "are differences in outcome associated with differences in exposure, after conditioning on measured covariates (Figure 2B)?" If there are differences in outcome after conditioning, this can suggest that

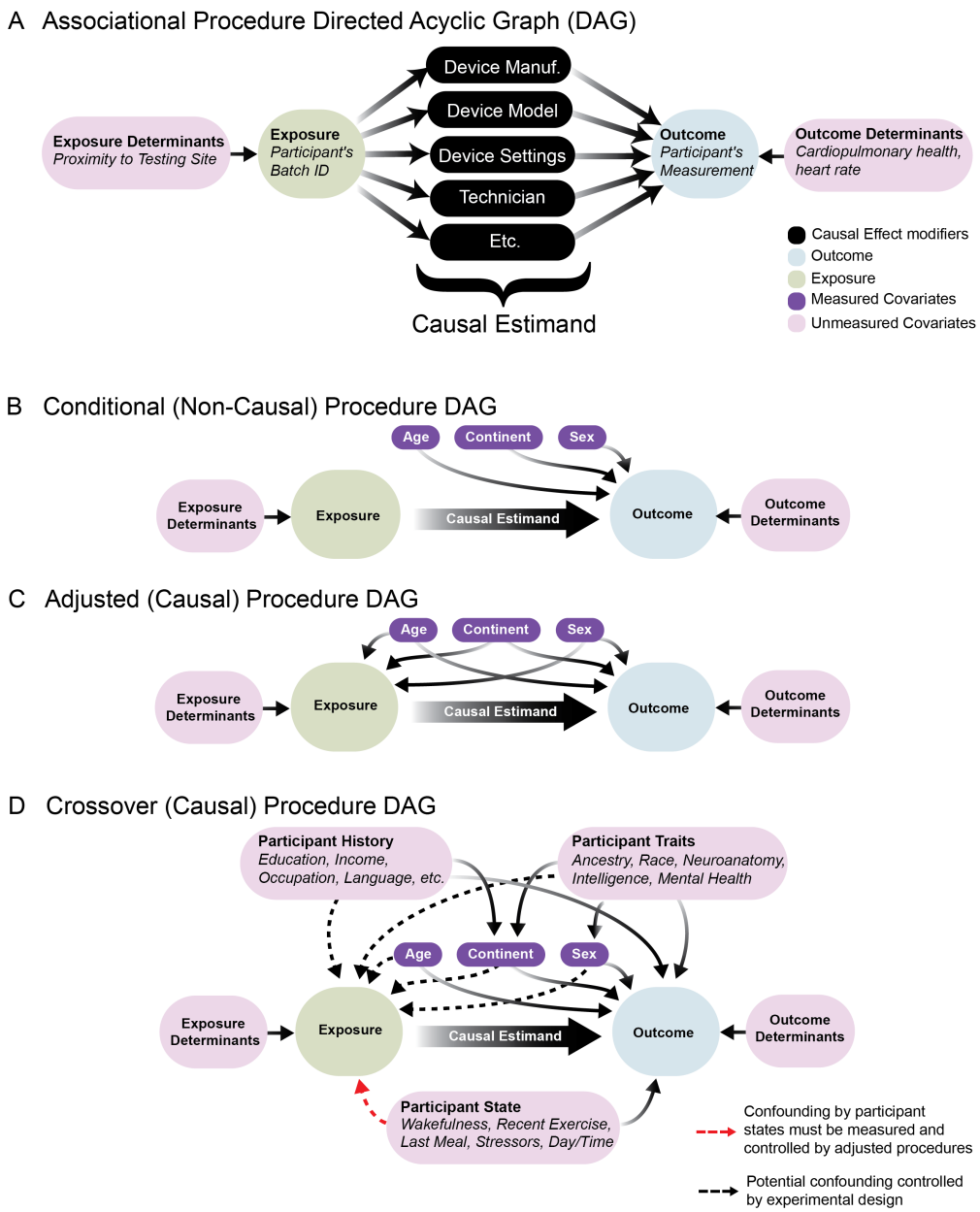

Figure 2: **Causal Graph of Study Covariates**. Causal Directed Acyclic Graphs (DAG) illustrating the underlying assumptions under which various procedures to detect or correct for batch effects are reasonable. Boxes represent variables of interest, and arrows indicate "cause and effect" relationships between variables. The causal estimand is the impact of the exposure (the batch) on the outcome (the participant's measurement) and is a cumulative effect of effect modifiers (black variables) both known and unknown that yield batch-specific differences. Variables are *confounding* if they affect both the exposure and the outcome. **(A)** Associational procedures and **(B)** conditional procedures are reasonable when measured or unmeasured covariates are non-confounding. **(C)** Adjusted causal procedures are reasonable when confounding variables are measured. **(D)** Crossover procedures are reasonable under many forms of potential confounding, both measured or unmeasured, so long as participant states are non-confounding. If participant states are confounding, the confounding states must be measured (red arrow).

the differences are due to different exposures. `cDCORR`[16] and `cComBat`[4] are methods designed to test and correct for conditional effects. However, this model is subject to errors in the presence of confounding variables. The covariates might also impact the exposure, and if so, also impact our ability to identify a causal estimand, but without our knowledge. Imagine, for instance, that the two

batches have covariates that partially overlap; for example, one sampled adolescents and adults, but not geriatrics, and the other sampled adults and geriatrics, but not adolescents. Conditional approaches make implicit assumptions about how to interpolate outcomes across covariate distributions potentially unobserved in a given batch. When those assumptions do not reflect the real data, the results can be erroneous. In this light, the identified effect may be a veridical batch effect, or an artificial covariate effect, thereby leading to either false positives or false negatives.

The "Adjusted" model (Figure 2C) enables us to mitigate this concern. Here, we seek valid answers to the following question: "conditioned on any impact of measured covariates on either exposure or outcome, do we see any residual differences in outcomes associated with differences in exposure?" This is effectively the model in most observational studies, which are a critical component of causal inference [14]. Our proposed `Causal cDcorr` and `Causal cComBat` (described below) are methods designed to test and correct for data following this model [15]. For example, to address the above concern, an adjusted model might discard all the adolescents and geriatrics, and simply focus on modeling the adults, for which we have data in both batches. While this reduces the sample size, it avoids the pitfalls of confidently concluding the presence or absence of a batch effect, even when it is not clear. However, it also suffers from a potential confounding issue. While it addresses the *observed* covariates, it can fail to address the *unobserved* covariates. Specifically, if there exist unobserved covariates that do not correlate with the observed ones and yet impact the outcomes, exposures, or both, then adjusted methods may lead to spurious results.

The "Cross-over" model (Figure 2D) addresses this concern. In a cross-over investigation, each participant is measured across *all* exposure groups. When properly performed, cross-over models enable us to seek valid answers to questions such as: "conditioned on all potential confounders, are differences in outcome associated with differences in exposure?" When the experimental design is a cross-over study and states are unchanging or randomized, even `Dcorr` and `ComBat` are adequate. If participant states change, then confounding remains unless the order participants are measured in each batch is randomized or participant states are measured and suitably adjusted for, as-per the above adjusted model.

Appendix A provides rigorous definitions and mathematical models contextualizing the above definitions.

2.2 Detecting and mitigating batch effects At present, literature approaches for detecting and mitigating batch effects tend to (implicitly or explicitly) model the effects as associational or conditional, which often fail to adequately account for confounding biases. To this end, we propose a simple technique to augment classical strategies (Figure 3, see Appendix B for details):

1. Use classical causal procedures to re-weight the measurements from the batches so that measured covariate distributions across all batches are overlapping. To re-weight, we use vector matching [21] for batch effect detection, and propensity trimming followed by nearest neighbor matching [14, 22] for batch effect correction.
2. Apply traditional procedures for detection or correction of batch effects post re-weighting. We use `cDcorr` for batch effect detection [15, 16], and `cComBat` for batch effect correction [3–9].

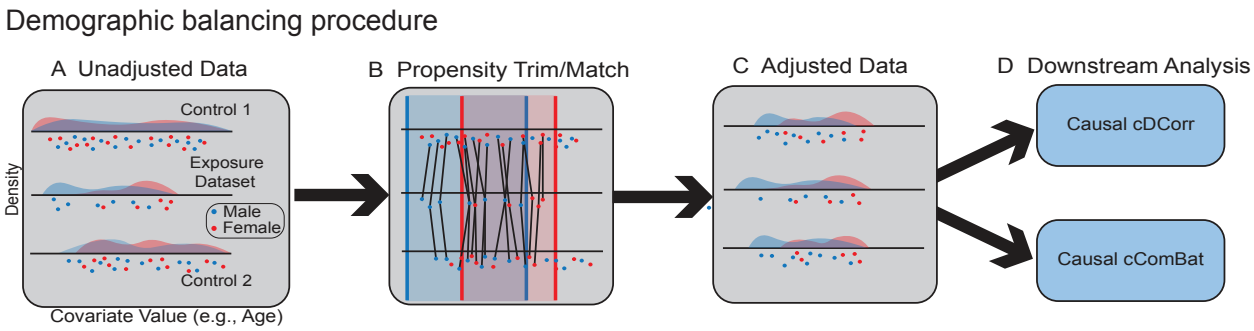

Figure 3: **The demographic balancing procedure serves to demographically align poorly balanced datasets using** CAUSAL CDCORR **or** `Causal cComBat`**. (A)** The unadjusted datasets are imbalanced in covariate distributions. The smallest dataset, the *reference*, has a narrower age distribution than the controls. **(B)** Propensity trimming (shaded boxes) and matching (lines) are conducted to identify reasonable matches between the reference and the control datasets. **(C)** The adjusted datasets have far improved demographic overlap after propensity trimming and matching. **(D)** Downstream analysis for batch effect correction or detection are applied via `Causal cComBat` or CAUSAL CDCORR.

2.3 Causal machinery helps mitigate limitations of traditional batch effect methods Here we quantify the performance of causal and non-causal methods generalizing the simulations from Figure 1. Figure 4A shows multiple simulation settings, including linear (left), nonlinear (center), and non-monotonic (right). The standard measure of effect size in the causal literature is called the 'Average Treatment Effect' (ATE) [12, 13]. The red bands in Figure 4A show the treatment effect for all possible values of the covariate, so the ATE is just the average width of that bar, which happens to be 0.5 here. Since the treatment here is batch, the goal is to remove the treatment effect, which requires correctly estimating the treatment effect and then removing it. We consider three different algorithms for this purpose: `cComBat`, `cComBatGAM`, and `Causal cComBat`. We show the estimates of the absolute ATE for each method and each simulation setting as we vary the amount of covariate overlap (top row, Figure 4C). When the relationship between covariate and outcome is linear, all the methods work regardless of the amount of overlap (Figure 4C.i). However, when the relationship is non-linear, the non-causal methods over-estimate the batch effect unless there is a nearly perfect overlap (Figure 4C.ii). In contrast, `Causal cComBat` frequently reports that there exists irreconcilable confounding when overlap is low (dotted red line), and correctly estimates and removes the ATE when overlap is high. The results are qualitatively similar when the relationship is non-monotonic.

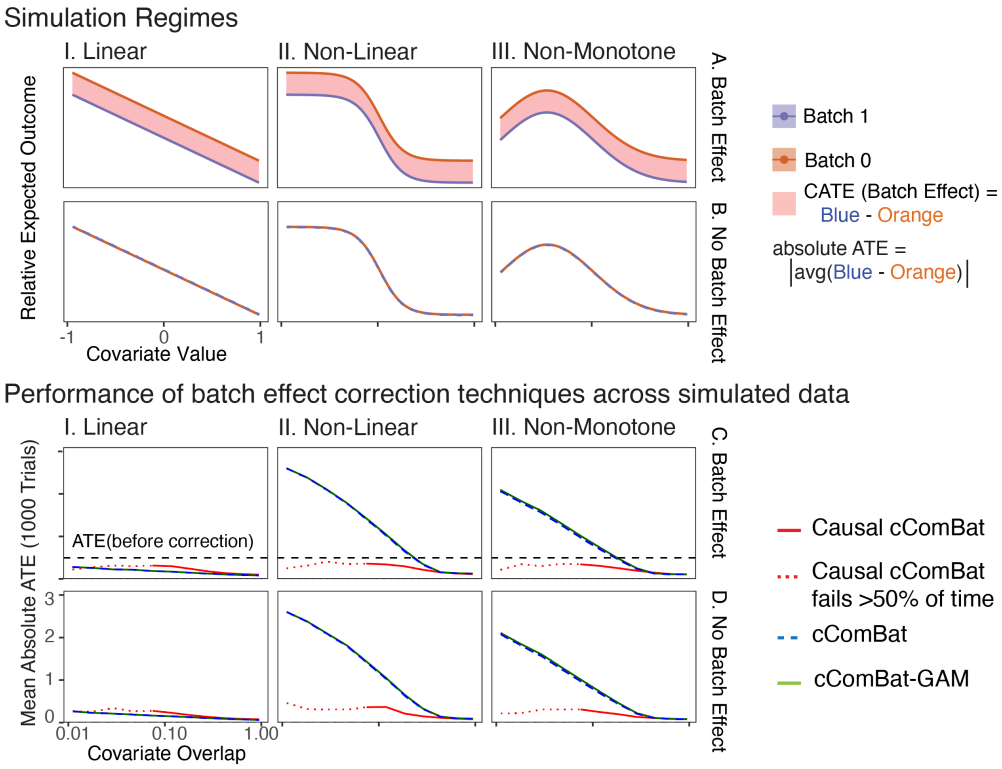

Figure 4: **Six simulation regimes illustrate that non-causal procedures are subject to strong biases without covariate matching**. **(A)** and **(B)** plots illustrating the relationship between the relative expected outcome and the covariate value, for each batch (color), across **(I.)** linear, **(II.)** non-linear, and **(III.)** non-monotone regimes. The conditional average treatment effect (red box) highlights the batch effect for each covariate value. The average treatment effect (ATE) is the average width of this box. We measure across regimes both with and without batch effects *a priori* (average treatment effect of 0.5 in **A** and 0 in **B**). **(C)** and **(D)** illustrate the effectiveness of different batch effect correction techniques. Techniques with high performance will have a mean absolute ATE after correction at or near 0 (there is no batch effect). Across all regimes, `Causal cComBat` struggles to execute more than 50% of the time when covariate overlap is insufficient (approximately 0.1). Across linear regimes, all techniques are effective for batch effect removal, regardless of covariate overlap. In both non-linear and non-monotone regimes, both `cComBat` and `cComBatGAM` fail to remove batch effects when present (top row) and introduce batch effects when none are present *a priori* (bottom row). Note in particular that the magnitude of the absolute ATE after correction exceeds the magnitude of the ATE before any correction was applied at all (0.5), indicating that the data for each batch are *more dissimilar* after correction than before correction when using `cComBat` or `cComBatGAM`.

Figure 4B shows the same simulations, but now there is no batch effect whatsoever. All the methods correctly estimate that there is no batch effect in the linear setting (Figure 4D.i). However, non-causal methods incorrectly estimate the presence of a batch effect for both the nonlinear and non-monotonic settings, except when the covariates are nearly perfectly overlapping (Figure 4D.ii and Figure 4D.iii). In these regimes, non-causal techniques tend to introduce a batch effect, when none is present *a priori*. The causal methods behave better here, too, reporting irreconcilable confounding when overlap is low, and correctly estimating the absence of a batch effect (thereby avoiding the introduction of a batch effect) when overlap is high.

3 Application of causal methods to a large neuroimaging multi-study

3.1 The CoRR Studies have disparate demographic characteristics

The motivation for our work is the neuroimaging mega-study produced by the Consortium for Reliability and Reproducibility [17], a collection of over $3{,}600$ functional neuroimaging measurements from over $1{,}700$ individuals spanning 27 separate datasets. Figure 5A 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 [23–25], age [26–28], and continent (as a surrogate for race and culture) [29, 30]

are variables that have been associated with brain connectivity, they serve as measured demographic covariates used in our investigation.

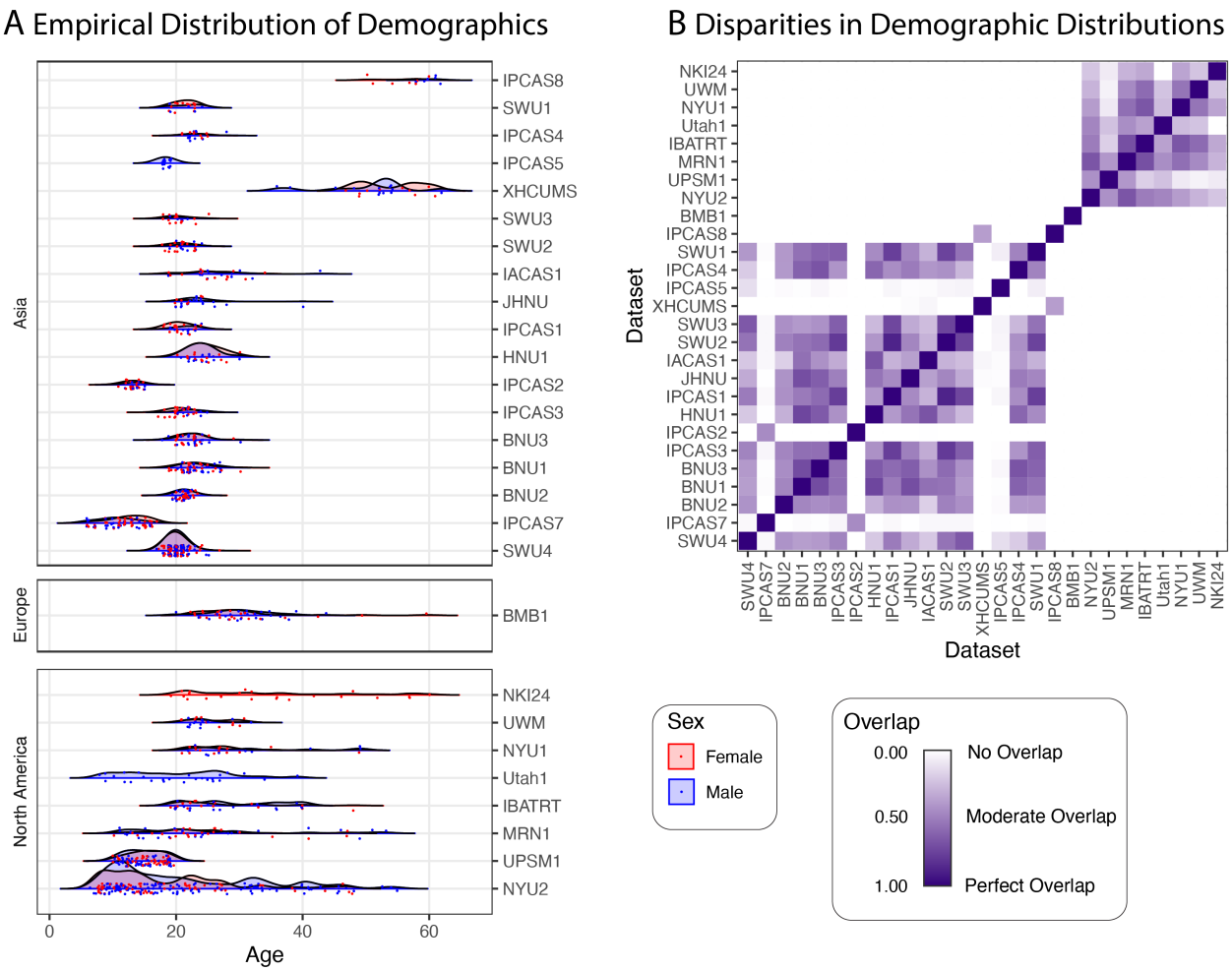

Figure 5: **Demographic data for the 27 studies from the CoRR Mega-Study**. **(A)** Each point represents the age of a participant corresponding to a single measurement. Rows are studies, boxes are continents, and color indicates sex. **(B)** Even with only 3 observed covariates (sex, age, and continent of measurement), the CoRR studies often show extremely limited covariate overlap [31]. This makes inference regarding batch effects difficult.

Figure 5B illustrates the level of demographic overlap in the CoRR mega-study, using a distribution-free overlapping index [31] (see Appendix E for details). The CoRR mega-study includes many pairs of datasets with varying degrees of overlap, from high (near 1) to low (near 0). Further, many of the datasets do not overlap at all (overlap of 0), making inference about batch effects impossible without making strong assumptions. This is particularly troublesome, as the covariate records for the CoRR mega-study common to all sites include only three covariates: age, sex, and continent of measurement. Additional measured covariates can only reduce the estimated overlap between the pairs of datasets, so having poor overlap on such a sparse set of covariates indicates that the actual demographic overlap is likely even lower.

3.2 Detecting Batch Effects in the CoRR mega-study Figure 6A focuses on discerning the viability of different types of effects one could use to test for batch effects from Section 2.1. For each pair of datasets in the CoRR study, we test whether (or not) a discernable effect is present, while controlling for age, sex, and continent differences between the individuals within each dataset. Intuitively, we would like to believe that if such a test rejects the null hypothesis in favor of the alternative, the data

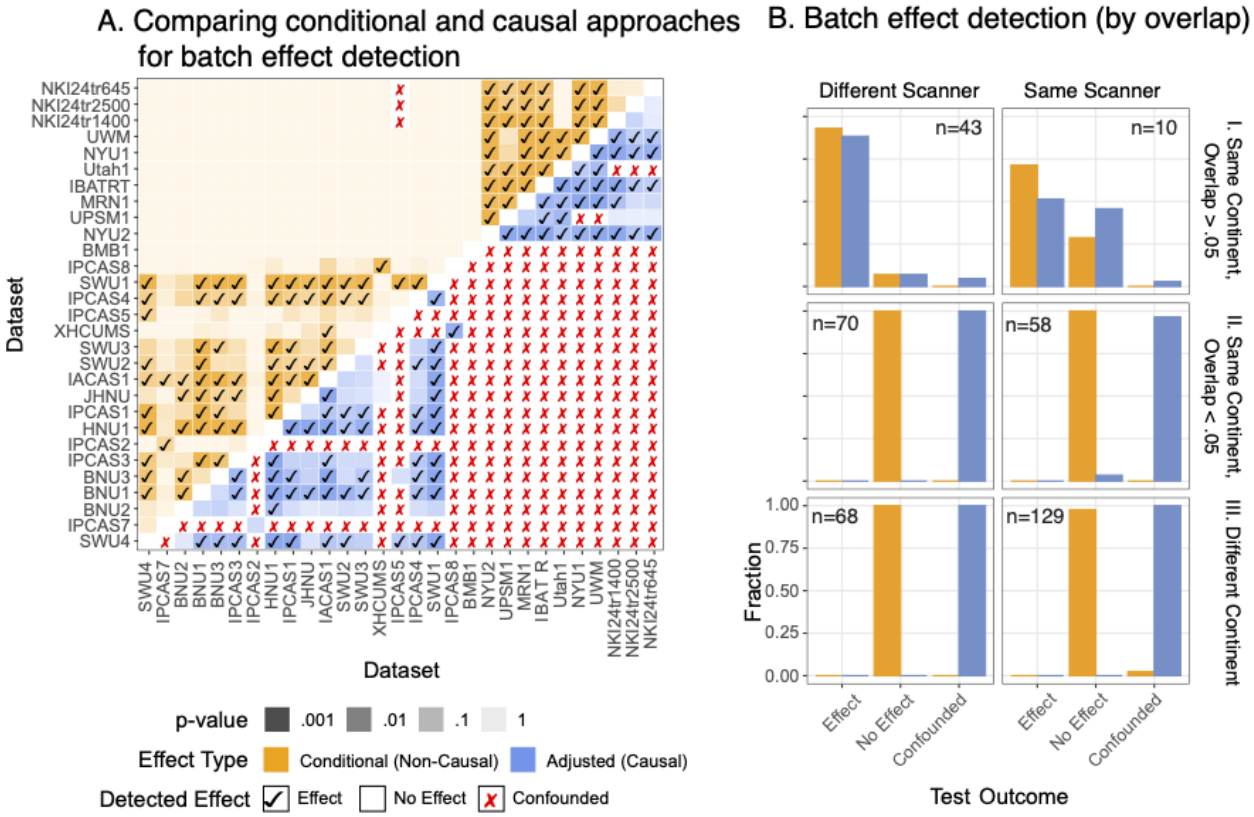

Figure 6: **Comparison of types of effects between datasets from the CoRR study**. **(A)** Heatmap of different types of effects (conditional and adjusted causal procedures) that can be used to detect differences between each pair of datasets in the CoRR study. Whereas most (25.8%) non-confounded conditional effects are not significant, most (66.1%) non-confounded adjusted effects are significant. **(B).I** delineation of how one possible source of batch effects, scanner model, impacts significance rates of batch effects. Almost all pairs of studies conducted on different scanners with high covariate overlap ($> .05$) have discernable batch effects. The rate of detected effects is lower when the scanner model is the same. **(B).II** and **(B).III** When the level of estimated covariate overlap is lower ($< .05$) or zero (different continent), conditional effects never detect a difference across datasets. On the other hand, adjusted causal procedures instead report that the data is too confounded for subsequent inference and avoid running entirely.

supports that a batch effect is present. When we account for demographic covariates using conditional (non-causal) approaches (orange squares, top left), discrepancies between the datasets are detected 25.8% of the time (checkboxes, cDCORR, BH correction, $\alpha = 0.05$). As explained in Section 2.1, the viability of these approaches as tests for batch effects is the assumption that the measured covariates are overlapping and non-confounding, which is overwhelmingly false for many of these comparisons, as shown in Figure 5B. Similar tests are performed using adjusted (causal) approaches in 6A (blue squares, bottom right). Many pairs of studies could not have an adjusted (causal) effect estimated due to poor demographic alignment (282 of 406 pairs of studies). Notably, adjusted (causal) procedures can be used to estimate effects between all pairs of the American Clique (a group of separate studies consisting of similar demographics collected in North America, consisting of NYU1, NYU2, IBATRT, UWM, MRN1). After adjustment for the sample characteristics of the demographics associated with individual studies, a majority of adjusted (causal) effects (66.1%) remain significant (CAUSAL cDCORR, BH correction, $\alpha = 0.05$).

Note that confounding (red "X"s) in Figure 6A for causal effects extremely closely align with datasets with low covariate overlap in Figure 5B. A possible source of a batch effect could be different types of scanners were used to collect the data [4, 9, 18], which is represented in our presentation as a causal effect modifier in Figure 2A. To this end, we aggregate the fraction of comparisons which report the

indicated test outcome for different levels of covariate overlap (rows) when the scanner model is the same (right column) or different (left column). When the scanners are different and the covariates overlap, both causal and non-causal methods reliably detect batch effects a majority of the time, and they do so at a higher rate than when the scanner is the same (Figure 6B.I). When the covariate overlap is low (Figure 6B.II), or when the continent of measurement is different entirely (and the overlap is zero, (Figure 6B.III), conditional (non-causal) procedures almost always fail to reject. On the other hand, adjusted (causal) procedures overwhelmingly report that a given pair of studies are demographically confounded, and are too different for any effect to be investigated.

Appendix Figure 14 delineates similar plots for the other two effect types (associational and causal crossover effects).

3.3 Traditional Approaches Produce Disparate Inference from Techniques Leveraging Causality We investigate the strength of the sex effect, conditional on age, in the connectomes before and after batch effect correction (Figure 7A). `ComBat` and `cComBat` are performed on the entire CoRR study (top row) with age and sex covariates, and subsequently all techniques are restricted to the subset of connectomes upon which `Causal cComBat` was executed (the *American Clique*, bottom row). The edges which show a significant sex effect conditional on age (partial distance correlation, BH correction, $\alpha = .05$) are colored from smallest test statistic to largest test statistic (rank = 1, dark purple), and edges without a significant conditional sex effect are colored white. 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 "X" indicates the analysis is not possible on the basis of continent, age, and sex). Using non-causal strategies for batch effect removal prior to restricting analysis to the American Clique produces substantially different inference than restricting analysis to the matched samples, as inference would suggest no sex effects versus many sex effects (bottom row).

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 matched American Clique using only `Causal cComBat` strategies. This investigation stems from the observation that the heatmaps appear, at face value, to be approximately similar in terms of the number of significant edges, and the visual patterns observed between the three left-most plots of the top row and the right-most plot of the bottom row. We compare the DICE overlap [32, 33] of the top 100 edges (by effect size) from each of the three non-causal approaches applied to all the data with those of `Causal cComBat` on the American Clique. A high DICE overlap indicates that the subsets of edges are similar. Using the raw connectomes, `ComBat`, and `cComBat` tend to produce similar inferences. In contrast, they all have low overlap with `Causal cComBat` (Figure 7C). This suggests that causal and non-causal strategies for batch effect correction produce somewhat orthogonal subsequent inference. The results are qualitatively similar when considering `cComBatGAM`, in Appendix Section G.

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 cComBat` and `Causal cDCORR`) yield empirically valid approaches to handling batch effects, or report that no inference can be made due to insufficient covariate overlap. This represents a principled approach over alternative strategies, which we show may introduce spurious biases and dramatically under-estimate or over-estimate batch effects. We illustrated how, ignorant of potential confounders, otherwise principled strategies may misrepresent batch effects. This illustrates that causal adjustment procedures

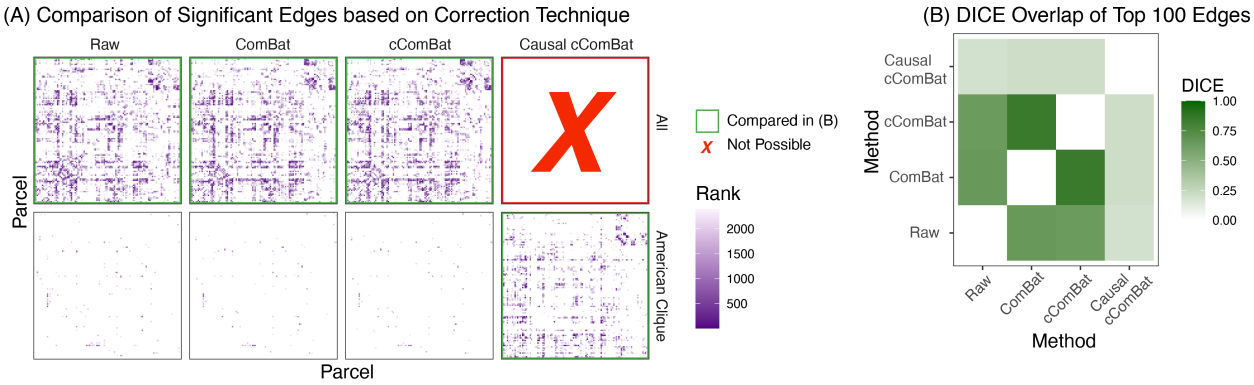

Figure 7: **Significant Edges Before and After Batch Effect Removal**. **(A)** The presence of a sex effect (conditional on individual age) is investigated for each edge in the connectome. The top 100 significant edges are shown in rank order from largest (rank = 1) to smallest sex effect ($\alpha = .05$, Benjamini Hochberg [34] 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 cComBat` 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 100 edges for all pairs of correction techniques indicated with a green square from **(A)**. Overlap of the top signal edges is high between no correction (raw), `ComBat`, and `cComBat`, but all techniques have low overlap with `Causal cComBat`, indicating that causal approaches yield measurements with signal distinct from that of non-causal approaches.

can augment hypothesis testing by indicating when the data may be unsuitable to provide answers to a particular question due to irreconcilable confounding, rather than permitting false positives or false negatives.

Finally, our neuroimaging use-case shows the utility of the methods developed herein for estimating and removing batch-related effects from large biological datasets. We demonstrate that a large number of non-detected effects in the CoRR study may, in fact, yield false conclusions on the basis of the fact that perhaps the level of confounding between the datasets is too great to yield any conclusion *at all*. We showed that, when demographic overlap was high and the scanner differed, batch effects were nearly always detectable using conditional (non-causal) or adjusted (causal) procedures. On the other hand, when demographic overlap is lower, even if the scanner still differs, effects are entirely undetectable by conditional (non-causal) procedures. In a situation where we expect batch effects (different scanners), we go from almost always detecting effects to never detecting effects simply by looking at pairs of comparisons with less demographic overlap. On the other hand, adjusted (causal) procedures do not suffer from this limitation, and explicitly delineate when the datasets are overly confounded by demographic covariates.

The final question of our analysis was: does it matter? Even if we disrupt or misrepresent veridical signals in the dataset, can we still detect them, irrespective of whether the signal is maintained in its unadulterated form? Our answer to this question is yes, as our methods identify different downstream demographic effects as compared to prior approaches, each of which identifies features similar to a conditional approach, which are well established in the causal literature to permit numerous false positives and false negatives [35, 36]. Together, our work rigorously defines and studies measurement bias decoupled from demographic confounders in complex, non-Euclidean data.

Limitations We do *not* claim that the specific choices for re-weighting measurements are 'optimal' in any sense. 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 contained herein and in our related manuscript [15] elucidate that even these simple balancing approaches are sufficient to highlight substantial shortcomings in the existing approaches for detecting or correcting batch effects, and that causal machinery mitigates these

shortcomings.

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 4(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 in regions of the covariate space in which the data *are able to support analysis*. This means that, although our conclusions may not apply to the *entire* region of covariate support in the observed data, they will be *statistically and methodologically valid* with respect to the re-weighted population (albeit, in a possibly narrower set of covariate values) that we *did* analyze.

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.

Another limitation of our proposed methods is that their adequacy to overcome confounding presupposes the measurement of a rich set of subject-specific covariates, which may often not be the case (and, likely, is not in the CoRR mega study). In this way, our real data investigation serves as illustrative of the wide disparities that can be found when leveraging causal techniques. On the other hand, with sufficient subject-specific covariates (e.g., more than age, biological sex, and continent as a loose proxy for race and culture) we are still likely to run into confounding issues which can be ameliorated by causal techniques such as those proposed herein.

Future Work With these limitations in mind, 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 [17, 37–39], in which potentially demographically disparate (or diagnostically) 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 5, and in the

ABIDE study [37, 38] detailed in Appendix Figure 13. While the SRBPS [39] has greater overlap on the basis of age, sex, and likely race (same continent) for many pairs of datasets (but still has two datasets that are dissimilar from the others), both ABIDE and SRBPS add the additional hurdle of neurodivergences being used in participant recruitment to batches. If the connectome is a proxy for brain function, neurodivergent brain features captured by the connectome may cause symptoms of neurodivergence [40–42], introducing a potential causal mediation effect if the neurodivergence is leveraged directly in participant recruitment [43, 44]. Intuitively, this can be conceptualized as participant illness status creating a *cycle* in a system-level directed graph (such as Figure 2), which yields a violation of the ignorability assumption that is critical for conditional or causal augmentations to yield stable estimates of batch effects [15]. While this problem has been defined for the wider problem of data fusion [45], is unclear the extent to which causal mediation impacts batch effect detection or correction, nor the suitability of alternative strategies, in high-dimensional biological studies.

Future work will focus on applying the methods developed here to correct for batch effects in mega-studies such as the Adolescent Brain Cognitive Development (ABCD) [46], which includes $N > 12{,}000$ demographically diverse individuals from across the United States, and for which 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. [47] 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. [47] and [39], 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 11) can be employed to mitigate the batch effects in other non-demographic-aligned datasets. We believe that a failure to attain a comprehensive understanding of batch effects and the extent to which they pervade statistical inferences 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.

Contributions EWB wrote the manuscript; EWB and JTV revised the manuscript; EWB, JTV, and MP conducted study conception and design; EWB, JTV, BC, and MP conducted interpretation of results; EWB, JC, SP, and RL analyzed raw neuroimaging and demographic data; EWB and JTV devised the statistical methods; EWB performed statistical analyses; MP, GK, SN, TX, MM, BC, and JTV provided valuable feedback and discussion throughout the work.

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 1RF1MH123233-01, 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. Scott Marek, Brenden Tervo-Clemmens, Finnegan J. Calabro, David F. Montez, Benjamin P. Kay, Alexander S. Hatoum, Meghan Rose Donohue, William Foran, Ryland L. Miller, Timothy J. Hen-

drickson, Stephen M. Malone, Sridhar Kandala, Eric Feczko, Oscar Miranda-Dominguez, Alice M. Graham, Eric A. Earl, Anders J. Perrone, Michaela Cordova, Olivia Doyle, Lucille A. Moore, Gregory M. Conan, Johnny Uriarte, Kathy Snider, Benjamin J. Lynch, James C. Wilgenbusch, Thomas Pengo, Angela Tam, Jianzhong Chen, Dillan J. Newbold, Annie Zheng, Nicole A. Seider, Andrew N. Van, Athanasia Metoki, Roselyne J. Chauvin, Timothy O. Laumann, Deanna J. Greene, Steven E. Petersen, Hugh Garavan, Wesley K. Thompson, Thomas E. Nichols, B. T. Thomas Yeo, Deanna M. Barch, Beatriz Luna, Damien A. Fair, and Nico U. F. Dosenbach. Reproducible brain-wide association studies require thousands of individuals. Nature, 603:654–660, March 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04492-9.

2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. Johanna M. M. Bayer, Paul M. Thompson, Christopher R. K. Ching, Mengting Liu, Andrew Chen, Alana C. Panzenhagen, Neda Jahanshad, Andre Marquand, Lianne Schmaal, and Philipp G. Sämann. Site effects how-to and when: An overview of retrospective techniques to accommodate site effects in multi-site neuroimaging analyses. Frontiers in Neurology, 13:923988, October 2022. ISSN 1664-2295. doi: 10.3389/fneur.2022.923988.
12. 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.

13. 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.
14. 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.
15. Eric W. Bridgeford, Jaewon Chung, Brian Gilbert, Sambit Panda, Adam Li, Cencheng Shen, Alexandra Badea, Brian Caffo, and Joshua T. Vogelstein. Learning sources of variability from high-dimensional observational studies. ArXiv e-prints, July 2023. doi: 10.48550/arXiv.2307.13868.
16. Xueqin Wang, Wenliang Pan, Wenhao Hu, Yuan Tian, and Heping Zhang. Conditional Distance Correlation. J. Am. Stat. Assoc., 110(512):1726, 2015. doi: 10.1080/01621459.2014.993081.
17. 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.
18. Andrew A. Chen, Joanne C. Beer, Nicholas J. Tustison, Philip A. Cook, Russell T. Shinohara, Haochang Shou, and Alzheimer's Disease Neuroimaging Initiative. Mitigating site effects in covariance for machine learning in neuroimaging data. Human Brain Mapping, 43(4):1179–1195, March 2022. ISSN 1097-0193. doi: 10.1002/hbm.25688.
19. 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.
20. 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.
21. Michael J. Lopez and R. Gutman. Matching to estimate the causal effects from multiple treatments, 2014. URL https://www.semanticscholar.org/paper/Matching-to-estimate-the-causal-effects-from-Lopez-Gutman/11ca8c8fa81498fe75349cff0001552c89115c8c. [Online; accessed 9. Feb. 2023].
22. 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.
23. 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.
24. 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.
25. 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.
26. 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.
27. 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.
28. 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.
29. 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.
30. Jianqiao Ge, Guoyuan Yang, Meizhen Han, Sizhong Zhou, Weiwei Men, Lang Qin, Bingjiang Lyu, Hai Li, Haobo Wang, Hengyi Rao, Zaixu Cui, Hesheng Liu, Xi-Nian Zuo, and Jia-Hong Gao. Increasing diversity in connectomics with the Chinese Human Connectome Project. Nature Neuroscience, 26(1):163–172, January 2023. ISSN 1546-1726. doi: 10.1038/s41593-022-01215-1.
31. 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.
32. Lee R. Dice. Measures of the Amount of Ecologic Association Between Species. Ecology, 26(3): 297–302, July 1945. ISSN 0012-9658. doi: 10.2307/1932409.
33. Thorvald Sørensen. A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity o Munksgaard in Komm., 1948. URL https://books.google.com/books/about/A_Method_of_Establishing_Groups_of_Equal.html?id=rpS8GAAACAAJ.
34. 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.
35. 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.
36. 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.
37. 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.
38. 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.
39. 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.
40. Joshua T. Vogelstein, Eric W. Bridgeford, Benjamin D. Pedigo, Jaewon Chung, Keith Levin, Brett Mensh, and Carey E. Priebe. Connectal coding: discovering the structures linking cognitive phenotypes to individual histories. Current Opinion in Neurobiology, 55:199–212, April 2019. ISSN 0959-4388. doi: 10.1016/j.conb.2019.04.005.
41. Kerstin Konrad and Simon B. Eickhoff. Is the ADHD brain wired differently? A review on structural and functional connectivity in attention deficit hyperactivity disorder. Human Brain Mapping, 31(6): 904, June 2010. doi: 10.1002/hbm.21058.
42. Sheema Hashem, Sabah Nisar, Ajaz A. Bhat, Santosh Kumar Yadav, Muhammad Waqar Azeem, Puneet Bagga, Khalid Fakhro, Ravinder Reddy, Michael P. Frenneaux, and Mohammad Haris. Genetics of structural and functional brain changes in autism spectrum disorder. Translational Psychiatry, 10, 2020. doi: 10.1038/s41398-020-00921-3.
43. Judea Pearl. Interpretation and identification of causal mediation. Psychological Methods, 19(4): 459–481, December 2014. ISSN 1939-1463. doi: 10.1037/a0036434.
44. Elias Bareinboim and Judea Pearl. Controlling Selection Bias in Causal Inference. In Artificial Intelligence and Statistics, pages 100–108. PMLR, March 2012. URL https://proceedings.mlr.press/v22/bareinboim12.html.
45. Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. Proceedings of the National Academy of Sciences, 113(27):7345–7352, July 2016. doi: 10.1073/pnas.1510507113.
46. 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.
47. 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.
48. Stephen R. Cole and Constantine E. Frangakis. The Consistency Statement in Causal Inference: A Definition or an Assumption? Epidemiology, 20(1):3, January 2009. ISSN 1044-3983. doi: 10.1097/EDE.0b013e31818ef366.
49. 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/.
50. 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.
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. Bradley Efron. Large-Scale Simultaneous Hypothesis Testing. J. Am. Stat. Assoc., 99(465):96–104, Mar 2004. ISSN 0162-1459. doi: 10.1198/016214504000000089.
53. 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.
54. 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.
55. 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.
56. 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.
57. 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.
58. 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.
59. 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.
60. 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.
61. 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.
62. 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.
63. Judea Pearl. Causal inference in statistics: An overview. ssu, 3(none):96–146, Jan 2009. ISSN 1935-7516. doi: 10.1214/09-SS057.
64. A. W. van der Vaart. Asymptotic Statistics. Cambridge University Press, Cambridge, England, UK, October 1998. ISBN 978-0-51180225-6. doi: 10.1017/CBO9780511802256.

Appendix A. Methods (Expanded). For a formal description of the below definitions, see Bridgeford et al. [15]. Informally, we adopt the following notation to describe batch effect correction. We assume that Y_i represents an observed measurement of interest (e.g., a connectome), T_i represents the batch in which the measurement is collected, X_i represents observed covariates, (e.g., age or biological sex), and Z_i represents unobserved covariates (e.g., height). The data (Y_i, T_i, X_i, Z_i) are n independent and identical samples from some true and unknown data generating distribution. The quantity $Y_i^{(t)}$ represents the measurement that would have been observed, had the measurement been collected in a given batch t. The key distinction is that Y_i represents the *actual* measurement, which is collected and studied. On the other hand, $Y_i^{(t)}$ is a potentially hypothetical measurement, which is only actually observed for the case where $T_i = t$ in most experimental contexts [48]. That is, individuals have a potential measurement $Y_i^{(t)}$ for every possible batch, but only a single of these potential measurements are studied under standard observational contexts. This is known as the *consistency assumption*, and can be written mathematically as:

$$Y_i = \sum_{t \in \mathcal{T}} \mathbf{1}_{\{T_i = t\}} Y_i^{(t)} = Y_i^{(T_i)},$$

where $\mathcal{T}$ is the set of batch labels. That the observed measurements Y_i given the batch (or measured covariates) and the potential measurements $Y_i^{(t)}$ do not, in general, have similar distributions is known as the "fundamental problem of causal inference." This problem requires resorting to additional assumptions to make conclusions on the basis of the observed data.

Using this convention, a batch effect is defined in Definition 1.

Definition 1 (Batch Effect). *A batch effect exists between two batches t and t' if $Y_i^{(t)}$ and $Y_i^{(t')}$ have different distributions.*

In light of this definition, a batch effect can be conceptualized as the *potential* measurements having different distributions between a given pair of batches. A batch effect is present if an individual's potential measurements (which are random quantities) would differ (in terms of their distributions) *by virtue of them being measured* in batches t and t'.

The major conceptual leap is that, under the manner in which most mega-studies are collected (where individuals are *unique* to each batch), Definition 1 is about *potential* measurements rather than *realized* measurements. Causal claims regarding potential measurements will only be valid on the basis of the observed (realized) data insofar as the reasonableness of the assumptions upon which they rest.

Associational Effects In an associational context, we observe measurements y_i and batches t_i for each individual $i \in [n]$, so effects can only be determined from realizations of Y_i and T_i.

Definition 2 (Associational Effect). *An associational effect exists between batches t and t' if $Y_i | T_i = t$ and $Y_i | T_i = t'$ have different distributions.*

Likewise, we can use this intuition to develop a suitable definition for associational batch effect correction:

Definition 3 (Associational Effect Correction). *Associational effect correction is a function g, where:*

$$g\left(\{(Y_i, T_i)\}_{i \in [n]}\right) = \left\{\tilde{Y}_i\right\}_{i \in [n]}$$

such that for all pairs of batches t and t', $\tilde{Y}_i | T_i = t$ and $\tilde{Y}_i | T_i = t'$ have the same distribution.

Whereas Y_i represents the measurement for an individual, the intuition is that $\tilde{Y}_i$ can be thought of as the "corrected" measurement, which does not have an associational effect for any pair of batches in the observed data.

Associational effects will often poorly characterize whether a batch effect is present. 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 $Y_i|T_i = t$ and $Y_i|T_i = t'$ 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 merely an associational effect. A sufficient condition for an associational effect to facilitate detecting or estimating a batch effect would be that individuals are randomized to each batch, in that individuals are randomly assigned to be measured in particular batches *pre-hoc*. Associational effect detection can be facilitated via DCORR [20], and associational effect correction can be facilitated via ComBat [4].

Conditional Effects In a conditional context, we observe measurements y_i, batches t_i, and covariates x_i for all individuals $i \in [n]$, so effects can be determined from realizations of Y_i, T_i, and X_i.

Definition 4 (Conditional Effect). *A conditional effect exists between batches t and t' if for some covariate x, $Y_i|T_i = t, X_i = x$ and $Y_i|T_i = t', X_i = x$ have different distributions.*

We can define conditional effect correction using this logic:

Definition 5 (Conditional effect correction). *Conditional effect correction is a function g, where:*

$$g\Big(\{(Y_i, T_i, X_i)\}_{i\in[n]}\Big) = \Big\{\tilde{Y}_i\Big\}_{i\in[n]}$$

such that for all pairs of batches t and t' and for all covariate levels x, $\tilde{Y}_i|T_i = t, X_i = x$ and $\tilde{Y}_i|T_i = t', X_i = x$ have the same distribution.

Conceptually, a conditional effect is equivalent to a batch effect if two conditions hold:

1. the measured covariates overlap in distribution between the batches (propensity overlap), and
2. the measured covariates provide all of the information regarding the mechanism about how people ended up in one batch versus the other (strong ignorability).

The former condition denotes that both batches must contain similar groups of people (in terms of measured covariates), and the latter condition specifies that the measured covariates X_i tell us all of the information needed to "exchange" measurements from one batch to the other. Borrowing the preceding example, even if we observe more young people in a batch t, we must still observe *some* young people in the other batch t'. In this sense, the measured covariates contain the information needed to "deconfound" disparities that might be batch effects or veridical effects due to upstream covariates. Therefore, when we make subsequent comparisons, we do not need to guess what people with similar covariates would have looked like in the other batch, and vice versa.

In this fashion, our comparisons can be thought of as locally (with respect to the covariates) exchanging a realized measurement Y_i in batch t with a realized measurement Y_i in batch t', where both individuals have similar covariates x. Intuitively, these comparisons can therefore be conceptualized as synthetically comparing $Y_i^{(t)}$ and $Y_i^{(t')}$ (the target estimand for establishing a batch effect) by using observed measurements Y_i from individuals who are similar on the basis of the covariates X_i between the two batches. Condition 1 ensures that we can make this intuitive step over the entire span of covariates in our batches. Conditional effect detection can be facilitated via cDCORR [16], and conditional effect correction can be facilitated via cComBat [4].

In practice, we never know whether the propensity distributions 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 propensity distributions. This motivates a third approach.

Adjusted Effects As before, we observe measurements y_i, batches t_i, and covariates x_i for all individuals $i \in [n]$, and we determine effects from realizations of Y_i, T_i, and X_i. Prior to assessing the equality of any distributions, however, we weight the observations such that the observed covariate distributions are rendered approximately overlapping.

Definition 6 (Adjusted Effect). *An adjusted effect exists between batches t and t' if after re-weighting samples such that $X_i|T_i = t$ and $X_i|T_i = t'$ are approximately overlapping (or, alternatively, approximately equal) in distribution, $Y_i|T_i = t, X_i = x$ and $Y_i|T_i = t', X_i = x$ have different distributions.*

We can similarly define adjusted effect correction as:

Definition 7 (Adjusted Effect Correction). *Assume that the samples are re-weighted via weights w_i (possibly 0 or 1) such that after re-weighting, $X_i|T_i = t$ and $X_i = t'$ are approximately overlapping (or alternatively, approximately equal) in distribution for all $w_i \neq 0$.*
Adjusted effect correction is a function g, where:

$$g\Big(\{(Y_i, T_i, X_i, w_i)\}_{i\in[n]}\Big) = \Big\{\tilde{Y}_i\Big\}_{i\in[n]}$$

such that for all pairs of batches t and t' and for all covariate levels x, $\tilde{Y}_i|T_i = t, X_i = x$ and $\tilde{Y}_i|T_i = t', X_i = x$ have the same distribution.

Adjusted effects, by default, satisfy the first criterion for a conditional effect to be a batch effect. This is because rendering measured covariate distributions approximately equal intuitively is a more strict criterion than simply ensuring that they approximately overlap. The reason that we believe that re-weighting to ensure approximate covariate distribution equality is desirable for effect correction, versus simply approximate covariate overlap, is discussed in Section 2.3 and Appendix D.2. Adjusted effect detection can be facilitated via `Causal cDcorr` [15], and adjusted effect correction can be facilitated with `Causal cComBat` (described in Section 2.2).

We still must satisfy the latter criterion for a conditional effect to be a batch effect; that is, given the measured covariates, we can ignore how people ended up in one batch versus the other. This assumption has the same interpretation as before.

Crossover Effects We observe measurements $y_i^{(t)}$ and covariates $x_i^{(t)}$, for all individuals $i \in [n]$ and for all batches $t \in \mathcal{T}$. In this case, we can make statements on the basis of $Y_i^{(t)}$ itself, because we actually observe outcomes for each possible batch. Therefore, we typically will not need to resort to local exchangeability (similar individuals/covariates across batches) as before, unless there are aspects of the individuals changing from one batch to another. *Crossover effects in general require the fewest assumptions to derive causal conclusions, since we directly observe all possible potential measurements for each individual.*

Definition 8 (Crossover Effect, constant states). *A crossover effect exists between batches t and t' if, given that $\left(X_i^{(t)}, Z_i^{(t)}\right)$ and $\left(X_i^{(t')}, Z_i^{(t')}\right)$ are sufficiently similar, $Y_i^{(t)}$ and $Y_i^{(t')}$ have the same distribution.*

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 (even if we did not measure these traits). However, *states* may differ as they may be functions of location or time. For example, if being measured impacts subsequent states, then a crossover effect may not be indicative of a batch effect without further assumptions and/or experimental design considerations (such as randomizing exposure order across participants, or resorting to adjusted effect strategies as before if these states are measured).

In the case where participant states are unchanging or are randomized, we can resort directly to associational strategies for subsequent detection or correction, and can make causal conclusions about batch effects with few additional considerations. In the case where participant states are changing and are not randomized but are measured, we can resort to adjusted strategies for adjusted effect detection or correction, via a crossover effect for non-constant states:

Definition 9 (Crossover Effect, non-constant states). *A crossover effect exists between batches t and t' if, after re-weighting samples such that $X_i^{(t)}$ and $X_i^{(t')}$ are approximately overlapping (or,*

alternatively, approximately equal) in distribution, $Y_i^{(t)}$ and $Y_i^{(t')}$ have the same distribution.

For this effect to be a true batch effect, we need to make the same assumptions as before; that is, that given the measured covariates (which include the changing states), we can ignore how people ended up in one batch versus the other.

A.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 [49]. 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 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 necessarily 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 [22].

A.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 require two-sample and conditional two-sample testing procedures. A natural test statistic for the two-sample testing procedure is the Distance Correlation [20], which is a non-parametric test for testing whether two variables are correlated. A simple augmentation of the distance correlation procedure [50, 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 conditional distance correlation [16], 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 conditional distance correlation, and the natural test statistic is $\texttt{cDCorr}(\mathbf{Y}, \vec{t}|\mathbf{X})$.

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 conditional distance correlation, and the natural test statistic is $\texttt{cDCorr}(\tilde{\mathbf{Y}}, \vec{\tilde{t}}|\tilde{\mathbf{X}})$.

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 conditional distance correlation, and the natural test statistic is $\texttt{cDCorr}(\mathbf{Y}, \vec{t}|\mathbf{X})$.

More Than Two Batches The above approaches generalize sufficiently to K batches using K-sample testing approaches [15]. 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 conditional 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 [52]. All p-values are estimated using $N = 1{,}000$ (detection, Figure 6) or $N = 10{,}000$ (correction, Figure 7) 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 [34].

A.3 Batch effect correction via z-scoring

We compare z-scoring per-feature within each batch [2] and traditional `ComBat` to Causal `ComBat`. For a given edge $y_i^{(t)}(k, l)$ (k, l) for an individual i in batch t, we define the z-score corrected edge-weight to be:

$$z_i^{(t)}(k, l) = \frac{y_i^{(t)}(k, l) - \bar{y}^{(t)}(k, l)}{s^{(t)}(k, l)},$$

where $\bar{y}^{(t)}(k, l)$ is the sample mean and $s^{(t)}(k, l)$ is the sample standard deviation for edge (k, l) in batch t. These strategies reduce known topological properties of human connectomes; particularly, homotopy, as shown in Figure 17 and Figure 16 and are therefore not considered for more careful analysis.

A.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 [53]. 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 $\texttt{pDCorr}(\vec{y}(k,l), \vec{s}|\vec{a})$.

A.5 Datasets The CoRR Mega-Study [17] is an aggregate dataset consisting of 27 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 [54], which provides a wrapper for the `CPAC` Pipeline [55]. fMRI scans for each individual are first processed to remove motion artifacts using `mcflirt` [56]. The fMRI scans are then registered to the corresponding individual's anatomical scan using FSL's boundary-based registration (BBR) via `epireg` [57]. A non-linear transformation from the individual's anatomical scan to the MNI152 [58] template is estimtaed using `FNIRT` [59]. Nuisance artifacts are removed by fitting the voxelwise timeseries to a regression model incorporating regressors for the Friston 24-parameter model [60], the top five principal components of the voxelwise timeseries in cerebrospinal fluid `aCompCor` [61], and a quadratic drift term. The adjusted voxelwise timeseries is downsampled to the regions of interest (ROIs) of the Automated Anatomical Labelling (AAL) parcellation [53] 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 A.1.

The NKI Rockland Sample [62] 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 B. Procedures for detecting and mitigating batch effects.

B.1 Detecting batch effects with CAUSAL cDCORR Many of the more direct types of detectable effects, such as associational and conditional effects, fail to adaequately account for confounding biases present in the data. We instead propose the use of CAUSAL cDCORR, in which a conditional K-sample test [16] is performed on samples with the same "range" of covariate values after propensity trimming via a strategy known as vector matching [21]. This strategy is the focus of a complementary theoretical manuscript, in which we illustrate the theoretical and empirical (via simulations) benefits of this technique over competing approaches for detecting causal effects between potential outcomes. This strategy maintains both high sensitivity and specificity under traditionally problematic data regimes (high-dimensionality, non-monotonicities, and non-linearities) in which other methods typically fail [15], making it a natural choice for causal discrepancy testing (of which "batch effect" detection, termed *causal unconditional discrepancy testing* in [15], as-defined herein is a special case).

Figure 5C illustrates visually the causal procedures employed for adjusting the batches. Rather than fully matching to produce the adjusted batches, samples are retained only such that they have approximately overlapping covariate distributions (propensity trimming, shaded boxes). A full schematic illustrating the use of vector matching is detailed in Appendix E of [15]. In the event that the batch assignment mechanism is ignorable given the covariates and that effects between datasets are in the same direction across all covariate levels, the adjusted effect detected by CAUSAL cDCORR is a causal batch effect, as proven in [15]. That the effects between datasets are in the same "direction" across all covariate levels can be best intuited via example. Consider the case where a batch effect exists, such that there is a signal disparity (a difference in the expected connectivity) in a particular edge of a connectome between two batches. If this signal disparity is a positive difference between batches 1 and 2 across all covariate levels (or, a negative difference across all covariate levels), CAUSAL cDCORR will detect a causal batch effect. In the event that the assignment mechanism is ignorable but that effects between datasets are not in the same direction across all covariate levels, the adjusted effect detected by CAUSAL cDCORR is a causal effect, but may reflect a *causal conditional discrepancy* (e.g., there are batch-specific differences, but they are isolated to particular covariate levels). In the event that the batch assignment mechanism is not ignorable, the effect may not reflect any causal effects (e.g., it may reflect unmeasured demographic differences between the batches).

B.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 A 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 [4] 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 [3, 63], Causal ComBat yields the removal of an internally valid causal effect and does not require extrapolation assumptions, unlike Conditional ComBat [12–14, 49]. 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 4. Appendix E depicts the impact on the empirical covariate distribution of the adjustment procedure.

Appendix C. Correction for Batch-Related Effects.

Appendix D. Simulations.

D.1 Batch Effect Detection Simulations Simulations illustrating the sensitivity (high testing power when a causal effect is present) and specificity (tests which do not falsely detect effects) of CAUSAL cDCORR for causal effect detection are in Bridgeford et al. [15].

D.2 Batch Effect Removal Simulations $n = 400$ points are sampled from Batch 0 or Batch 1 with probability 0.5; e.g. $T_i \overset{iid}{\sim} \text{Bern}(0.5)$.

Covariate distributions For each row, the covariate distributions are:

I. Overlap The covariate distribution is $X_i \overset{iid}{\sim} 2\text{Beta}(2, 2) - 1$.

II. Moderate Overlap The covariate distribution is:

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

III. Limited Overlap The covariate distribution is:

$$X_i|T_i = t \overset{d}{\sim} \begin{cases} 2\text{Beta}(2, 12) - 1 & t = 0 \\ 2\text{Beta}(12, 2) - 1 & t = 1 \end{cases}$$

Simulation contexts We investigate these in four simulation contexts, where for all simulations, $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, 1)$:

Non-linearity A sigmoidal relationship between the covariate and the outcome. The outcome is:

$$Y_i = 4\text{sigmoid}(8X_i) + T_i + \frac{1}{2}\epsilon_i,$$

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

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

The effectiveness of batch effect correction methods for the non-linearity setting are shown in Figure 4.

Symmetric non-monotonicity A gaussian non-monotonic relationship between the covariate and the outcome. The outcome is:

$$Y_i = \varphi\left(X_i, \mu = 0, \sigma = \frac{1}{2}\right) + T_i + \frac{1}{2}\epsilon_i,$$

where $\varphi(x, \mu, \sigma)$ is the probability density function of the Normal distribution:

$$\varphi(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\left(\frac{x - \mu}{\sigma}\right)^2\right\}$$

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

We call it "symmetric" because $\mu = 0$, which leads to the both the covariate/outcome relationship and the covariates themselves being symmetric about $x = 0$. The effectiveness of batch effect correction methods for the symmetric non-monotonicity setting are shown in Figure 8.

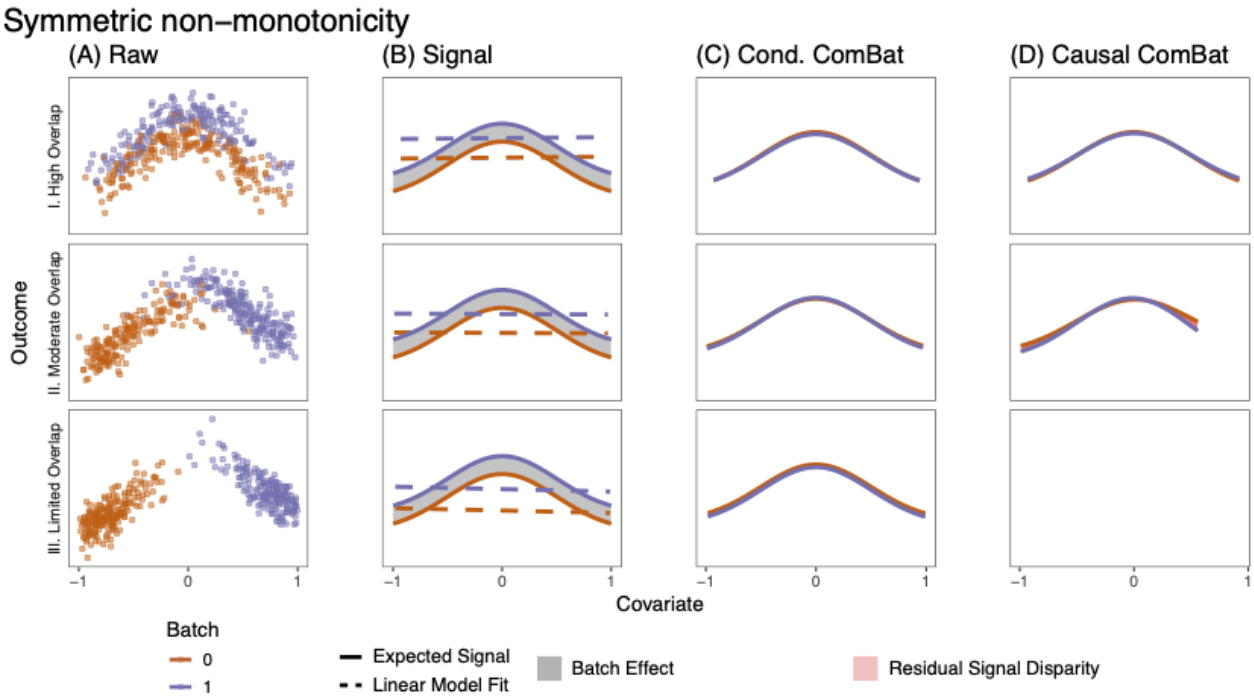

Figure 8: A symmetric non-monotonicity about $x = 0$. `cComBat` provides reasonable results when the non-monotonicity and the covariate distributions (for each batch) are symmetric about the same value (here, $x = 0$). In high balance the estimated ATE of `cComBat` and `Causal cComBat` are 0.076 and 0.051 (`Causal cComBat` better 67.6% of the time), in moderate balance are 0.179 and 0.173 (`Causal cComBat` better 50.9% of the time), and in limited overlap are 0.358 and 0.289 (`Causal cComBat` better 66.7% of the time) with `Causal cComBat` only executing in 36 simulations.

Asymmetric non-monotonicity A gaussian non-monotonic relationship between the covariate and the outcome. The outcome is:

$$Y_i = \varphi\left(X_i, \mu = -0.5, \sigma = \frac{1}{2}\right) + T_i + \frac{1}{2}\epsilon_i.$$

We call it "asymmetric" because $\mu = -0.5$, which leads to the effect not being symmetric about $x = 0$ (whereas the covariates are, by construction, symmetric about $x = 0$). The effectiveness of batch effect correction methods for the asymmetric non-monotonicity setting are shown in Figure 9.

Symmetric non-monotonicity, asymmetric covariates A gaussian non-monotonic relationship between the covariate and the outcome, and a different covariate distribution from the rest of the simulations. The covariate distribution is:

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

where $b = 2$, $b = 6$, and $b = 12$ for the overlap, moderate, and limited conditions, respectively.

The outcome is:

$$Y_i = \varphi\left(X_i, \mu = 0, \sigma = \frac{1}{2}\right) + T_i + \frac{1}{2}\epsilon_i.$$

This time, the non-monotonicity is symmetric, but the covariate distributions are increasingly asymmetric. The effectiveness of batch effect correction methods for the asymmetric non-monotonicity setting are shown in Figure 10.

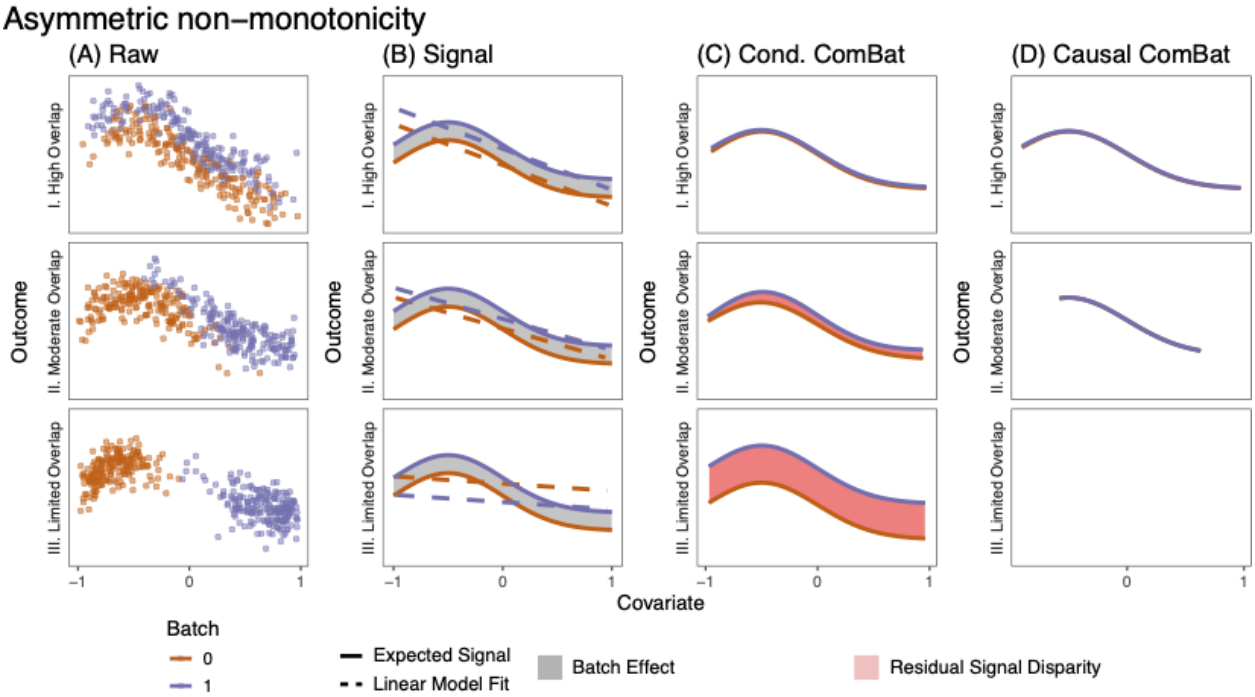

Figure 9: An asymmetric non-monotonicity in which the non-monotonicity is not symmetric about $x = 0$, while the covariate distributions (for each batch) are symmetric about $x = 0$. `cComBat` struggles and worsens the batch effect as the covariate balance decreases, while `Causal cComBat` is able to work as long as there are moderate amounts of covariate overlap. In high balance the estimated ATE of `cComBat` and `Causal cComBat` are 0.0497 and 0.0421 (`Causal cComBat` better 59.5% of the time), in moderate balance are 0.074 and 0.130 (`Causal cComBat` better 99.7% of the time), and in limited overlap are 2.61 and 0.407 (`Causal cComBat` always better) with `Causal cComBat` only executing in 33 simulations.

Linear A linear relationship between the covariate and the outcome. The outcome is:

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

The effectiveness of batch effect correction methods for the linear setting are shown in Figure 11.

D.3 Estimated Average Treatment Effect To evaluate the effectiveness of each batch effect correction technique on simulated data, we compute the true data expected signal for each covariate level; e.g., $\mathbb{E}\left[Y_i^{(t)}|x\right]$ for each batch t and each covariate level x. Since ϵ_i has mean 0, this would be the quantity:

$$\mathbb{E}\left[Y_i^{(t)}|x\right] = f(x) + t,$$

where f is the covariate/outcome relationship (possibly incorrectly modeled). Since x is continuous, we compute this value for x across $B = 200$ evenly spaced breakpoints for evaluation. Given a set of samples, we train a batch effect correction model, and fit the trained model to the expected signal for each covariate level, leaving us with corrected expected signal for each batch, which we denote by $v(t, x)$. In theory, if the batch effect correction technique removes the batch effect as modeled, $v(1, x) \approx v(0, x)$ for all x. To evaluate each technique, we consider the estimated average treatment effect (estimated ATE) for each trial r of 1000 trials:

$$d_r = \frac{1}{B} \sum_{x \in \mathcal{X}} |v(1, x) - v(0, x)|,$$

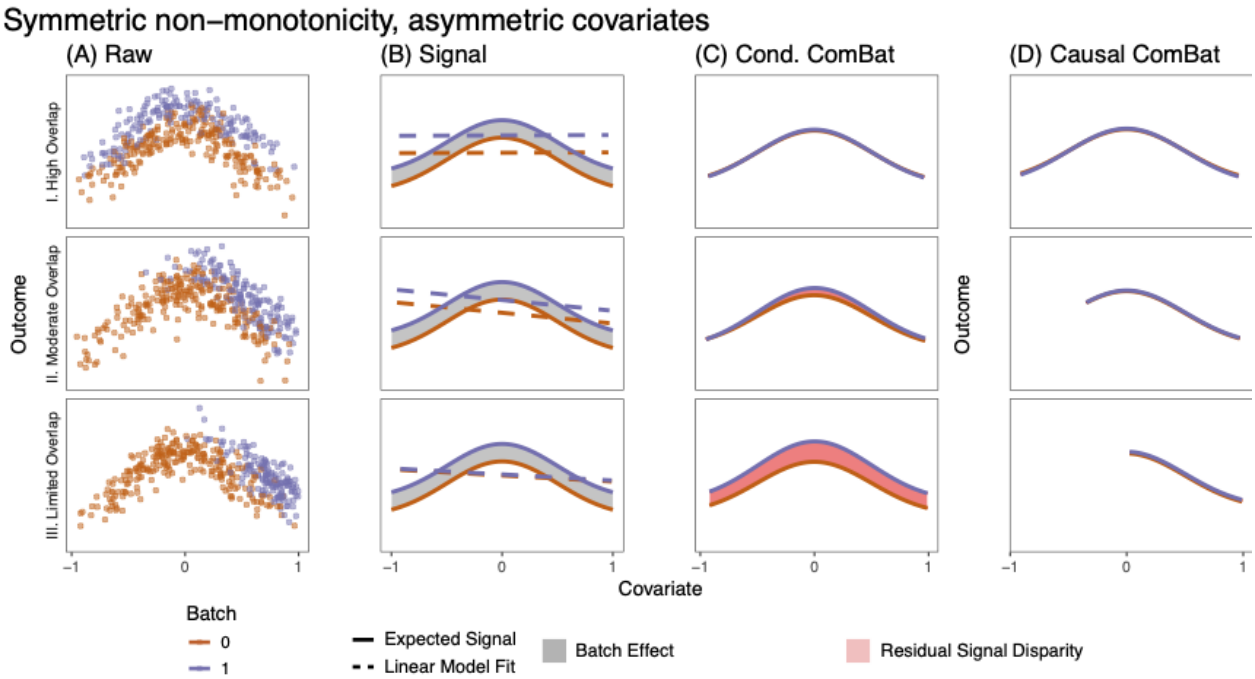

Figure 10: A symmetric non-monotonicity about $x = 0$, but the covariate distribution is no longer symmetric about $x = 0$. `cComBat` struggles to remove the batch effect and leaves a residual signal disparity, but `Causal cComBat` is able to address the batch effect within a range of covariate overlap. In high balance the estimated ATE of `cComBat` and `Causal cComBat` are 0.075 and 0.0506 (`Causal cComBat` better 65.8% of the time), in moderate balance are 0.181 and 0.0751 (`Causal cComBat` better 96.0% of the time), and in limited overlap are 0.84 and 0.106 (`Causal cComBat` always better). Due to the fact that the covariate distribution for batches 0 and 1 overlap for this simulation, `Causal cComBat` always executes.

and the magnitudes of d_r are annotated in the plots (shaded red boxes) for a single simulation. We compute the mean estimated average treatment effect as:

$$d = \frac{1}{r} \sum_{r \in [R]} d_r.$$

A value of 0 corresponds to the batch effect being completely eliminated (the expected signal for each batch after correction is identical), a value of 1 would equate to the residual disparity between the expected signals being the same as before batch effect correction, and a value > 1 corresponds to the residual disparity between the expected signals *exceeding* the original batch effect.

D.4 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 11), non-linear (Figure 4), or non-monotonic (Figure 8, Figure 9, and Figure 10) covariate/outcome relationship, with a batch-specific offset. We performed these simulations for cases where the covariate distributions overlap (row I), moderately overlap (row II), and have limited overlap (row III). 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 column (D), when further conditions are satisfied, such as:

1. covariate distributions are matched: 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. `Causal cComBat` attempts to enforce this approach through matching. We explored the

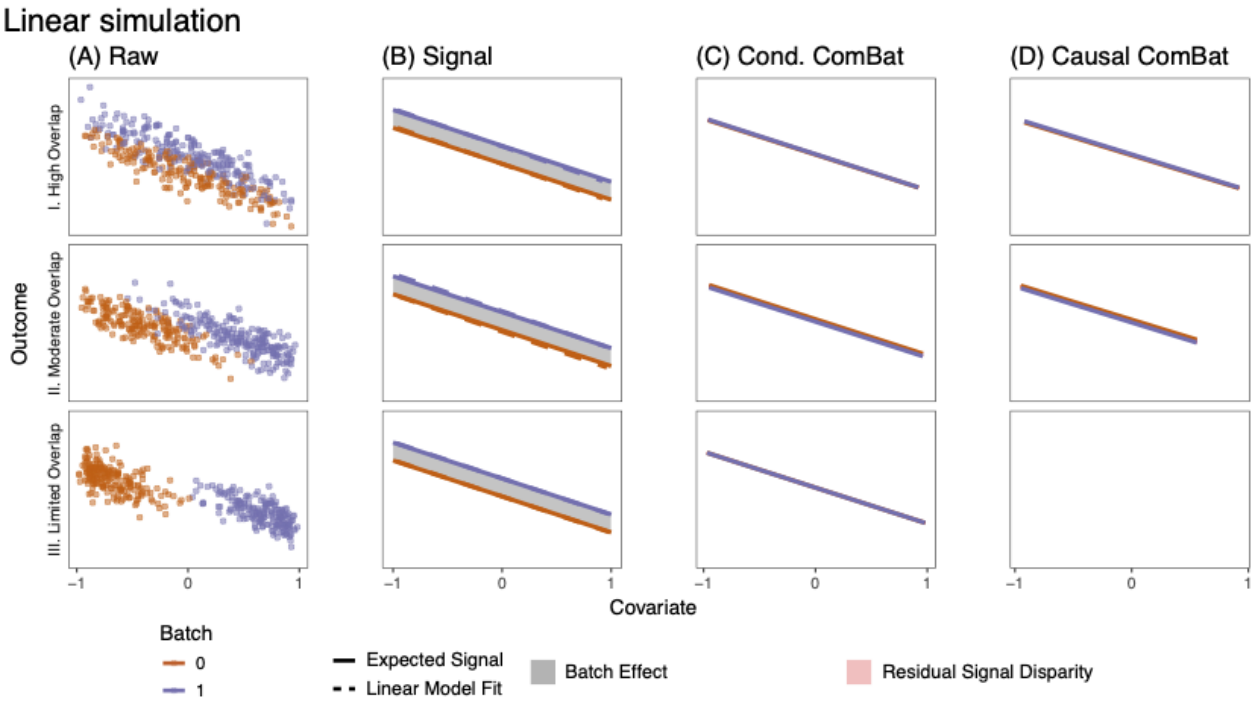

Figure 11: A linear simulation in which the assumptions of `cComBat` are perfectly satisfied. In high balance the estimated ATE of `cComBat` and `Causal cComBat` are 0.035 and 0.0375 (`Causal cComBat` better 42.5% of the time), in moderate balance are 0.074 and 0.113 (`Causal cComBat` better 33.1% of the time), and in limited overlap are 0.145 and 0.292 (`Causal cComBat` better 23.1% of the time) with `Causal cComBat` only executing in 39 simulations.

implications of this in Figure 4 and Figure 9.

2. the covariate distributions (and the covariate/outcome relationship) are symmetric about the same point: again, the idea here is that intuitively, the the inaccurate assumptions of the model must be uniformly misrepresented by the model in the two batches. This is a pre-hoc condition and relates directly to the specific underlying true veridical signal, and is difficult to exploit practically. We explore the implications of this in Figure 8 and Figure 10.

As shown in the simulations in Figure 4 and Figure 9, these artifacts can be arbitrarily large depending on the nature of the incorrect assumption. In these examples, the residual signal disparity introduced by `cComBat` in limited overlap cases (row III) dwarfs the magnitude of the original batch effect, and is of similar magnitude with moderate covariate overlap (row II).

On the other hand, `Causal cComBat` first pre-aligns the covariate distributions, and even with model misspecifications, is able to successfully remove a batch effect for the subset of the covariate space with sufficient covariate overlap. Importantly, 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 is typically the case for most of the illustrated similations in the "Limited Overlap" condition. 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 effectively impossible to tell whether the implicit or explicit assumptions of the model for batch effect correction are sufficient or not, and almost impossible for our model to account for all ways a batch effect could manifest in high-dimensional data (such as higher order correlations, non-linearities, variance differences) given a finite sample. 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 modeling assumptions are within reason on a dimension-by-dimension (or, higher-order) basis.

Addressing the batch effect One way to characterize the batch effect in univariate contexts (and within our simulations) would be analogous to the average treatment effect [12]; e.g.:

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)]$$

For all simulations, by construction the ATE is 1 and does not depend on the covariates.

By definition, a difference in the expectations implies that $Y(1) \overset{\mathcal{D}}{\neq} Y(0)$, 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 [12] or f is known, the adjustment factor $\hat{\beta}_1 - \hat{\beta}_0$ will be a consistent and unbiased estimate of the ATE [12, 64]. For a non-technical explanation in words, we can obtain reasonable estimates of the ATE in one of three ways (and consequently, account for the batch effect):

1. Extrapolation: if we know $g(x)$, then we can "guess" what $\mathbb{E}[Y_i(t)|x]$ 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 Figure 11; particularly, noting cells (C).II and (C).III.
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 CATE. Proofs of this result can be found in [12, 13]. In this particular case, $X_i \perp\!\!\!\perp T_i$, which implies the strong ignorability condition [12]. For more trivial reading specific to the case where $X_i \perp\!\!\!\perp T_i$, see [64]. This result is exhibited in the simulations for Figure 4 and Appendix Figure 9 looking at column (D) compared to column (C), and noting that causal pre-processing procedures tend to produce a reasonable result (while non-causal conditional procedures with a model misspecification do not).
3. The covariate distribution is symmetric about the same point that a non-monotonicity is symmetric: in this extremely special case, we have that $f(x|1) = f(-x|0)$, and $f(y(0)|x) = f(y(1)|-x)$ for every x. This result is exhibited by way of comparison of Figures 8 and Figure 9, and noting that while Figure 8(C).III is able to arrive at a reasonable solution, Figure 9(C).III does not.

By a similar argument, the estimated batch disparity 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.

Simulations in which the assumption about $f(x)$ 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.

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 [31]. 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 $f_d(a|S = s, C = c)$ is the conditional density of age, conditional on the individual's sex being s, continent of measurement being c for a given dataset d. The mass $\mathbb{P}_d(S = s|C = c)$ is the conditional mass of sex, conditional on the individual's continent of measurement being c, for dataset d.

Finally, the mass $\mathbb{P}_d(C = c)$ represents the mass of an individual's continent of measurement being c, for dataset 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:

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

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.

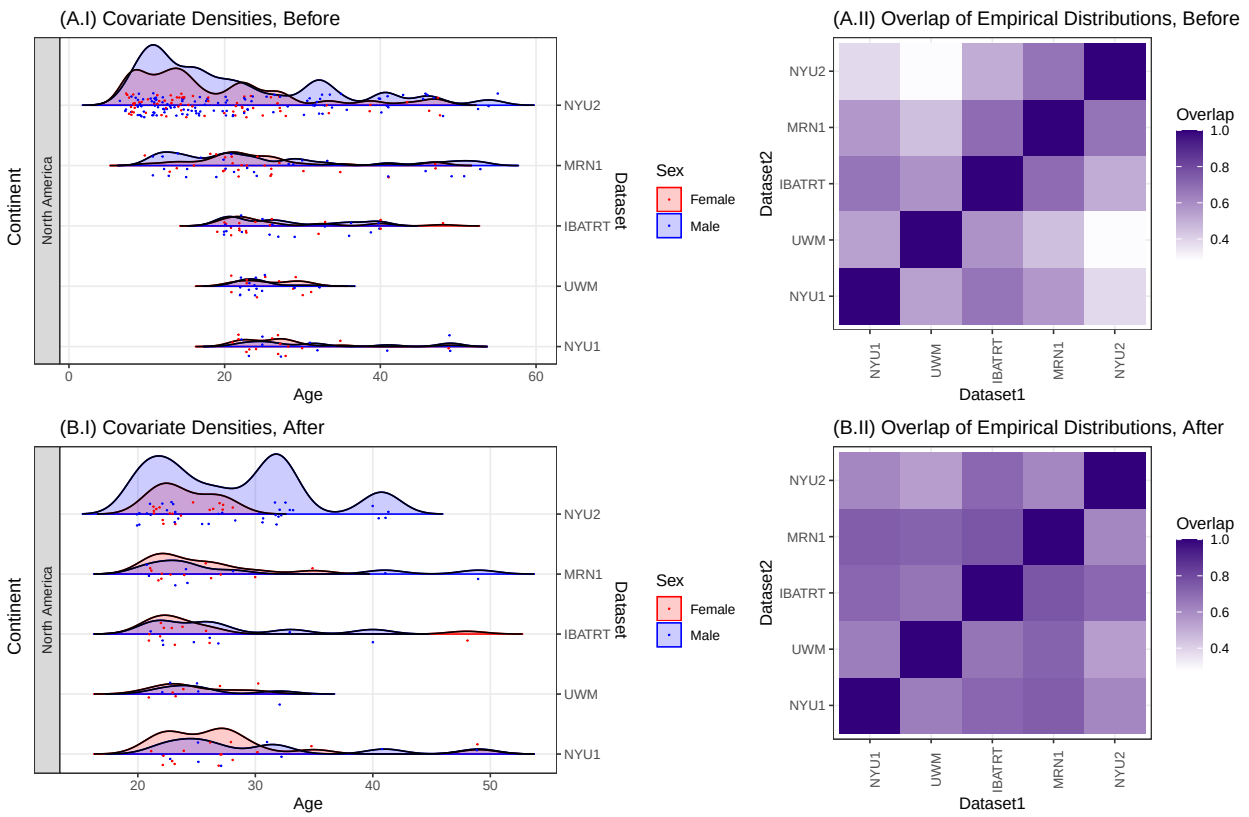

Figure 12: **The overlap of the empirical covariate distributions for the American Clique, before and after adjustment**. **(A)** The empirical distributions of the covariates before adjustment. **(B)** The empirical overlap of the covariate distributions after the adjustment procedure is applied, as discussed in Appendix C.

Associational and Crossover effect detection As for conditional and adjusted effects in Figure 6, we can perform a similar investigation for associational or causal crossover effects, which we add in Appendix Figure 14. Associational effects can almost always be detected (405 of 406 pairs of datasets), regardless of whether or not any demographic overlap is present or any comparison seems reasonable (distance correlation, BH correction, $\alpha = 0.05$). This highlights that associational effects, perhaps, can be best thought of as determining whether or not the data is simply different across a given pair of batches, and are not particularly useful for attribution of the source of a detected difference (such as to batch effects, or demographic effects). 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,

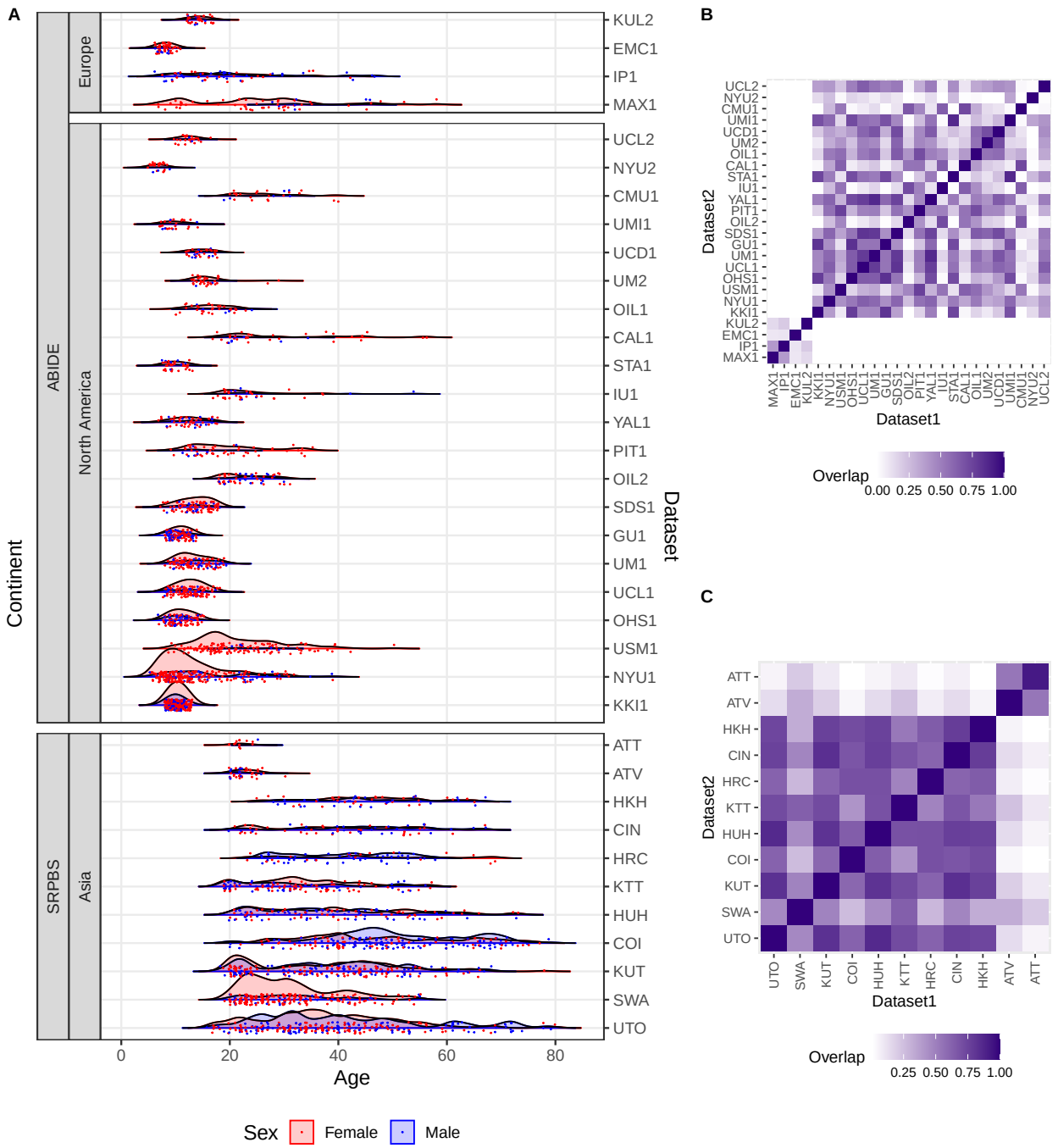

Figure 13: **The overlap of the empirical covariate distributions for two mega-studies. (A)** The empirical distribution of covariates. **(B)** The overlap of covariate distributions given by the distribution-free overlapping index for ABIDE mega-study [37, 38]. **(C)** The overlap of covariate distributions given by the distribution-free overlapping index for the SRPBS mega-study. [39]. 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 4, Appendix 9, or Appendix Figure 10 if modeling assumptions are not reasonable.

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. Causal crossover effects, which require the fewest assumptions to be bonified causal batch effects, are present for 2 of

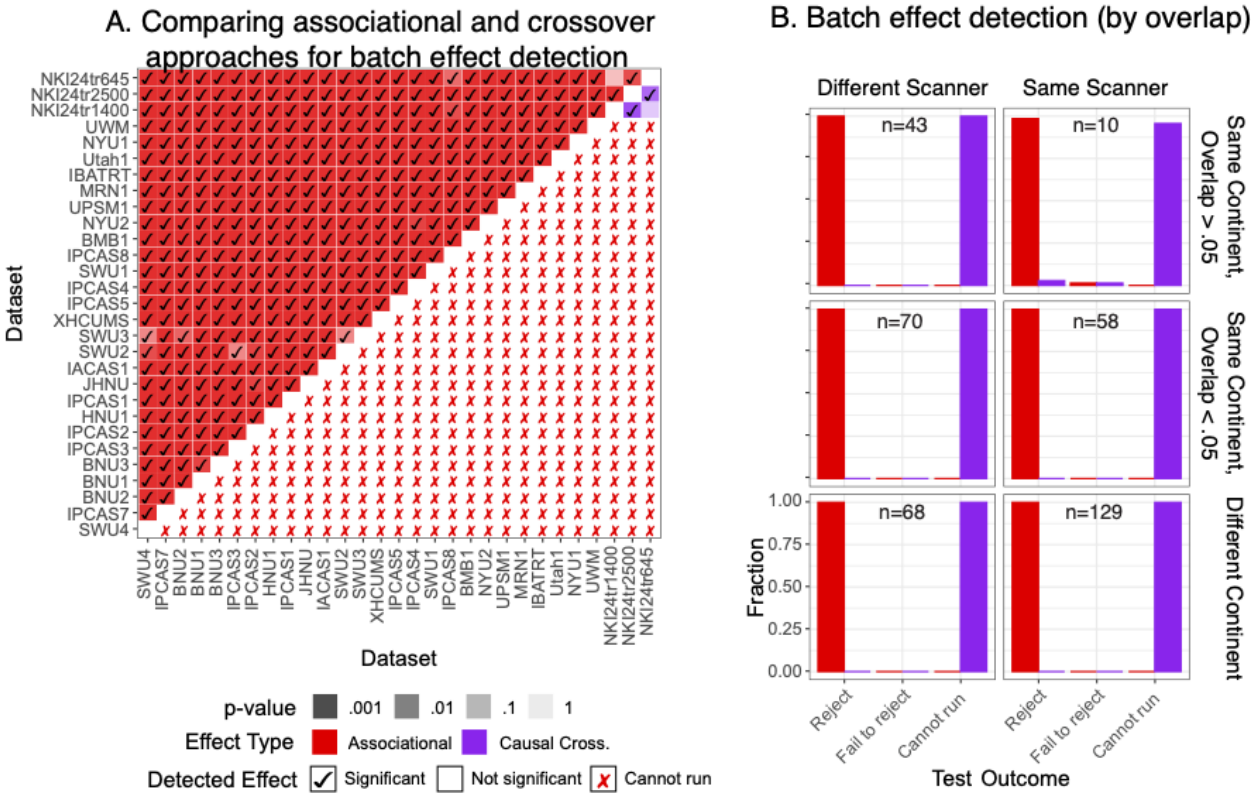

Figure 14: **Comparison of associational and causal crossover effects between pairs of datasets from the CoRR study**. **(A)** A heatmap analogous to Figure 6(A), but for associational and causal crossover effects. Most associational effects are significant (405 of 406) and most causal crossover effects are siginficant (2 of 3 pairs of datasets with the same individuals measured multiple times). **(B)** analogous to Figure 6(B), but for associational and conditional effects.

3 possible combinations of datasets (distance correlation, BH correction, $\alpha = 0.05$). Unfortunately, this requires a very specific recruitment setup (that is, the same individuals measured across the batches) which is not met for the 403 remaining possible pairs of datasets.

Batch Effects are reduced with Causal `ComBat` Here we investigate batch effects before and after correction. Figure 15 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. 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 16. The left column of Figure 16A 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 16A 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

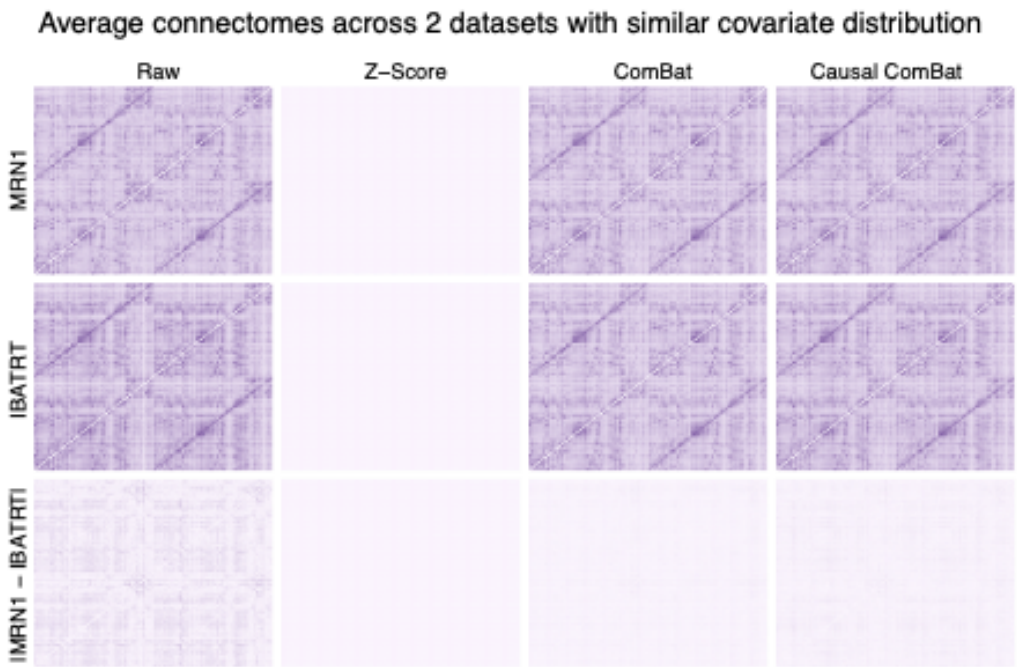

Figure 15: **Distribution of Detected Effects Before and After Correction**. A comparison of the average connectomes from two studies from the same continent, with similar sex and age distributions conditional on sex. While the *Raw* average connectomes appear similar, the absolute difference between the two is appreciable. All of the three adjustment strategies reduce the absolute difference between the average connectomes for the two batches.

Figure 16B, these two properties are examined statistically for the individuals from the American Clique (Mann-Whitney U Statistic, see Methods A.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 17 explores the distribution of edge weights within these two community structures before and after correction. The conclusions are similar in that Causal `ComBat` and `ComBat` preserve the property that the homotopic edges have far higher connectivity than non-homotopic edges.

Appendix G. Figure 7 (Extended). This section extends Figure 7 from the paper to include `cComBatGAM` and an adaptation of `cComBatGAM`, Causal `cComBatGAM`, which prepends the causal pre-processing to `cComBatGAM`. In Figure 18B, note that `cComBatGAM` tends to produce similar inference to other non-causal strategies, and `cComBatGAM` tends to produce similar inference to `Causal cComBat`. This illustrates that considerations of causality heavily influence subsequent inference.

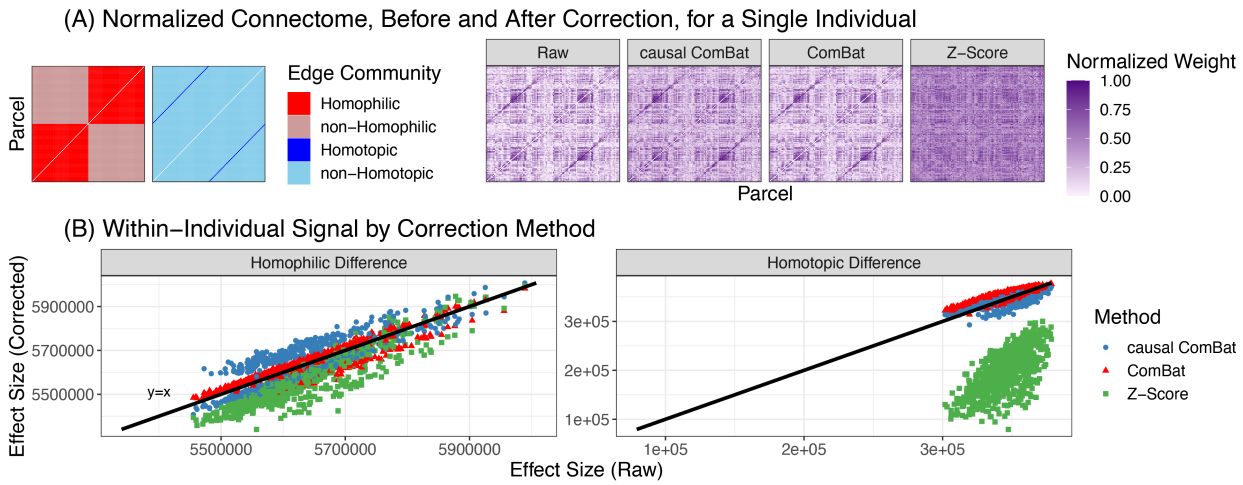

Figure 16: **Preservation of Topological Properties of Connectomes after Batch Correction**. The presence of topological effects is explored in the context of the homophilic and homotopic differences in connectivity in the American Clique. **(A)** (*Left*) A map indicating which edges are associated 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.

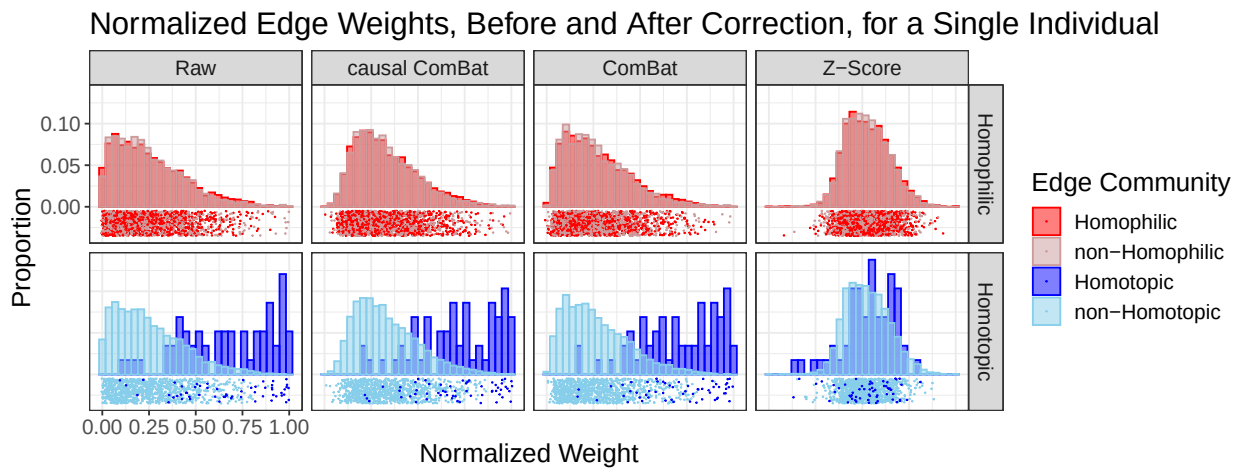

Figure 17: Histogram and rug plot indicating the empirical distribution of the normalized edge weights associated with each of the four edge communities for a single individual's connectome.

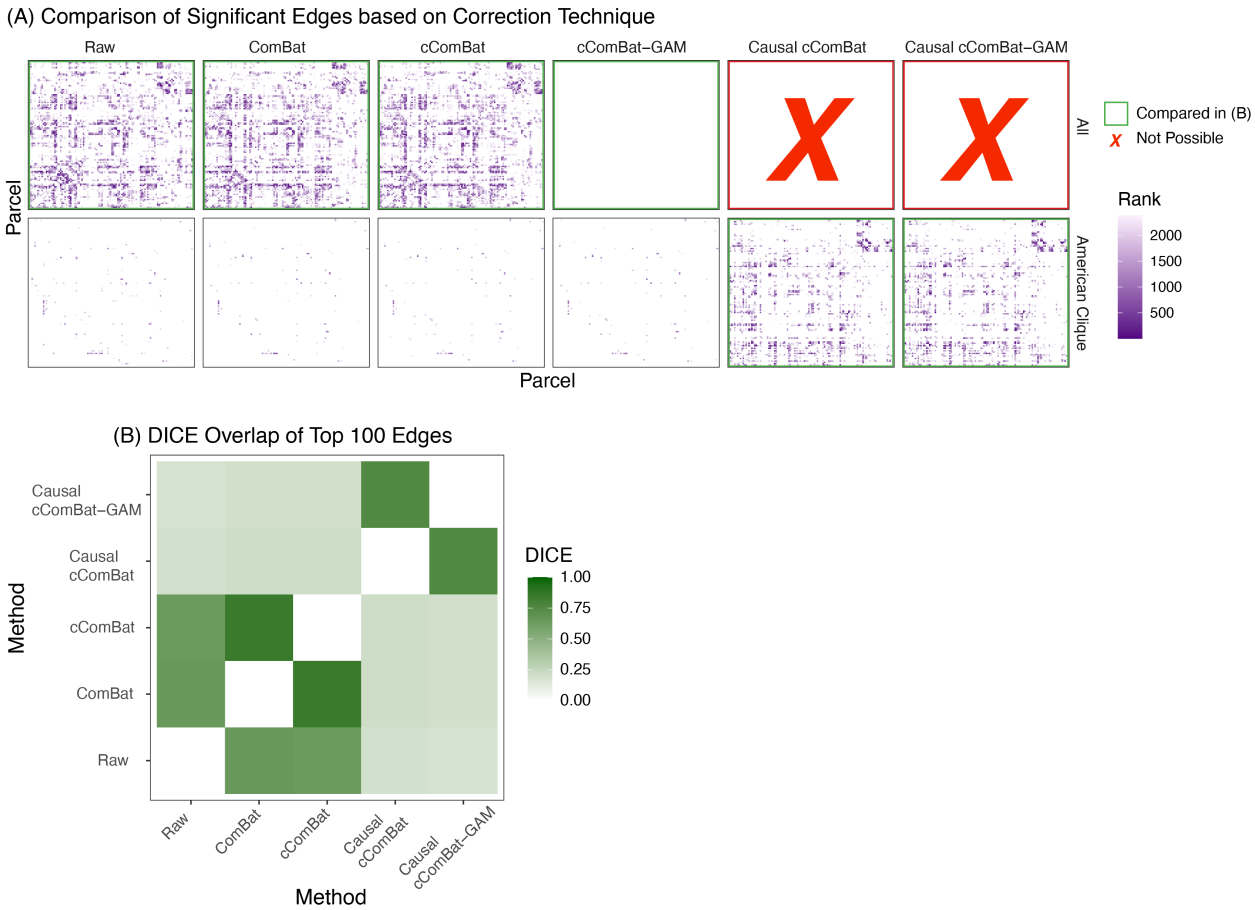

Figure 18: **Extended Version of Figure 7 to include `cComBatGAM`. (A)** illustrates the heatmaps of significant edges, ranked from smallest p-value (dark purple) to largest significant p-value (light purple). **(B)** illustrates the DICE overlap of the top 100 most significant edges across all strategies. Non-causal strategies tend to produce similar inference, and causal strategies tend to produce similar inference to one-another, but non-causal strategies and causal strategies produce dissimilar inference.