

Collaborative Indirect Treatment Comparisons with Multiple Distributed Single-arm Trials

Yuru Zhu

Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania

Huiyuan Wang

Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania

Haitao Chu

Global Biometrics and Data Management, Pfizer Inc.

Division of Biostatistics and Health Data Science, University of Minnesota Twin Cities

Yumou Qiu

School of Mathematical Sciences, Peking University

and

Yong Chen*

Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania

September 30, 2025

Abstract

When randomized controlled trials are impractical or unethical to simultaneously compare multiple treatments, indirect treatment comparisons using single-arm trials offer valuable evidence for health technology assessments, especially for rare diseases and early-phase drug development. In practice, each sponsor conducts a single-arm trial on its own drug with restricted data-sharing and targets effects in its trial population, which can lead to unfair comparisons. This motivates methods for fair treatment comparisons across a range of target populations in distributed networks of single-arm trials sharing only aggregated data. Existing federated methods, which assume at least one site contains all treatments and allow pooling of treatment groups within the same site, cannot address this problem. We propose a novel distributed augmented calibration weighting (DAC) method to simultaneously estimate the pairwise average treatment effects (ATEs) across all trial population combinations in a distributed network of multiple single-arm trials. Using two communication rounds, DAC estimators balance covariates via calibration weighting, incorporate flexible nuisance parameter estimation, achieve doubly robust consistency, and yield results identical to pooled-data analysis. When nuisance parameters are estimated parametrically, DAC estimators are enhanced to achieve doubly robust inference with minimal squared first-order asymptotic bias. Simulations and a real-data application show good performance.

*corresponding author

Keywords: Calibration weighting; Distributed inference; Double robustness; Indirect treatment comparison; Single-arm trials.

1 Introduction

Generating evidence for relative effectiveness of multiple interventions is crucial in modern healthcare, with applications in drug development, public health policy evaluation, and precision medicine (Miksad et al., 2019; Dagenais et al., 2022; Williams and Evans, 2023). While randomized controlled trials that jointly evaluate all treatments remain the gold standard, they are often infeasible in practice due to strict eligibility criteria, the growing number of treatment options, or ethical concerns. For example, guidelines developed by the U.S. Food and Drug Administration (2001) discourage placebo controls when an available treatment is known to prevent serious harm, such as death or irreversible morbidity. In such cases, single-arm trials, where all participants receive a predetermined treatment, offer an alternative. These trials are widely used for rare diseases and early-phase drug development, and have even supported regulatory approvals (Vickers et al., 2007; Seymour et al., 2010; Hatzwell et al., 2016).

Because single-arm trials lack concurrent comparators, auxiliary data are often used to impute the “missing” comparator group, enabling indirect treatment comparisons (ITCs) in an “unanchored” manner (Phillippo et al., 2016; Thorlund et al., 2020; Jahanshahi et al., 2021). ITCs play a critical role in health technology assessment (HTA) and joint clinical assessment (JCA) in pharmaceutical research, where timely and fair evaluation of new therapies is important. To support it, the HTA Regulation (European Union, 2021; European Commission, 2024a,b) in the European Union mandates consistent, transparent, and comparative evaluation of clinical evidence. In drug development, each sponsor typically evaluates only its own investigational drug, while access to competitors’ data is restricted by commercial confidentiality or prohibitive licensing costs. Sponsors, assessors and regulators often conduct ITCs by emulating comparator groups using summary statistics (Phillippo et al., 2016; Patel et al., 2021), leading to distributed networks of single-arm trials, particularly in rare diseases where placebo controls are infeasible. Further-

more, equitable assessment requires treatment comparison on a common target population, but sponsors usually prioritize their own trial population, yielding misaligned evaluations.

Distributed networks permit sharing aggregated data but face challenges, including efficiency loss due to distributed inference and heterogeneous covariate distributions across sites, and communication burden (Gu and Chen, 2023). Most existing federated causal inference settings consider that at least one site contains all treatments of interest (Xiong et al., 2023; Hu et al., 2024; Han et al., 2025; Guo et al., 2025). In such cases, the causal effect can be identified using data from that site alone, and federation primarily improves efficiency. Moreover, these methods estimate the effect for one prespecified target population per run, requiring repeated executions for multiple populations and thus increasing communication costs.

These challenges motivate developing methods that (i) support ITCs within distributed networks of multiple single-arm trials, and (ii) enable simultaneously comparing multiple treatments across a range of target populations, including all possible combinations of single-arm trial populations, while respecting data-sharing constraints and minimizing communication burden with no efficiency loss. To this end, we propose a novel Distributed Augmented Calibration weighting (DAC) method for ITCs in a distributed network of multiple single-arm trials, where each site retains its data locally and shares aggregated data with a coordinating center.

Our contributions are twofold. First, we introduce a new federated causal inference setting where each site conducts a single-arm trial for a unique treatment and individual-level data cannot be shared across treatment groups. This setting, unlike existing work that allows pooling of treatment groups within the same site, imposes stricter data-sharing constraints, making federation essential for identification. It reflects practical scenarios in HTA submissions for comparative effectiveness evidence of new drugs and naturally extends to federated causal inference settings where each site evaluates a subset of treatments.

Second, we develop the DAC estimation method to address the unique challenges of this setting. DAC leverages calibration weighting (Deville and Särndal, 1992) with summary statistics to balance covariates. DAC estimators accommodate flexible parametric or nonparametric nuisance estimation, and achieve doubly robust consistency. Under parametric models, our bias-reduced DAC provides the first adaptation of the bias-reduction technique (Vermeulen and Vansteelandt, 2015) to the distributed network of single-arm trials with a tailored algorithm, enabling doubly robust inference with minimal squared first-order asymptotic bias under strict data-sharing constraints. Crucially, our DAC estimator exactly recovers the pooled augmented calibration weighting estimator using all individual-level data, ensuring no information loss. With only two communication rounds of aggregated data, DAC addresses heterogeneous covariate distributions, protects privacy, avoids information loss, and delivers valid inference for all pairwise average treatment effects (ATEs) across all possible target populations formed from site combinations.

1.1 Related Work

The basis of using the internal single-arm trial data with its external auxiliary data to conduct indirect comparison is the assumption that links mechanisms of generating potential outcomes across data sources, allowing the generalization of potential outcomes from a study sample to a target population. When individual-level auxiliary data on alternative treatments are available, the distributional exchangeability assumption, which ensures counterfactual outcomes across sources have the same conditional distribution given observed confounders, allows combining all data as a multi-treatment study. Causal methods such as stratification, matching, modeling, and weighting can then estimate target ATEs (Lee, 2018; Li and Song, 2020; Shi et al., 2023; Wang et al., 2024).

When only limited summary statistics are available, estimation becomes more restrictive. Meta-analysis methods rely on an assumption that the expected counterfactual outcome has same

value or distribution across sources, but may yield biased estimates when there are systematic differences between the single-arm trial and auxiliary data populations (Schmidli et al., 2014; Woolacott et al., 2017; Weber et al., 2018). The unanchored matching-adjusted indirect comparison (MAIC) approach (Signorovitch et al., 2010) combines the internal single-arm trial data with some fixed external summary statistics of the alternative treatment via calibration weighting (Phillippo et al., 2018; Cheng et al., 2020). MAIC addresses covariate heterogeneity but only identifies the treatment effect in the external data population and depends on correctly specifying the selection model for inclusion in the internal single-arm trial, as poorly chosen or irrelevant variables can introduce bias (Deville and Särndal, 1992; Chan et al., 2015; Zubizarreta, 2015). However, summary statistics for calibration are often limited in practice, especially when comparing an increasing number of treatments simultaneously under complex selection mechanisms.

In contrast, our DAC approach requires two communication rounds to transfer specific aggregated data, balancing patient privacy with the limitations of fixed summary statistics. DAC handles heterogeneous covariate distributions, supports simultaneous pairwise comparisons across a range of target populations and achieves the same results as pooled individual-level analysis.

1.2 Overview of the Paper

Section 2 outlines the problem setup and identification of the estimands in a distributed network. Section 3 introduces the general DAC estimation approach and its algorithm with nonparametric methods for estimating working models. Section 4 provides the bias-reduced DAC estimation for doubly robust inference under parametric working models. Section 5 illustrates the asymptotic properties of the proposed DAC estimators. Sections 6 and 7 show performance of the DAC estimator in simulations and empirical data analysis. Section 8 concludes with a discussion.

2 Preliminaries

We consider a distributed network consisting of K finite sites. Let the categorical random variable $D \in \{1, 2, \dots, K\}$ index the site and X denote a vector of pre-treatment covariates that explain treatment assignments, such as demographic characteristics and inclusion/exclusion criteria. After selecting eligible patients, each site conducts a single-arm trial and collects the outcome Y . At site k , we observe n_k independent and identically distributed samples $\{(D_i, Y_i, X_i)\}_{D_i=k}$ drawn from its underlying single-arm trial population whose distribution is denoted as P_k . Let $f_{X|D=k}(x)$ be the density of covariates X at site k . Denote the total sample size of all sites as $N = \sum_{d=1}^K n_d$ with the limiting proportion $\mathbb{P}(D = k) := \lim_{N \rightarrow \infty} n_k/N > 0$ for each site $k \in \{1, 2, \dots, K\}$. For better illustration, we focus on the case where no treatment assignments are repeated across sites, thus D can also represent the treatment assignment. Extensions to cases of repeated treatments at different sites or multiple treatments assigned within a local site are discussed in Section S6 of the Supplementary Material (SM).

This paper aims to develop an estimation procedure that, in a single implementation, enables simultaneously comparing the effectiveness of various treatments across different subpopulations. Specifically, we address the following questions for each site:

(Q.1) How would outcomes change if its local patients received a different treatment?

(Q.2) What would a relative effect be if treatments were compared in another patient population?

We approach these questions using the potential outcome framework (Rubin, 1974). Let $Y(d)$ be the potential outcome observed if the subject were assigned treatment d . The causal estimand

$$\tau_{(k,k')|\mathcal{I}} = \mathbb{E}\{Y(k') \mid D \in \mathcal{I}\} - \mathbb{E}\{Y(k) \mid D \in \mathcal{I}\}, \quad k, k' = 1, \dots, K, k \neq k', \quad (1)$$

represents the comparative effectiveness of treatment k' over k in the subpopulation of patients

in the sites indexed by a subset $\mathcal{I} \subseteq \{1, \dots, K\}$, whose covariate density is $\{\sum_{d \in \mathcal{I}} \mathbb{P}(D = d)\}^{-1} \sum_{d \in \mathcal{I}} \mathbb{P}(D = d) f_{X|D=d}(x)$. The cases $\mathcal{I} = \{k\}$ and $\mathcal{I} \neq \{k\}$ correspond to Questions (Q.1) and (Q.2), respectively. Rather than focusing on one pre-specified target population, we consider all subpopulations corresponding to site subsets $\mathcal{I} \subseteq \{1, \dots, K\}$. Since the estimand in (1) is a difference in expectations of two potential outcomes, it suffices to identify the expected potential outcome $\mu_{k|\mathcal{I}} \equiv \mathbb{E}\{Y(k) \mid D \in \mathcal{I}\}$ in the target population. We introduce the following assumptions to establish the identifiability of $\mu_{k|\mathcal{I}}$ under the distributed data structure.

Following the stable unit treatment value assumption that there is no interference between units and no multiple versions of treatment, we make the consistency assumption:

Assumption 1 (Consistency). *The observed outcome is $Y = \sum_{d=1}^K Y(d)I(D = d)$.*

By Assumption 1, the observed outcome at site k is the potential outcome under treatment k .

Assumption 2 (Exchangeability). *The conditional distributions of potential outcomes given covariates are identical across sites: $(Y(1), \dots, Y(K)) \mid (X, D = j) \stackrel{d}{=} (Y(1), \dots, Y(K)) \mid (X, D = k)$ for any $j, k \in \{1, \dots, K\}$ with $j \neq k$.*

This assumption posits that treatments operate via the same mechanism across different single-arm trial populations given observed covariates, ensuring no unmeasured confounding between sites (treatments) and potential outcomes (Stuart et al., 2011; Shi et al., 2023). If local data were pooled, Assumption 2 implies that the K single-arm trials would collectively resemble a hypothetical K -arm randomized trial. For single-arm trials investigating different treatments for the same disease, aligned eligibility criteria within similar time periods make this assumption reasonable in our context (Gray et al., 2020). Although the exchangeability assumption is generally not testable, Section S6.3 of the SM discusses its assessment using negative control outcomes and provides practical guidance for mitigating bias when this assumption may be violated. To ensure comparability across sites, we impose the following positivity assumption.

Assumption 3 (Positivity). *For a positive constant ε , $\min_{j,k=1,\dots,K} \frac{f_{X|D=j}(x)}{f_{X|D=k}(x)} \geq \varepsilon$ almost surely.*

Assumption 3 is plausible, as eligibility criteria for single-arm trials conducted in close time periods are often similar, leading to identical supports for covariate distributions across different single-arm trials. Let $Z = (D, Y, X^T)^T$ and P denote the distribution of Z for the overall population combining all single-arm trial populations, i.e., $P = \sum_{d=1}^K \mathbb{P}(D = d) \times P_d$. Let $m_k(X) \equiv \mathbb{E}(Y \mid X, D = k)$ be the conditional mean function of outcome Y given X at the local site k and $w_{jk}(x) \equiv \{f_{X|D=j}(x)\mathbb{P}(D = j)\} / \{f_{X|D=k}(x)\mathbb{P}(D = k)\}$ be a weighting function for $j, k \in \{1, \dots, K\}$. Then, the identification is presented as follows.

Proposition 1. *Under Assumptions 1–3, for any $k \in \{1, \dots, K\}$ and $\mathcal{I} \subseteq \{1, \dots, K\}$, the expected potential outcome $\mu_{k|\mathcal{I}}$ can be identified in two forms:*

$$\mu_{k|\mathcal{I}} = \mathbb{E}_P \left\{ \frac{m_k(X)I(D \in \mathcal{I})}{\mathbb{P}(D \in \mathcal{I})} \right\}. \quad (2)$$

$$\mu_{k|\mathcal{I}} = \mathbb{E}_P \left\{ \frac{YI(D = k)}{\mathbb{P}(D \in \mathcal{I})} \times \sum_{j \in \mathcal{I}} w_{jk}(X) \right\}. \quad (3)$$

Proposition 1 leads to two plug-in estimators for $\mu_{k|\mathcal{I}}$. The first is the outcome regression (OR) estimator $\hat{\mu}_{k|\mathcal{I}}^{\text{OR}} = \{\sum_{i=1}^N \hat{m}_k(X_i)I(D_i \in \mathcal{I})\} / \{\sum_{j \in \mathcal{I}} n_j\}$, which averages predictions of the estimated conditional mean function $\hat{m}_k(X)$ across patients in sites indexed by $\mathcal{I}$. The second is the propensity score weighting (PS) estimator $\hat{\mu}_{k|\mathcal{I}}^{\text{PS}} = \{\sum_{i=1}^N Y_i I(D_i = k) \sum_{j \in \mathcal{I}} \hat{w}_{jk}(X_i)\} / \{\sum_{j \in \mathcal{I}} n_j\}$, which plugs the estimated weighting function $\hat{w}_{jk}(X)$ in (3). These methods are widely used for causal effect estimation (Shi et al., 2023). However, consistency of OR or PS estimators relies on correct specifications of regression models or propensity score models, respectively (Lunceford and Davidian, 2004; Funk et al., 2011). Moreover, estimating weighting functions in distributed settings without sharing individual-level data poses challenges. To address these issues, the next two sections introduce our Distributed Augmented

Calibration weighting (DAC) method. DAC leverages the two forms of identification in Proposition 1 to provide robust estimators for simultaneously comparing treatment effects across various target populations, requiring only aggregated data shared within minimal communication rounds.

3 Distributed Augmented Calibration Method

In this section, we propose the general DAC approach for doubly robust estimation of $\tau_{(k,k')|\mathcal{I}}$ for any $k, k' \in \{1, \dots, K\}, k \neq k'$ and $\mathcal{I} \subseteq \{1, \dots, K\}$, which can incorporate nonparametric methods for estimating outcome regression models and weighting functions. We begin by deriving the efficient influence functions (EIFs) for target estimands. The EIF for $\mu_{k|\mathcal{I}}$ is

$$\varphi_{k|\mathcal{I}}(Z; P) = \frac{1}{\mathbb{P}(D \in \mathcal{I})} \left[I(D \in \mathcal{I}) \{m_k(X) - \mu_{k|\mathcal{I}}\} + I(D = k) \{Y - m_k(X)\} \sum_{j \in \mathcal{I}} w_{jk}(X) \right],$$

which corrects the first-order bias of plug-in estimators (see Section 2.2.5 in Chernozhukov et al. 2018). Similarly, the EIF for $\tau_{(k,k')|\mathcal{I}}$ is $\phi_{(k,k')|\mathcal{I}}(Z; P) = \varphi_{k'|\mathcal{I}}(Z; P) - \varphi_{k|\mathcal{I}}(Z; P)$. The derivation of these EIFs is detailed in Theorem 1 in Section 5.

From above EIFs, our DAC requires estimating two working models: the outcome model $m_k(X)$ and weighting function $w_{jk}(X)$. While $m_k(X)$ can be estimated locally at site k , estimating $w_{jk}(X)$ without accessing individual-level data is challenging due to data-sharing restrictions and population heterogeneity in distributed settings. Instead of directly modeling the sampling propensity score, we adopt the calibration weighting (CW) approach (Deville and Särndal, 1992), which uses only summary-level statistics to balance covariates and avoids extreme weights (Qin and Zhang, 2007; Hainmueller, 2012; Zubizarreta, 2015). It is based on the identity $\mathbb{E}[I(D = k)w_{jk}(X)g(X)/\mathbb{P}(D = j)] = \mathbb{E}\{g(X) \mid D = j\}$, where $g(X)$ is a vector of functions of X (e.g., moments of X) to be calibrated. Its empirical analog leads to the calibration

constraints $\sum_{D_i=k} w_{jk}(X_i)/n_j g(X_i) = \bar{g}_j(X)$, where $\bar{g}_j(X) = \sum_{D_i=j} g(X_i)/n_j$ is the sample mean of $g(X)$ at site j , suggesting aligning the weighted mean of $g(X)$ from site k with the sample mean of $g(X)$ at site j to achieve covariate balance between the two sites. Weights are estimated by minimizing a loss function subject to these balancing constraints:

$$\begin{aligned} \min_{\{w_{jk}(X_i), D_i=k\}} \sum_{D_i=k} d(w_{jk}(X_i)/n_j, 1/n_j) \\ \text{subject to } w_{jk}(X_i) \geq 0, \sum_{D_i=k} w_{jk}(X_i)/n_j = 1 \text{ and } \sum_{D_i=k} w_{jk}(X_i)/n_j g(X_i) = \bar{g}_j(X). \end{aligned} \quad (4)$$

Here, we use the negative entropy loss $d(w_{jk}(X_i)/n_j, 1/n_j) = w_{jk}(X_i) \log\{w_{jk}(X_i)\}/n_j$, which measures the distance between the weight $w_{jk}(X_i)/n_j$ and the uniform weight $1/n_j$ to minimize the variability from heterogeneous weights, corresponding to the entropy balancing method (Hainmueller, 2012). By introducing Lagrange multiplier γ_{jk} , the estimated calibration weight is

$$\hat{w}_{jk}(X_i)/n_j = \exp\{\hat{\gamma}_{jk}^T g(X_i)\} / \sum_{D_i=k} \exp\{\hat{\gamma}_{jk}^T g(X_i)\} \text{ for } D_i = k, \quad (5)$$

where $\hat{\gamma}_{jk}$ solves the estimating equation

$$\sum_{D_i=k} \exp\{\gamma_{jk}^T g(X_i)\} \{g(X_i) - \bar{g}_j(X)\} = 0. \quad (6)$$

It is the Lagrangian dual of optimization problem (4) under the negative entropy loss, which implies a log-linear model $\log\{w_{jk}(X)\} = \alpha_{jk} + \gamma_{jk}^T g(X)$ (Cheng et al., 2020).

The DAC estimator based on the estimated outcome model $\hat{m}_k$ and weighting function $\hat{w}_{jk}$ is

$$\hat{\mu}_{k|\mathcal{I}} = \sum_{j \in \mathcal{I}} \frac{n_j}{\sum_{j \in \mathcal{I}} n_j} (A_{jk}^1 + A_{kj}^2), \quad (7)$$

$$\hat{\tau}_{(k,k')|\mathcal{I}} = \sum_{j \in \mathcal{I}} \frac{n_j}{\sum_{j \in \mathcal{I}} n_j} (A_{jk'}^1 - A_{jk}^1 + A_{k'j}^2 - A_{kj}^2). \quad (8)$$

where

$$A_{jk}^1 = \frac{1}{n_j} \sum_{i:D_i=j} \hat{m}_k(X_i), \quad A_{kj}^2 = \frac{1}{n_j} \sum_{i:D_i=k} \hat{w}_{jk}(X_i) \{Y_i - \hat{m}_k(X_i)\} \quad (9)$$

are two types of aggregated data. The first is the sample mean of $\hat{m}_k(X)$ at local site j , the second is a weighted average of residuals of $\hat{m}_k$ at local site k . And $\hat{\mu}_{k|\mathcal{I}}$ is a weighted average of the sums of these two aggregated data across sites in $\mathcal{I}$, enabling privacy-preserving estimation.

Remark 1. Using individual-level data from all single-arm trials, an estimator by solving $\mathbb{E}_P \{ \phi_{(k,k')|\mathcal{I}}(Z; P) \} = 0$ for $\tau_{(k,k')|\mathcal{I}}$ empirically based on respective estimators $\hat{m}_l(X)$ and $\hat{w}_{jl}(X)$ for $m_l(X)$ and $w_{jl}(X)$, $l = k, k', j \in \mathcal{I}$, has an explicit form $\hat{\tau}_{(k,k')|\mathcal{I}}^{\text{pool}} = \hat{\mu}_{k'|\mathcal{I}}^{\text{pool}} - \hat{\mu}_{k|\mathcal{I}}^{\text{pool}}$, where

$$\hat{\mu}_{k|\mathcal{I}}^{\text{pool}} = \frac{1}{\sum_{j \in \mathcal{I}} n_j} \sum_{i=1}^N \left\{ \hat{m}_k(X_i) I(D_i \in \mathcal{I}) + I(D_i = k) \{Y_i - \hat{m}_k(X_i)\} \sum_{j \in \mathcal{I}} \hat{w}_{jk}(X_i) \right\}.$$

Notably, our DAC estimator for $\tau_{(k,k')|\mathcal{I}}$ in (8) using only aggregated data replicates $\hat{\tau}_{(k,k')|\mathcal{I}}^{\text{pool}}$ using pooled data. Thus, our approach is as efficient and robust as $\hat{\tau}_{(k,k')|\mathcal{I}}^{\text{pool}}$ without information loss.

Remark 2. The DAC estimator for $\mu_{k|\mathcal{I}}$ in (7) differs from the distributed calibration weighting (DCW) estimator, $\hat{\mu}_{k|\mathcal{I}}^{\text{DCW}} := \sum_{j \in \mathcal{I}} n_j B_{kj}^2 / (\sum_{j \in \mathcal{I}} n_j)$, where $B_{kj}^2 = \sum_{i:D_i=k} \hat{w}_{jk}(X_i) Y_i / n_j$, by

$$\sum_{j \in \mathcal{I}} \frac{n_j}{\sum_{j \in \mathcal{I}} n_j} \left\{ \sum_{i:D_i=j} \frac{\hat{m}_k(X_i)}{n_j} - \sum_{i:D_i=k} \frac{\hat{w}_{jk}(X_i)}{n_j} \hat{m}_k(X_i) \right\}. \quad (10)$$

This term serves as the bias-correction component, ensuring that the DAC estimator which incorporates local average of $\hat{m}_k(X)$ remains consistent under a correctly specified outcome model even when the weighting function is misspecified, in contrast to the DCW estimator. This difference is zero if calibration constraints in (4) also balance $\hat{m}_k(X)$, making the DAC and DCW estimators equivalent when CW balances both summary statistics $g(X)$ and the fitted outcome model $\hat{m}_k(X)$. However, this requires an additional communication round to share local averages

of $\hat{m}_1(X), \dots, \hat{m}_K(X)$ at each site for obtaining $\hat{\tau}_{(k,k')|\mathcal{I}}^{\text{DCW}} = \hat{\mu}_{k'|\mathcal{I}}^{\text{DCW}} - \hat{\mu}_{k|\mathcal{I}}^{\text{DCW}}$, as shown in Figure S1 in the SM. To improve communication efficiency, our DAC method combines the OR with the CW approach that balances $g(X)$ alone, requiring two rounds as shown in Figure 1.

The DAC estimator in (8) combines the OR estimator with a bias correction term and is doubly robust consistent if either $m_k(X)$ or $w_{jk}(X)$ is consistently estimated; see Theorem 2 for details. The semiparametric efficiency bound for $\tau_{(k,k')|\mathcal{I}}$ is $B_{(k,k')|\mathcal{I}} = \mathbb{E}_P\{\phi_{(k,k')|\mathcal{I}}^2(Z; P)\}$, which means no estimator can have a smaller mean squared error than $B_{(k,k')|\mathcal{I}}$ in the local minimax sense (see Corollary 2.6 in Van der Vaart 2002). As stated in Theorem 3 in Section 5, if both outcome models and weighting functions are consistently estimated with a convergence rate of at least $N^{1/4}$, the asymptotic variance of $\hat{\tau}_{(k,k')|\mathcal{I}}$ attains $B_{(k,k')|\mathcal{I}}$, yielding the variance estimation:

$$\begin{aligned} \widehat{\text{Var}}(\hat{\tau}_{(k,k')|\mathcal{I}}) &= \hat{B}_{(k,k')|\mathcal{I}}/N \\ &= \frac{1}{(\sum_{l \in \mathcal{I}} n_l)^2} \sum_{l \in \mathcal{I}} A_{lkk'}^3 + \frac{1}{(\sum_{l \in \mathcal{I}} n_l)^2} \sum_{l,h \in \mathcal{I}} A_{k'lh}^4 + \frac{1}{(\sum_{l \in \mathcal{I}} n_l)^2} \sum_{l,h \in \mathcal{I}} A_{klh}^4 \\ &\quad + \frac{2}{(\sum_{l \in \mathcal{I}} n_l)^2} \mathbb{I}(k' \in \mathcal{I}) \sum_{l \in \mathcal{I}} A_{k'kl}^5 + \frac{2}{(\sum_{l \in \mathcal{I}} n_l)^2} \mathbb{I}(k \in \mathcal{I}) \sum_{l \in \mathcal{I}} A_{kk'l}^5 \quad (11) \\ &\quad + \frac{1}{\sum_{l \in \mathcal{I}} n_l} \hat{\tau}_{(k,k')|\mathcal{I}}^2 - \frac{2}{(\sum_{l \in \mathcal{I}} n_l)^2} \hat{\tau}_{(k,k')|\mathcal{I}} \left\{ \sum_{l \in \mathcal{I}} n_l (A_{lk'}^1 - A_{lk}^1) \right\} \\ &\quad - \mathbb{I}(k' \in \mathcal{I}) \frac{2\hat{\tau}_{(k,k')|\mathcal{I}}}{(\sum_{l \in \mathcal{I}} n_l)^2} \sum_{l \in \mathcal{I}} n_l A_{k'l}^2 + \mathbb{I}(k \in \mathcal{I}) \frac{2\hat{\tau}_{(k,k')|\mathcal{I}}}{(\sum_{l \in \mathcal{I}} n_l)^2} \sum_{l \in \mathcal{I}} n_l A_{kl}^2, \end{aligned}$$

where

$$\begin{aligned} A_{lkk'}^3 &= \sum_{D_i=l} \{\hat{m}_k(X_i) - \hat{m}_{k'}(X_i)\}^2, \quad A_{k'lh}^4 = \sum_{D_i=k} \hat{w}_{lk}(X_i) \hat{w}_{hk}(X_i) \{Y_i - \hat{m}_k(X_i)\}^2, \\ A_{kk'l}^5 &= \sum_{D_i=k} \hat{w}_{lk}(X_i) \{\hat{m}_k(X_i) - \hat{m}_{k'}(X_i)\} \{Y_i - \hat{m}_k(X_i)\} \end{aligned} \quad (12)$$

represent three additional types of aggregated data that contribute to the variance estimation.

Various machine learning methods, such as random forests (Wager and Athey, 2018) and neural

networks (Chen and White, 1999) can meet the convergence rate requirement of at least $N^{1/4}$ for nuisance function estimation. Therefore, our DAC estimator accommodates flexible semi-parametric or nonparametric methods for estimating nuisance functions while maintaining the parametric-rate consistency for $\hat{\tau}_{(k,k')|\mathcal{I}}$.

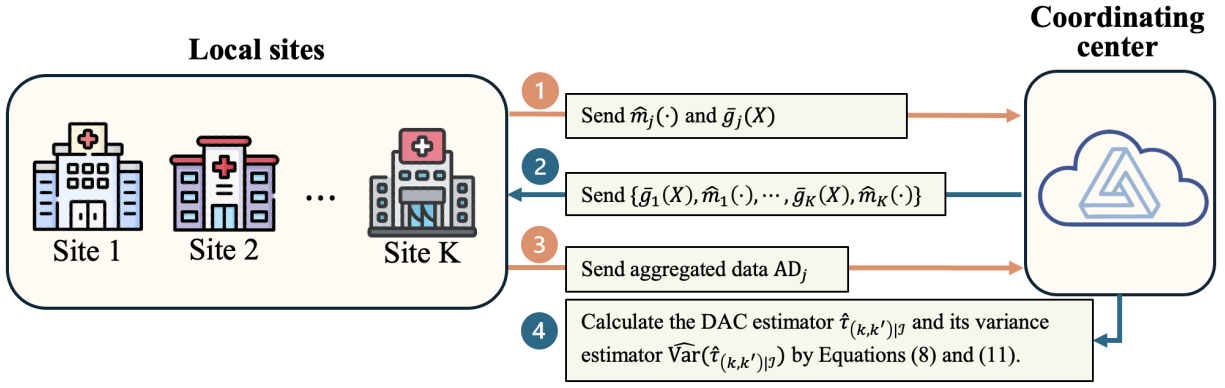
Remark 3. DAC estimators using nonparametric methods, which, although asymptotically efficient, can exhibit high variability when the sample size of local single-arm trials is limited; see, for example, the summary of several published applications regarding single-arm studies in Section 3.1 of Phillippo et al. (2016). Alternatively, parametric models can be employed.

We propose a procedure for implementing the DAC method with nonparametric estimation of nuisance parameters in a distributed network of K distinct single-arm trials. The goal is to estimate ATEs $\tau_{(k,k')|\mathcal{I}} = \mathbb{E}(Y(k') - Y(k)|D \in \mathcal{I})$ across all comparator groups (k, k') for $k, k' \in \{1, \dots, K\}$ and all target populations $\mathcal{I} \subseteq \{1, \dots, K\}$ simultaneously. Sites are allowed to interact with a coordinating center by transmitting specific aggregated data. Guided by prior knowledge and specific data characteristics, appropriate nonparametric method can be applied to local samples at site k to obtain $\hat{m}_k(X)$, such as random forests and the method of sieves, etc. The method of sieves can be used in the CW approach for estimating weighting functions with general sieve basis functions such as power series, splines, wavelets, or artificial neural networks (Chen, 2007). Each site j is required to provide the aggregated data

$$\text{AD}_j = \{A_{jk}^1, A_{jk}^2, A_{jkk'}^3, A_{jkk'}^4, A_{jkk'}^5, k, k' \in \{1, \dots, K\}\} \quad (13)$$

based on Equations (9) and (12) to calculate DAC estimators $\hat{\tau}_{(k,k')|\mathcal{I}}$ and variance estimators.

The DAC procedure with nonparametric nuisance parameter estimation is outlined in Figure 1 and Algorithm 1. First, each site $j \in 1, \dots, K$ locally regresses the outcome on covariates,



$\hat{m}_j(\cdot)$: Fitted outcome model at local site j .

$\bar{g}_j(X)$: summary statistics of covariates at local site j .

$\tau_{(k,k')|\mathcal{I}}$: ATE of Treatment k' versus Treatment k on the target population represented by $\mathcal{I}$.

AD_j is calculated by Equation (13) with $\{\hat{m}_1(\cdot), \dots, \hat{m}_K(\cdot)\}$ and weights obtained from calibration weighting using $\{\bar{g}_1(X), \dots, \bar{g}_K(X)\}$.

Figure 1: Workflow of the algorithm on the distributed augmented calibration weighting (DAC) estimation under nonparametric working models for the collaborative indirect comparisons in a distributed network where each site conducts a single-arm trial on a distinct treatment.

Algorithm 1: DAC with nonparametric methods for estimating working models

- for** site $j = 1$ to K **do**
 Regress the outcome on covariates at site j to obtain the fitted outcome model function $\hat{m}_j(\cdot)$.
 Send the summary statistics of covariates $\bar{g}_j(X)$ and the fitted outcome model function $\hat{m}_j(\cdot)$ to the coordinating center.
 - The coordinating center collects and sends $\{\bar{g}_1(X), \hat{m}_1(\cdot), \bar{g}_2(X), \hat{m}_2(\cdot), \dots, \bar{g}_K(X), \hat{m}_K(\cdot)\}$ to all sites.
 - for** site $j = 1$ to K **do**
 Obtain the calibration weights $\hat{w}_{kj}(X_s)/n_k$ via the calibration weighting with $\bar{g}_k(X)$ by solving (5) and (6) for $k \neq j$; set $\hat{w}_{jj}(X_s)/n_j = 1/n_j$.
 Compute the aggregated data AD_j in (13) by Equations (9) and (12) and send them to the coordinating center.
 - For $k, k' \in \{1, \dots, K\}$ and $\mathcal{I} \subseteq \{1, \dots, K\}$, the coordinating center calculates the DAC estimator $\hat{\tau}_{(k,k')|\mathcal{I}}$ of the ATE by (8) and its variance estimate by (11).
-

and transmits the fitted outcome model function $\hat{m}_j(\cdot)$ and summary statistics of covariates $\bar{g}_j(X)$ to the coordinating center. In our algorithm, transmitting a function means sharing the function structure and parameters; for example, a spline method requires sending the order and knots of the spline function, and estimated coefficients of the basis functions. Second, the coordinating center collects these fitted outcome models and summary statistics, and sends them to each site. Third,

each site j uses summary statistics $\bar{g}_k(X)$ from other sites $k \neq j$ to perform calibration weighting $K-1$ times to generate $K-1$ sets of weights $\{\hat{w}_{kj}(X_s), D_s = j\}_{k \neq j}$ by solving Equations (5) and (6). Then aggregated data AD_j in Equation (13) is computed and sent to the coordinating center. Finally, the coordinating center utilizes aggregated data from all sites to compute DAC estimators $\hat{\tau}_{(k,k')|\mathcal{I}}$ of the ATEs and their variance estimators by Equations (8) and (11) for $k, k' \in 1, \dots, K$ and $\mathcal{I} \subseteq \{1, \dots, K\}$. This algorithm estimates pairwise ATEs across target populations in just two communication rounds, producing identical results to pooled individual-level analysis, as DAC estimators are equivalently expressed in aggregated data form.

Remark 4. Section S5 of the SM presents the distributed outcome regression (DOR) estimation in a distributed network, requiring two communication rounds in four steps. Under linear models, DOR reduces to a single round. In contrast, the DAC approach combines calibration weighting and outcome regression, enhancing robustness without increasing communication costs.

4 Distributed Doubly Robust Inference

Section 3 presents the general DAC method for doubly robust estimation rather than doubly robust inference. In this section, we strengthen the DAC to achieve doubly robust inference under parametric working models.

With parametric working models, the DAC estimator is consistent if either the outcome model or the weighting function is correctly specified. However, the choice of nuisance parameter estimators affects the asymptotic variance of $\hat{\tau}_{(k,k')|\mathcal{I}}$ under potential model misspecification; see Section S3 of the SM for details. Section S4 of the SM provides complex variance estimation for DAC under general parametric nuisance parameter estimation methods with potential misspecification, requiring transmitting more aggregated data as shown in Equation (S.8). Fortunately, valid doubly robust inference with parametric working models can be achieved in

various contexts (Kim and Haziza, 2014; Vermeulen and Vansteelandt, 2015; Tan, 2020; Ning et al., 2020), ensuring a generic variance estimation form if at least one working model is correctly specified. Let the working model for outcome regression and weighting functions be $\{m_k(X) = m(X; \beta_k) : \beta_k \in \mathbb{R}^p\}$ and $\{w_{jk}(X) = w(X; \gamma_{jk}) : \gamma_{jk} \in \mathbb{R}^q\}$, respectively. And $\Gamma_{\mathcal{I}k} = (\gamma_{jk}, j \in \mathcal{I})$. Denote nuisance parameter estimators as $\hat{\beta}_k$ and $\hat{\gamma}_{jk}$. And $\hat{\mu}_{k|\mathcal{I}}$ is the DAC estimator in (7) using estimated working models $m(X; \hat{\beta}_k)$ and $w(X; \hat{\gamma}_{jk})$. To achieve doubly robust inference, Vermeulen and Vansteelandt (2015) suggest to adapt the nuisance parameter estimation to the EIF by solving the following estimating equations for estimates $\hat{\beta}_k^{\text{BR}}$ and $\hat{\Gamma}_{\mathcal{I}k}^{\text{BR}}$,

$$\hat{\mathbb{E}}_N \left\{ \frac{\partial \varphi_{k|\mathcal{I}}}{\partial \beta_k}(Z; \mu_{k|\mathcal{I}}, \hat{\beta}_k^{\text{BR}}, \hat{\Gamma}_{\mathcal{I}k}^{\text{BR}}) \right\} = 0, \quad \hat{\mathbb{E}}_N \left\{ \frac{\partial \varphi_{k|\mathcal{I}}}{\partial \gamma_{jk}}(Z; \mu_{k|\mathcal{I}}, \hat{\beta}_k^{\text{BR}}, \hat{\Gamma}_{\mathcal{I}k}^{\text{BR}}) \right\} = 0, j \in \mathcal{I}, \quad (14)$$

where $\hat{\mathbb{E}}_N f(Z) = \sum_{i=1}^N f(Z_i)/N$ represents empirical mean. The DAC estimator $\hat{\mu}_{k|\mathcal{I}}^{\text{BR}}$ using $m(X; \hat{\beta}_k^{\text{BR}})$ and $w(X; \hat{\gamma}_{jk}^{\text{BR}})$ has a generic asymptotic variance $\mathbb{E}_P \{ \varphi_{k|\mathcal{I}}^2(Z; \mu_{*k|\mathcal{I}}^{\text{BR}}, \beta_{*k}^{\text{BR}}, \Gamma_{*\mathcal{I}k}^{\text{BR}}) \}$, where $\mu_{*k|\mathcal{I}}^{\text{BR}}, \beta_{*k}^{\text{BR}}$ and $\Gamma_{*\mathcal{I}k}^{\text{BR}}$ are probabilistic limits, and the squared first-order asymptotic bias $\text{bias}^2(\beta_k, \Gamma_{\mathcal{I}k}; \mu_{k|\mathcal{I}}) = [\mathbb{E}\{\varphi_{k|\mathcal{I}}(Z; \mu_{k|\mathcal{I}}, \beta_k, \Gamma_{\mathcal{I}k})\}]^2$ is minimized in the working model class.

We apply technique (14) to the DAC under log-linear weighting functions and linear outcome models, i.e., $m_\ell(X; \beta_\ell) = \beta_\ell^T g(X)$ and $w_{j\ell}(X; \gamma_{j\ell}) = \exp(\gamma_{j\ell}^T a(X))$ for $\ell = k, k'$. Components in both sets of basis functions $g(X) = \{g_1(X), \dots, g_p(X)\}^T$ and $a(X) = \{a_1(X), \dots, a_q(X)\}^T$ are linearly independent. Then, nuisance parameters are estimated by solving

$$\frac{1}{N} \sum_{i=1}^N I(D_i = j) g(X_i) - \frac{1}{N} \sum_{i=1}^N I(D_i = \ell) \exp(\gamma_{j\ell}^T a(X_i)) g(X_i) = 0, \quad (15)$$

$$\frac{1}{N} \sum_{i=1}^N \left\{ I(D_i = \ell) \sum_{j \in \mathcal{I}} \exp(\gamma_{j\ell}^T a(X_i)) \{Y_i - \beta_\ell^T g(X_i)\} a(X_i) \right\} = 0. \quad (16)$$

Identification of nuisance parameters requires $p = q$ and full-rank conditions on the matrices $\mathbb{E}\{I(D = \ell) \sum_{j \in \mathcal{I}} \exp(\gamma_{j\ell}^T a(X)) a(X) g^T(X)\}$ for $\mathcal{I} \subseteq \{1, \dots, K\}$. Equation (15) aligns with

the CW method to balance $g(X)$. The solution to (15) and (16) by generalized method of moments (GMM) is denoted as $\hat{\beta}_{\ell|\mathcal{I}}^{\text{BR}}$ and $\hat{\Gamma}_{\mathcal{I}\ell}^{\text{BR}} = (\hat{\gamma}_{j\ell}^{\text{BR}}, j \in \mathcal{I})$. Then, the DAC estimator $\hat{\mu}_{\ell|\mathcal{I}}^{\text{BR}}$ is the estimator in Equation (7) with $\hat{w}_{j\ell}(X) = \exp\{(\hat{\gamma}_{j\ell}^{\text{BR}})^T a(X)\}$ and $\hat{m}_{\ell}(X) = (\hat{\beta}_{\ell|\mathcal{I}}^{\text{BR}})^T g(X)$, and the variance estimate for $\hat{\tau}_{(k,k')|\mathcal{I}}^{\text{BR}} = \hat{\mu}_{k'|\mathcal{I}}^{\text{BR}} - \hat{\mu}_{k|\mathcal{I}}^{\text{BR}}$ is $\hat{\mathbb{E}}_N\{\phi_{(k,k')|\mathcal{I}}^2(Z; \hat{\tau}_{(k,k')|\mathcal{I}}^{\text{BR}}, \hat{\beta}_{\ell|\mathcal{I}}^{\text{BR}}, \hat{\Gamma}_{\mathcal{I}\ell}^{\text{BR}}, \ell = k, k')\}/N$.

Remark 5. Notably, if $g(X) = a(X)$, Equation (15) aligns with the entropy balancing method to balance $g(X)$, suggesting solving Equation (6) to obtain the estimate $\hat{\gamma}_{j\ell}^{\text{BR}}$. Equation (16) indicates performing weighted least squares with weights $\sum_{j \in \mathcal{I}} \exp\{(\hat{\gamma}_{j\ell}^{\text{BR}})^T a(X_i)\}$ at site l to obtain $\hat{\beta}_{\ell|\mathcal{I}}^{\text{BR}}$. And $\hat{\tau}_{(k,k')|\mathcal{I}}^{\text{BR}}$ simplifies to an outcome regression estimator with $\hat{m}_{\ell}(X) = \{\hat{\beta}_{\ell|\mathcal{I}}^{\text{BR}}\}^T g(X)$, as components $\{A_{lj}^2, l \in \{k, k'\}, j \in \mathcal{I}\}$ of the bias correction term in (9) are weighted averages of residuals from weighted least squares which equal zero. From Equation (10), the DAC estimator $\hat{\tau}_{(k,k')|\mathcal{I}}^{\text{BR}}$ is also equivalent to the DCW estimator $\hat{\tau}_{(k,k')|\mathcal{I}}^{\text{DCW}}$ with $g(X)$ to be calibrated, as balancing $g(X)$ automatically balances $\hat{m}_k(X)$ under linear outcome models. Thus, in this case, Equation (16) is actually unnecessary for the point estimation of $\tau_{(k,k')|\mathcal{I}}$, as the DAC estimator with $\hat{w}_{j\ell}(X) = \exp\{(\hat{\gamma}_{j\ell}^{\text{BR}})^T g(X)\}$ remains the DCW estimator regardless of how β_{ℓ} is estimated. This finding not only confirms doubly robust consistency (Zhao and Percival, 2016) but also reveals the doubly robust inference property of the entropy balancing estimator with respect to the linear outcome regression model and log-linear weighting function with same basis functions.

Let the aggregated data from site j be

$$\text{AD}_j^{\text{BR}} = \{\hat{\beta}_{j|\mathcal{I}}^{\text{BR}}, O_j^1, O_{jk|\mathcal{I}}^2, O_j^3, O_{jkk'|\mathcal{I}}^4, O_{jk|\mathcal{I}}^5, k, k' \in \{1, \dots, K\}, \mathcal{I} \subseteq \{1, \dots, K\}\}, \quad (17)$$

where $O_j^1 = \sum_{i:D_i=j} g(X_i)$, $O_{jk|\mathcal{I}}^2 = \sum_{i:D_i=j} \exp\{\hat{\gamma}_{kj}^T a(X_i)\} \{Y_i - (\hat{\beta}_{j|\mathcal{I}}^{\text{BR}})^T g(X_i)\}$, $O_j^3 = \sum_{i:D_i=j} g(X_i)g^T(X_i)$, $O_{jlk|\mathcal{I}}^4 = \sum_{i:D_i=j} \exp\{(\hat{\gamma}_{lj} + \hat{\gamma}_{hk})^T a(X_i)\} \{Y_i - (\hat{\beta}_{j|\mathcal{I}}^{\text{BR}})^T g(X_i)\}^2$ and $O_{jk|\mathcal{I}}^5 = \sum_{D_i=j} \exp\{\hat{\gamma}_{kj}^T a(X_i)\} \{Y_i - (\hat{\beta}_{j|\mathcal{I}}^{\text{BR}})^T g(X_i)\} g(X_i)$. The DAC estimator in (8) under log-linear

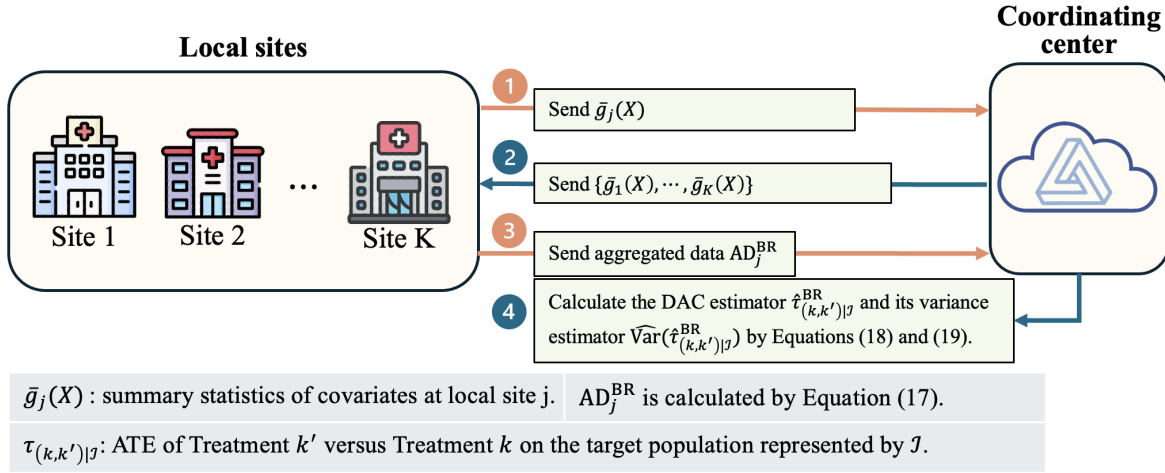


Figure 2: Workflow of the distributed augmented calibration weighting (DAC) estimation under a log-linear weighting function and linear outcome regression model to achieve doubly robust inference in a distributed network of multiple single-arm trials.

weighting functions and linear outcome models, along with its variance estimator, simplifies to

$$\hat{\tau}_{(k,k')|I}^{BR} = \frac{1}{\sum_{j \in I} n_j} \{ (\hat{\beta}_{k'|I}^{BR} - \hat{\beta}_{k|I}^{BR})^T (\sum_{j \in I} O_j^1) + \sum_{j \in I} O_{k'j|I}^2 - \sum_{j \in I} O_{kj|I}^2 \}, \quad (18)$$

$$\begin{aligned} \widehat{Var}(\hat{\tau}_{(k,k')|I}^{BR}) &= \frac{2}{(\sum_{l \in I} n_l)^2} \hat{\tau}_{(k,k')|I}^{BR} (\hat{\beta}_{k'|I}^{BR} - \hat{\beta}_{k|I}^{BR})^T (\sum_{l \in I} O_l^1) + \frac{1}{\sum_{l \in I} n_l} \{ \hat{\tau}_{(k,k')|I}^{BR} \}^2 \\ &\quad + \frac{2\hat{\tau}_{(k,k')|I}^{BR}}{(\sum_{l \in I} n_l)^2} \{ \mathbb{I}(k \in I) \sum_{l \in I} O_{kl|I}^2 - \mathbb{I}(k' \in I) \sum_{l \in I} O_{k'l|I}^2 \}, \\ &\quad + \frac{1}{(\sum_{l \in I} n_l)^2} (\hat{\beta}_{k'|I}^{BR} - \hat{\beta}_{k|I}^{BR})^T (\sum_{l \in I} O_l^3) (\hat{\beta}_{k'|I}^{BR} - \hat{\beta}_{k|I}^{BR}) + \frac{\sum_{l,h \in I} (O_{k'lh|I}^4 + O_{klh|I}^4)}{(\sum_{l \in I} n_l)^2} \\ &\quad + \frac{2}{(\sum_{l \in I} n_l)^2} (\hat{\beta}_{k'|I}^{BR} - \hat{\beta}_{k|I}^{BR})^T \{ \mathbb{I}(k' \in I) \sum_{l \in I} O_{k'l|I}^5 - \mathbb{I}(k \in I) \sum_{l \in I} O_{kl|I}^5 \}. \end{aligned} \quad (19)$$

The variance estimation in (19) holds even if one working model is misspecified. Figure 2 and Algorithm 2 outline a four-step DAC estimation procedure with two communication rounds for this case. For variable selection of $a(X)$ and $g(X)$, we can consider variables that exhibit distinct distributions across sites and are important for local outcome models.

Generally, under generalized linear working models, $m_\ell(X)$ and $w_{j\ell}(X)$ for $\ell = k, k'$ are modeled as $\{h_m(\beta_\ell^T g(X)) : \beta \in [-M, M]^p\}$ and $\{h_w(\gamma_{j\ell}^T a(X)) : \gamma_{j\ell} \in [-M, M]^q\}$, respec-

Algorithm 2: DAC under a log-linear weighting function and a linear outcome regression model to achieve doubly robust inference

1. **for** site $j = 1$ to K **do**
 └ Send the summary statistics of covariates $\bar{g}_j(X)$ to the coordinating center.
 2. The coordinating center collects and sends $\{\bar{g}_1(X), \bar{g}_2(X), \dots, \bar{g}_K(X)\}$ to all sites.
 3. **for** site $\ell = 1$ to K **do**
 └ Solve (15) and (16) by GMM to obtain $\hat{\gamma}_{j\ell}^{\text{BR}}$ and $\hat{\beta}_{\ell|\mathcal{I}}^{\text{BR}}$ for $j \in \{1, \dots, K\}$ and $\mathcal{I} \subseteq \{1, \dots, K\}$, using local data at site ℓ and $\{\bar{g}_1(X), \bar{g}_2(X), \dots, \bar{g}_K(X)\}$.
 └ Compute the aggregated data $\text{AD}_{\ell}^{\text{BR}}$ in (17) and send them to the coordinating center.
 4. For $k, k' \in \{1, \dots, K\}$ and $\mathcal{I} \subseteq \{1, \dots, K\}$, the coordinating center calculates the DAC estimator $\hat{\tau}_{(k,k')|\mathcal{I}}^{\text{BR}}$ of ATE by (18) and its variance estimate by (19).
-

tively, where $h_m(\cdot)$ and $h_w(\cdot)$ are link functions, and M is a large positive constant to ensure compactness of the parameter space. This setup supports various outcome types, including continuous and categorical data, by appropriate choices of $h_m(\cdot)$. To ensure valid weights, $h_w(\cdot)$ is constrained to be positive. The technique in (14) motivates the following estimating equations:

$$\frac{1}{N} \sum_{i=1}^N h'_m(\beta_\ell^T g(X_i)) g(X_i) \{I(D_i = j) - I(D_i = \ell) h_w(\gamma_{j\ell}^T a(X_i))\} = 0 \quad (20)$$

$$\frac{1}{N} \sum_{i=1}^N \left\{ I(D_i = \ell) \sum_{j \in \mathcal{I}} h'_w(\gamma_{j\ell}^T a(X_i)) \{Y_i - h_m(\beta_\ell^T g(X_i))\} a(X_i) \right\} = 0, \quad (21)$$

where $j \in \mathcal{I}$ and $h'_w(\cdot)$ denotes the first-order derivative of $h_w(\cdot)$. And some conditions (Assumption 7 in the SM) are required for the identification of nuisance parameters in above estimating equations. Equation (20) corresponds to calibration weighting with variables $h'_m(\beta_\ell^T g(X))g(X)$ to be calibrated. The solution to (20) and (21) is $(\hat{\beta}_\ell^{\text{BR}}, \hat{\gamma}_{j\ell}^{\text{BR}}, j \in \mathcal{I})$. Then the bias-reduced DAC estimator $\hat{\mu}_{\ell|\mathcal{I}}^{\text{BR}}$ is the estimator in Equation (7) with estimated working models $\hat{w}_{j\ell}(X) = h_w(\{\hat{\gamma}_{j\ell}^{\text{BR}}\}^T a(X))$ and $\hat{m}_\ell(X) = h_m(\{\hat{\beta}_\ell^{\text{BR}}\}^T g(X))$. As stated in Theorem 4, the treatment effect estimator $\hat{\tau}_{(k,k')|\mathcal{I}}^{\text{BR}} = \hat{\mu}_{k'|\mathcal{I}}^{\text{BR}} - \hat{\mu}_{k|\mathcal{I}}^{\text{BR}}$ is asymptotically normal with a simple and generic form of asymptotic variance, even when one of the working models is misspecified. Moreover, it minimizes the squared first-order asymptotic bias when both models are misspecified. For com-

plex models, solving Equations (20) and (21) may require iterative methods, and developing a communication-efficient computation scheme is left for future work.

Remark 6. The DAC estimator introduced in Section 3 is a general estimator accommodating both parametric and nonparametric estimation of the two nuisance working models. The bias-reduced DAC estimator is a specialized version under parametric nuisance models, achieving doubly robust inference with minimal squared first-order asymptotic bias. Its key technique is to estimate nuisance parameters by enforcing that the derivative of the efficient influence function equals zero (Equations (15) and (16)), which makes outcome model estimation depend on calibration weights and thus requires weights to be estimated first. A direct implementation of Algorithm 1 using the aggregated data in Equation (13) would require an extra communication round. To avoid this, we rewrite estimators with an alternative aggregated data structure in Equation (17), and perform bias-reduced DAC using Algorithm 2 with two communication rounds. In summary, Algorithm 1 applies to nonparametric nuisance estimation, whereas Algorithm 2 enables bias-reduced inference under parametric models with minimal communication cost.

5 Asymptotic Properties

This section formally illustrates the EIFs for target estimands and asymptotic properties of DAC and bias-reduced DAC estimators. The following theorem presents the EIFs of $\mu_{k|\mathcal{I}}$ and $\tau_{(k,k')|\mathcal{I}}$.

Theorem 1. *Under Assumptions 1–3, the efficient influence function for $\mu_{k|\mathcal{I}}$ is*

$$\varphi_{k|\mathcal{I}}(Z; P) = \frac{1}{\mathbb{P}(D \in \mathcal{I})} \left[I(D \in \mathcal{I}) \{m_k(X) - \mu_{k|\mathcal{I}}\} + I(D = k) \{Y - m_k(X)\} \sum_{j \in \mathcal{I}} w_{jk}(X) \right],$$

where P in $\varphi_{k|\mathcal{I}}(Z; P)$ indicates that all the parameters $\mu_{k|\mathcal{I}}$, $m_k(X)$, and $w_{jk}(X)$ are defined

under the distribution P . The efficient influence function for $\tau_{(k,k')|\mathcal{I}}$ is

$$\phi_{(k,k')|\mathcal{I}}(Z; P) = \varphi_{k'|\mathcal{I}}(Z; P) - \varphi_{k|\mathcal{I}}(Z; P).$$

Remark 7. Extensions for EIFs in cases of repeated treatments across different sites or multiple treatments assigned within a local site are detailed in Section S6 of the SM.

To formally state doubly robust consistency of the proposed distributed estimator, let $\mathcal{F}$ and $\mathcal{G}$ be the working model classes for estimating outcome regression and weighting function, respectively. The squared $L_2(P)$ -norm is denoted as $\|f\|_P^2 = \mathbb{E}_P\{f(Z)\}^2$. Let $w_{\mathcal{I}\ell} = (w_{j\ell}, j \in \mathcal{I})^T$.

Theorem 2. *Under Assumptions 1–3 and some additional technical conditions (Assumptions 4 and 5 in Section S1.3 of the SM), the DAC estimator $\hat{\tau}_{(k,k')|\mathcal{I}}$ in (8) based on $\hat{m}_\ell$ and $\hat{w}_{j\ell}$ has doubly robust consistency with respect to $(\mathcal{F}, \mathcal{G})$ in the sense that it is consistent if either $m_\ell(X) = m_\ell^*(X)$ or $w_{\mathcal{I}\ell}(X) = w_{\mathcal{I}\ell}^*(X)$ holds for $\ell = k$ and k' , where $m_\ell^*(X)$ and $w_{j\ell}^*(X)$ are probabilistic limits of $\hat{m}_\ell(X)$ and $\hat{w}_{j\ell}(X)$, i.e., $\|\hat{m}_\ell - m_\ell^*\|_{P \rightarrow P} \rightarrow 0$ and $\|\hat{w}_{j\ell} - w_{j\ell}^*\|_{P \rightarrow P} \rightarrow 0$.*

Under some assumptions on convergence rates for working models, the following theorem establishes the basis for inference on the DAC estimator $\hat{\tau}_{(k,k')|\mathcal{I}}$ in (8).

Theorem 3. *Under Assumptions 1–3 and additional conditions (Assumptions 4–6 in the SM),*

$$\sqrt{N} \{ \hat{\tau}_{(k,k')|\mathcal{I}} - \tau_{(k,k')|\mathcal{I}} \} \rightarrow_d N(0, B_{(k,k')|\mathcal{I}})$$

if, for $\ell = k, k'$, $\|\hat{m}_\ell - m_\ell\|_P = o_{\mathbb{P}}(1)$, $\|\hat{w}_{\mathcal{I}\ell} - w_{\mathcal{I}\ell}\|_P = o_{\mathbb{P}}(1)$, and $\|\hat{m}_\ell - m_\ell\|_P \times \|\hat{w}_{\mathcal{I}\ell} - w_{\mathcal{I}\ell}\|_P = o_{\mathbb{P}}(N^{-1/2})$. Moreover, $B_{(k,k')|\mathcal{I}}$ can be consistently estimated by the sample second moment of the efficient influence function; that is, $\hat{B}_{(k,k')|\mathcal{I}} \rightarrow_P B_{(k,k')|\mathcal{I}}$ with $\hat{B}_{(k,k')|\mathcal{I}} = \frac{1}{N} \sum_{i=1}^N \{ \phi_{(k,k')|\mathcal{I}}(Z_i; \hat{P}) \}^2$, where $\hat{P}$ represents that nuisance parameters in $\phi_{(k,k')|\mathcal{I}}$ are plugged in by their respective estimators.

We demonstrate that the bias-reduced DAC estimator $\hat{\tau}_{(k,k')|\mathcal{I}}^{\text{BR}}$ under generalized linear models exhibits the doubly robust inference and bias reduction properties in the following theorem.

Theorem 4. *Under Assumptions 1–3 and some additional technical conditions (Assumptions 4, 6 and 7 in the SM), the following two statements hold: The solution to (20) and (21) is $\hat{\beta}_\ell^{\text{BR}}$ and $\hat{\Gamma}_{\mathcal{I}\ell}^{\text{BR}} = (\hat{\gamma}_{j\ell}^{\text{BR}}, j \in \mathcal{I})$, and $\beta_{*\ell}^{\text{BR}}$ and $\Gamma_{*\mathcal{I}\ell}^{\text{BR}}$ are their probabilistic limits for $\ell \in \{k, k'\}$.*

(1) *If for $\ell = k, k'$, either $m_\ell(X) = h_m(\beta_\ell^T g(X))$ or $w_{j\ell}(X) = h_w(\gamma_{j\ell}^T a(X))$ holds for some $\beta_\ell \in [-M, M]^p, \gamma_{j\ell} \in [-M, M]^q$, then*

$$\sqrt{N}\{\hat{\tau}_{(k,k')|\mathcal{I}}^{\text{BR}} - \tau_{(k,k')|\mathcal{I}}\} \rightarrow_d N(0, \mathbb{E}_P\{\phi_{(k,k')|\mathcal{I}}^2(Z; \tau_{(k,k')|\mathcal{I}}, \beta_{*\ell}^{\text{BR}}, \Gamma_{*\mathcal{I}\ell}^{\text{BR}}, \ell = k, k')\}),$$

and the asymptotic variance can be consistently estimated by $\frac{1}{N} \sum_{i=1}^N \{\phi_{(k,k')|\mathcal{I}}(Z_i; \hat{\tau}_{(k,k')|\mathcal{I}}^{\text{BR}}, \hat{\beta}_\ell^{\text{BR}}, \hat{\Gamma}_{\mathcal{I}\ell}^{\text{BR}}, \ell = k, k')\}^2$.

(2) *The squared first-order asymptotic bias of $\hat{\tau}_{(k,k')|\mathcal{I}}$ under working models $\{h_m(\beta_\ell^T g(X)) : \beta_\ell \in [-M, M]^p\}$ and $\{h_w(\gamma_{j\ell}^T a(X)) : \gamma_{j\ell} \in [-M, M]^q\}$ is*

$$\text{bias}^2(\beta_\ell, \Gamma_{\mathcal{I}\ell}, \ell = k, k'; \tau_{(k,k')|\mathcal{I}}) = [\mathbb{E}_P\{\phi_{(k,k')|\mathcal{I}}(Z; \tau_{(k,k')|\mathcal{I}}, \beta_\ell, \Gamma_{\mathcal{I}\ell}, \ell = k, k')\}]^2.$$

We have $\text{bias}^2(\beta_{\ell}^{\text{BR}}, \Gamma_{*\mathcal{I}\ell}^{\text{BR}}, \ell = k, k'; \tau_{(k,k')|\mathcal{I}}) \leq \text{bias}^2(\beta_\ell, \Gamma_{\mathcal{I}\ell}, \ell = k, k'; \tau_{(k,k')|\mathcal{I}})$ for all β_ℓ and $\Gamma_{\mathcal{I}\ell}$ within the parameter space $(-M, M)^{2q \times |\mathcal{I}| + 2p}$.*

6 Simulation

We conduct simulations to evaluate the finite sample performance of DAC estimators with linear outcome models and entropy balancing (Figure 2), and compare them to DOR estimators which use ordinary least squares to estimate coefficients of linear models (Figure S3 in the SM).

The number of sites in the distributed network is set to $K = 4$, with a total sample size of $N \in \{800, 2400, 4000\}$ subjects across all sites. Baseline covariates $X = (X_1, X_2, X_3)^T$ are generated by a multivariate normal distribution with mean $(0.6, 0.6, 0.6)$, variances $(0.64, 1, 1.44)$ and correlations $\rho_{X_1X_2} = \rho_{X_2X_3} = 0.001$, $\rho_{X_1X_3} = 0$. The site index D is generated by the following two strategies using the multinoulli distribution with the parameters of probabilities being

$$P(D = k | X) = \exp\{\zeta_{k1}X_1 + \zeta_{k2}X_2 + \zeta_{k3}X_3\} / \sum_{i=1}^K \exp\{\zeta_{i1}X_1 + \zeta_{i2}X_2 + \zeta_{i3}X_3\} \text{ and}$$

$$P(D = k | X) = \exp\{\nu_{k0} + \nu_{k1}X_1^2 + \nu_{k2}X_2^2 + \nu_{k3}X_3^2\} / \sum_{i=1}^K \exp\{\nu_{i0} + \nu_{i1}X_1^2 + \nu_{i2}X_2^2 + \nu_{i3}X_3^2\},$$

respectively, where $\zeta_{11} = 0.2, \zeta_{12} = -0.2, \zeta_{13} = -0.1, \zeta_{21} = 0.36, \zeta_{22} = -0.28, \zeta_{23} = -0.16, \zeta_{31} = 0.5, \zeta_{32} = -0.4, \zeta_{33} = -0.2, \zeta_{41} = 0.65, \zeta_{42} = -0.5, \zeta_{43} = -0.25, \nu_{10} = -0.4, \nu_{11} = 1.6, \nu_{12} = 1, \nu_{13} = 0.25, \nu_{20} = -0.3, \nu_{21} = 1.4, \nu_{22} = 0.7, \nu_{23} = 0.5, \nu_{30} = -0.5, \nu_{31} = 1.55, \nu_{32} = 1.1, \nu_{33} = 0.3, \nu_{40} = -0.1, \nu_{41} = 1.5, \nu_{42} = 0.6, \nu_{43} = 0.4$. Under both strategies, each site accounts for approximately 25% of the overall population. Similarly, two strategies are used to generate the observed outcomes Y with normal distribution,

$$Y | (X, D = k) \sim N(\theta_{k0} + \theta_{k1}X_1 + \theta_{k2}X_2 + \theta_{k3}X_3, 0.04|X_2|^{0.4}) \text{ and}$$

$$Y | (X, D = k) \sim N(\psi_{k0} + \psi_{k1}X_1^2 + \psi_{k2}X_2^2 + \psi_{k3}X_3^2, 0.04|X_2|^{0.4}),$$

respectively, where $\theta_{10} = 1, \theta_{11} = 0.3, \theta_{12} = 0.2, \theta_{13} = 0.2, \theta_{20} = 4, \theta_{21} = 0.5, \theta_{22} = 0.5, \theta_{23} = 0.6, \theta_{30} = 7, \theta_{31} = 0.7, \theta_{32} = 0.8, \theta_{33} = 1, \theta_{40} = 10, \theta_{41} = 0.9, \theta_{42} = 1.1, \theta_{43} = 1.5, \psi_{10} = -0.5, \psi_{11} = -1, \psi_{12} = -1, \psi_{13} = -0.5, \psi_{20} = 0.2, \psi_{21} = 0.5, \psi_{22} = 1, \psi_{23} = 0.5, \psi_{30} = 1.5, \psi_{31} = 1, \psi_{32} = 2, \psi_{33} = 1, \psi_{40} = 3, \psi_{41} = 1.5, \psi_{42} = 2.5, \psi_{43} = 2$. We evaluate two estimation methods under four data-generating scenarios outlined in Table 1 to estimate $\tau_{(k,k')|I}$, simulating 600 replicates for each scenario.

Table 1: Descriptions of (a) four data-generating scenarios in simulations, and (b) two estimation methods, their types, the random vector $g(X)$ to be calibrated in the calibration weighting and the working model $m_k(X)$ for outcome regression. A “—” indicates the working model is not involved in the estimation method.

(a) Data-generating scenarios			
Scenario	Details		
(i)	generating D and Y both by the first strategy, respectively		
(ii)	generating D by the first strategy but Y by the second		
(iii)	generating D by the second strategy while Y by the first		
(iv)	generating D and Y both by the second strategy, respectively		

(b) Estimation methods			
Method	Estimator type	Calibrated random vector $g(X)$	Outcome model $m_k(X)$
DOR	outcome regression	—	linear model
DAC	augmented calibration weighting	$(X_1, X_2, X_3)^T$	linear model

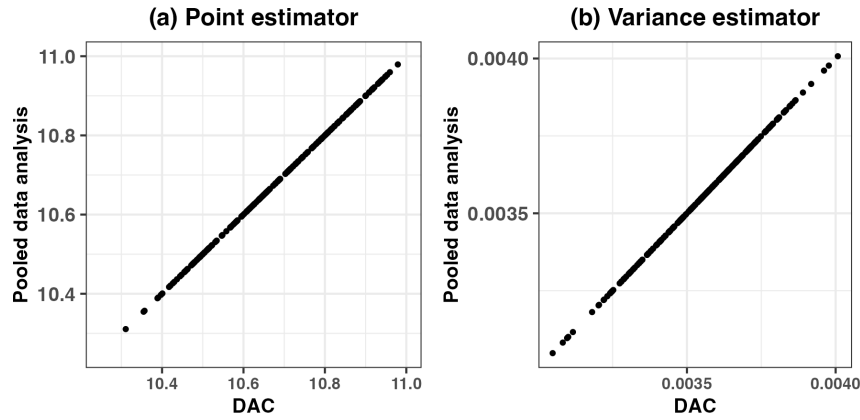


Figure 3: Scatterplots of (a) point and (b) variance estimates from the augmented calibration weighting estimation obtained in distributed setting (DAC) and in pooled data analysis.

Table 2: Simulation results of comparing Treatment 4 versus Treatment 1 on the overall population in a distributed network with four single-arm trials under three different values of the total sample size N based on 600 Monte Carlo replications.

Scenario	True value	Method	$N = 800$				$N = 2400$				$N = 4000$			
			Bias $\times 10^3$	SD	ESE	CP	Bias $\times 10^3$	SD	ESE	CP	Bias $\times 10^3$	SD	ESE	CP
(i) 10.68		DOR	-0.57	0.06	0.07	97.17	-0.83	0.04	0.04	94.33	-0.29	0.03	0.03	96.33
		DAC	-0.59	0.06	0.07	97.17	-0.83	0.04	0.04	94.67	-0.28	0.03	0.03	96.17
(ii) 15.26		DOR	-273.60	0.47	0.48	89.83	-258.13	0.29	0.28	81.83	-237.96	0.22	0.22	78.33
		DAC	-9.33	0.47	0.48	95.50	-10.86	0.29	0.28	94.00	4.45	0.22	0.22	94.83
(iii) 10.68		DOR	0.17	0.07	0.07	95.33	0.61	0.04	0.04	95.17	0.35	0.03	0.03	94.17
		DAC	0.17	0.07	0.07	95.50	0.60	0.04	0.04	94.50	0.34	0.03	0.03	94.17
(iv) 15.27		DOR	-1112.65	0.46	0.44	31.17	-1059.41	0.26	0.25	1.50	-1048.71	0.20	0.20	0.17
		DAC	-803.27	0.46	0.44	54.33	-770.35	0.27	0.26	17.17	-760.74	0.20	0.20	4.17

Note: Scenarios and Methods are described in Table 1. Bias (magnified 1000 times for presentation) and SD are the Monte Carlo bias and standard deviation across the 600 simulations of the points estimates. ESE and CP(%) are the averages of estimated asymptotic standard errors and coverage proportions of 95% confidence intervals based on the plug-in method, respectively.

Figure 3 shows that augmented calibration weighting estimates calculated by aggregated data in the distributed setting are identical those from pooled data analysis, confirming the lossless property of the DAC method. Table 2 summarizes the performance of DAC estimators and DOR estimators in estimating ATEs for treatment 4 versus treatment 1 on the overall population $\{1, 2, 3, 4\}$. Results for target populations $\{3\}$, $\{1, 3\}$, $\{1, 3, 4\}$ are detailed in Tables S1-S3 in the SM. For DAC estimators, the averages of estimated asymptotic standard errors (ESEs) are close to the Monte Carlo standard deviations (SDs), validating the asymptotic variance estimation and the doubly robust inference property. DAC estimators with at least one correctly specified working model under Scenarios (i)-(iii) exhibit very small bias and comparable Monte Carlo standard deviations relative to DOR estimators, even for small sample sizes. In contrast, DOR estimators in Scenarios (ii) and (iv) display large bias when their working models are misspecified. DAC reduces DOR’s bias by 25.3%–31.3% when neither working model is correctly specified. Coverage proportions based on DAC estimators approximate the nominal level of 0.95 when at least one working model is correctly specified in Scenarios (i)-(iii), while OR estimators achieve nominal coverage only under correctly specified outcome models.

7 Real Data Analysis

We conducted our empirical analysis using data from the Magnetic Resonance Imaging (MRI) sub-study of the Memory in Diabetes (MIND) substudy within the Action to Control Cardiovascular Risk in Diabetes (ACCORD) trial. ACCORD was a randomized trial of 10,251 middle-aged and older participants with type 2 diabetes mellitus at high cardiovascular risk due to prevalent cardiovascular disease or additional risk factors. Participants in the main ACCORD trial were randomized to either a systolic blood pressure goal of less than 120 mm Hg (intensive BP therapy) versus less than 140 mm Hg (standard BP therapy), or to a fibrate versus placebo for patients with

low-density lipoprotein cholesterol levels less than 100 mg/dL. 632 participants from ACCORD trial who agreed to MRI procedure were enrolled in MRI substudy to examine whether intensive hypertension therapy and combination therapy with a statin plus fibrate could reduce the risk of total brain volume (TBV) decline in patients with type 2 diabetes mellitus. A flow chart in Figure S5 of the SM shows the cohort participation; see Williamson et al. (2014) for details on the study design. Of these participants, 614 completed the baseline TBV measurement within 45 days after randomization, and 503 completed a 40-month follow-up TBV measurement. This final subset comprised 314 participants from the blood pressure (BP) trial (161 standard BP and 153 intensive BP) and 189 from the lipid trial (100 placebo and 89 fibrate). Due to participant attrition during follow-up, there were some imbalances in baseline characteristics of these 503 participants, as shown in Figure S6 of the SM. Williamson et al. (2014) analyzed these patient-level data using analysis of covariance with the adjustment on baseline characteristics, equivalent to outcome regression estimation with linear models, reporting a 4.4 cm^3 [95% CI, 1.1 to 7.8 cm^3 ; P-value = 0.01] greater TBV decline in the intensive BP group compared to the standard BP group at 40 months, while fibrate therapy had no effect on TBV changes compared to placebo.

Here we investigate effects of the four interventions (standard BP, intensive BP, lipid placebo and lipid fibrate) using DAC method (Figure 1), treating each treatment group as a single-arm trial within a distributed network. This causal framework enables fair treatment comparisons across various target populations, addressing baseline imbalances and enhancing estimation robustness. Specifically, we estimate the expected counterfactual 40-month TBV changes for the four interventions and compare the effects of intensive BP versus standard BP, fibrate versus placebo and intensive BP versus fibrate across the BP trial population, lipid trial population, and overall population combining both trial populations. DAC estimates are compared with DOR estimates.

In our analysis, outcome models for 40-month TBV changes use cubic B-splines for contin-

uous variables and include all covariates selected via forward selection in each therapy group. These covariates include gender, history of cardiovascular disease, myocardial infarction, stroke, smoking status, education level, self-report history of depression, glycemia intervention, clinical center network, and spline functions of baseline TBV, intracranial volume, age, height and body mass index. The same covariates are used for calibration weighting. The outcome models achieved R-squared values of 67.4%, 66.8%, 71.4% and 76.5% for the standard BP, intensive BP, lipid placebo, and lipid fibrate groups, respectively.

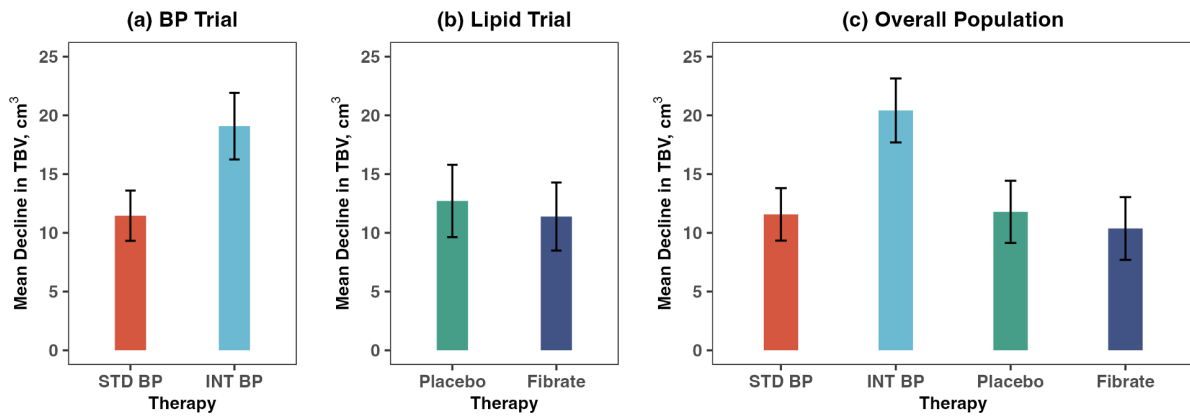


Figure 4: Means of counterfactual 40-month declines in total brain volume (TBV) under four treatments: standard (STD) blood pressure, intensive (INT) blood pressure, lipid placebo and lipid fibrate in (a) the blood pressure (BP) trial population, (b) the lipid trial population and (c) the overall population combining BP and lipid trials. Whiskers mark 95% confidence intervals.

Figure 4 displays DAC estimates of counterfactual means of 40-month TBV declines under the four interventions in the BP trial, lipid trial, and overall populations, revealing significant TBV declines for all interventions. Estimates for the overall population differ slightly from trial-specific estimates due to heterogeneity between the BP and lipid trials, consistent with results in Williamson et al. (2014). Table 3 summarizes the ATEs of 40-month TBV changes from baseline between different treatments. For intensive versus standard BP therapy, the DAC estimate in the overall population shows an 8.8 cm³ [95% CI, 5.5 to 12.2 cm³; P-value = 2×10^{-7}] greater TBV decline under intensive therapy. No significant difference was observed between fibrate therapy

and placebo. Both estimation methods yield consistent conclusions, aligning with findings of Williamson et al. (2014). And DAC estimators exhibited slightly smaller variances than DOR estimators. These highlight the robustness and improved efficiency of DAC.

Table 3: The distributed outcome regression (DOR, Figure S3 in the SM) and distributed augmented calibration weighting (DAC, Figure 1) estimators with their 95% confidence intervals for average treatment effects (ATEs) on the 40-month changes of total brain volume from baseline between standard (STD) blood pressure, intensive (INT) blood pressure, lipid placebo and lipid fibrate treatments in the blood pressure (BP) trial, the lipid trial, and the overall populations. P-values from two-sided hypothesis tests are listed in the last column.

Target population	Comparators	Method	ATE (95% CI), cm ³	P-value
BP trial population	INT BP v.s. STD BP	DOR	-7.7 (-11.3, -4.1)	< 0.001
		DAC	-7.6 (-10.9, -4.3)	< 0.001
Lipid trial population	Fibrate v.s. Placebo	DOR	1.6 (-2.3, 5.5)	0.41
		DAC	1.3 (-2.5, 5.2)	0.50
Overall population	INT BP v.s. STD BP	DOR	-9.2 (-13.1, -5.2)	< 0.001
		DAC	-8.8 (-12.2, -5.5)	< 0.001
	Fibrate v.s. Placebo	DOR	2.1 (-1.6, 5.8)	0.27
		DAC	1.4 (-2.2, 5)	0.44
	INT BP v.s. Fibrate	DOR	-11.1 (-15.5, -6.8)	< 0.001
		DAC	-10.0 (-13.9, -6.2)	< 0.001

8 Conclusion

In this paper, we propose a novel collaborative indirect treatment comparison method to simultaneously estimate ATEs between any pair of treatments across various target populations in a distributed network of multiple single-arm trials without overlapping treatments. It introduces the DAC estimation, which is doubly robust consistent, accommodates data heterogeneity, and requires only aggregated data from each site. For parametric working models, the asymptotic variance of the DAC estimator depends on how nuisance parameters are estimated. To address this, we develop a bias-reduced DAC estimator in which nuisance parameters are estimated by adapting to the EIF, to achieve doubly robust inference and reduced first-order asymptotic bias. We provide the four-step, lossless DAC estimation procedure which requires two communication rounds between the coordinating center and participating sites.

We extend the proposed estimation method to handle multi-arm trials with overlapping treatments across different trials and consider the relaxation of the exchangeability assumption in Section S6 of the SM. A potential future direction is adapting the method for high-dimensional data, where numerous covariates are used for confounding adjustment via outcome regression and propensity score models. Approaches for doubly robust inference in high-dimensional settings have been developed for centralized frameworks (Tan, 2020; Ning et al., 2020) and could be adapted for distributed networks. Another extension involves creating a debiased calibration estimator using generalized entropy (Kwon et al., 2024) to improve estimation accuracy, especially in scenarios with complex relationships between the calibrated variables and outcomes.

References

- Chan, K., Yam, S., and Zheng, Z. (2015). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 78(3):673–700.
- Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. volume 6 of *Handbook of Econometrics*, page 5549–5632. Elsevier.
- Chen, X. and White, H. (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2):682–691.
- Cheng, D., Ayyagari, R., and Signorovitch, J. (2020). The statistical performance of matching-adjusted indirect comparisons: estimating treatment effects with aggregate external control data. *The Annals of Applied Statistics*, 14(4):1806–1833.
- Chernozhukov, V., Chetverikov, D., Demirer, M., et al. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21:C1–C68.

- Dagenais, S., Russo, L., Madsen, A., et al. (2022). Use of real-world evidence to drive drug development strategy and inform clinical trial design. *Clinical Pharmacology & Therapeutics*, 111(1):77–89.
- Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.
- European Commission (2024a). Guidance on outcomes for joint clinical assessments.
- European Commission (2024b). Guidance on the validity of clinical studies for joint clinical assessments.
- European Union (2021). Regulation (EU) 2021/2282 of the European Parliament and of the Council of 15 December 2021 on Health Technology Assessment.
- Food and Drug Administration (2001). Guidance for industry: E10 choice of control group and related issues in clinical trials. <https://www.fda.gov/media/71349/>. Accessed September 25, 2024.
- Funk, M. J., Westreich, D., Wiesen, C., et al. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7):761–767.
- Gray, C. M., Grimson, F., Layton, D., et al. (2020). A framework for methodological choice and evidence assessment for studies using external comparators from real-world data. *Drug Safety*, 43:623–633.
- Gu, J. and Chen, S. X. (2023). Distributed statistical inference under heterogeneity. *Journal of Machine Learning Research*, 24(387):1–57.
- Guo, Z., Li, X., Han, L., et al. (2025). Robust inference for federated meta-learning. *Journal of the American Statistical Association*, 0(0):1–16.

- Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 20(1):25–46.
- Han, L., Hou, J., Cho, K., et al. (2025). Federated Adaptive Causal Estimation (FACE) of target treatment effects. *Journal of the American Statistical Association*.
- Hatswell, A. J., Baio, G., Berlin, J. A., et al. (2016). Regulatory approval of pharmaceuticals without a randomised controlled study: Analysis of EMA and FDA approvals 1999–2014. *BMJ Open*, 6(6):e011666.
- Hu, M., Shi, X., and Song, P. X.-K. (2024). Collaborative inference for treatment effect with distributed data-sharing management in multicenter studies. *Statistics in Medicine*, 43(11):2263–2279.
- Jahanshahi, M., Gregg, K., Davis, G., et al. (2021). The use of external controls in fda regulatory decision making. *Therapeutic Innovation & Regulatory Science*, 55(5):1019–1035.
- Kim, J. K. and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, 24(1):375–394.
- Kwon, Y., Kim, J. K., and Qiu, Y. (2024). Debiased calibration estimation using generalized entropy in survey sampling. *arXiv*.
- Lee, Y.-Y. (2018). Efficient propensity score regression estimators of multivalued treatment effects for the treated. *Journal of Econometrics*, 204(2):207–222.
- Li, X. and Song, Y. (2020). Target population statistical inference with data integration across multiple sources — an approach to mitigate information shortage in rare disease clinical trials. *Statistics in Biopharmaceutical Research*, 12(3):322–333.

- Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960.
- Miksad, R. A., Samant, M. K., Sarkar, S., et al. (2019). Small but mighty: the use of real-world evidence to inform precision medicine. *Clinical Pharmacology & Therapeutics*, 106(1):87–90.
- Ning, Y., Sida, P., and Imai, K. (2020). Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika*, 107(3):533–554.
- Patel, D., Grimson, F., Mihaylova, E., et al. (2021). Use of external comparators for health technology assessment submissions based on single-arm trials. *Value in Health*, 24(8):1118–1125.
- Phillippo, D., Ades, T., Dias, S., et al. (2016). NICE DSU technical support document 18: Methods for population-adjusted indirect comparisons in submissions to NICE. <https://research-information.bris.ac.uk/ws/portalfiles/portal/94868463>. Accessed September 25, 2024.
- Phillippo, D. M., Ades, A. E., Dias, S., et al. (2018). Methods for population-adjusted indirect comparisons in health technology appraisal. *Medical Decision Making*, 38(2):200–211.
- Qin, J. and Zhang, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(1):101–122.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.
- Schmidli, H., Gsteiger, S., Roychoudhury, S., et al. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032.

- Seymour, L., Ivy, S. P., Sargent, D., et al. (2010). The design of phase II clinical trials testing cancer therapeutics: Consensus recommendations from the clinical trial design task force of the national cancer institute investigational drug steering committee. *Clinical Cancer Research*, 16(6):1764–1769.
- Shi, X., Pan, Z., and Miao, W. (2023). Data integration in causal inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(1):e1581.
- Signorovitch, J., Wu, E., Yu, A., et al. (2010). Comparative effectiveness without head-to-head trials a method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. *Pharmacoeconomics*, 28(10):935–945.
- Stuart, E., Cole, S., Bradshaw, C., et al. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society Series A*, 174(2):369–386.
- Tan, Z. (2020). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *Annals of Statistics*, 48(2):811–837.
- Thorlund, K., Dron, L., Park, J. J., et al. (2020). Synthetic and external controls in clinical trials—a primer for researchers. *Clinical epidemiology*, 12:457–467.
- Van der Vaart, A. (2002). *Semiparametric statistics*, pages 331–457. Number 1781 in Lecture Notes in Math. Springer.
- Vermeulen, K. and Vansteelandt, S. (2015). Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110(511):1024–1036.
- Vickers, A. J., Ballen, V., and Scher, H. I. (2007). Setting the bar in phase II trials: The use of

- historical data for determining “go/no go” decision for definitive phase III testing. *Clinical Cancer Research*, 13(3):972–976.
- Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.
- Wang, H., Wu, F., and Chen, Y.-F. (2024). A comparison of methods for estimating the average treatment effect on the treated for externally controlled trials. *arXiv*.
- Weber, K., Hemmings, R., and Koch, A. (2018). How to use prior knowledge and still give new data a chance? *Pharmaceutical Statistics*, 17(4):329–341.
- Williams, D. M. and Evans, M. (2023). The evolution of real-world evidence in healthcare decision making. *Expert Opinion on Drug Safety*, 22(6):443–445.
- Williamson, J. D., Launer, L. J., Bryan, R. N., et al. (2014). Cognitive function and brain structure in persons with type 2 diabetes mellitus after intensive lowering of blood pressure and lipid levels: a randomized clinical trial. *JAMA internal medicine*, 174(3):324–333.
- Woolacott, N., Corbett, M., Jones-Diette, J., et al. (2017). Methodological challenges for the evaluation of clinical effectiveness in the context of accelerated regulatory approval: An overview. *Journal of Clinical Epidemiology*, 90:108–118.
- Xiong, R., Koenecke, A., Powell, M., et al. (2023). Federated causal inference in heterogeneous observational data. *Statistics in Medicine*, 42(24):4418–4439.
- Zhao, Q. and Percival, D. (2016). Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1):20160010.
- Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922.