Received: 7 September 2023 | Revised: 1 February 2024 | Accepted: 14 March 2024

DOI: 10.1002/sim.10068

RESEARCH ARTICLE

Statistics in Medicine WILEY

# Collaborative inference for treatment effect with distributed data-sharing management in multicenter studies

**Mengtong Hu** | **Xu Shi** | **Peter X.-K. Song**

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan,

**Correspondence**
Peter X.-K. Song, Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA.
Email: pxsong@umich.edu

**Funding information**
National Institutes of Health, Grant/Award Numbers: R01ES033656, R01GM139926, U24ES028502; National Science Foundation, Grant/Award Number: DMS2113564

Data sharing barriers present paramount challenges arising from multicenter clinical studies where multiple data sources are stored and managed in a distributed fashion at different local study sites. Merging such data sources into a common data storage for a centralized statistical analysis requires a data use agreement, which is often time-consuming. Data merging may become more burdensome when propensity score modeling is involved in the analysis because combining many confounding variables, and systematic incorporation of this additional modeling in a meta-analysis has not been thoroughly investigated in the literature. Motivated from a multicenter clinical trial of basal insulin treatment for reducing the risk of post-transplantation diabetes mellitus, we propose a new inference framework that avoids the merging of subject-level raw data from multiple sites at a centralized facility but needs only the sharing of summary statistics. Unlike the architecture of federated learning, the proposed collaborative inference does not need a center site to combine local results and thus enjoys maximal protection of data privacy and minimal sensitivity to unbalanced data distributions across data sources. We show theoretically and numerically that the new distributed inference approach has little loss of statistical power compared to the centralized method that requires merging the entire data. We present large-sample properties and algorithms for the proposed method. We illustrate its performance by simulation experiments and the motivating example on the differential average treatment effect of basal insulin to lower risk of diabetes among kidney-transplant patients compared to the standard-of-care.

**KEYWORDS**
data privacy, distributed inference, federated learning, meta-analysis, renewable estimation

## 1 | INTRODUCTION

Multicenter clinical studies are undertaken widely in practice, as this study design empowers practitioners to enhance statistical power, increase population diversity, and improve generalizability, thereby enhancing the replicability of clinical findings. Distributed data management is typically adopted in multicenter studies for data collection and storage, as

data-sharing barriers can hinder the timely creation of a centralized database containing all variables needed for statistical analyses. These operational challenges call for new methods that enable statistical inference to evaluate covariate effects (eg, treatment effect) and adjust for different sources of bias[1] while considering data privacy protection. We propose a new inference analytic, termed *collaborative inference*, which utilizes a different data communication architecture from existing federated learning methods. As evidenced in the article, this new inference method offers clear advantages over conventional meta-analysis, an exemplary approach in the federated learning paradigm.

Our development of collaborative inference is motivated by one of our collaborative projects, the Insulin Therapy for the Prevention of New-Onset Diabetes after Transplantation (ITP-NODAT) trial, which involves four kidney transplant centers in three European countries to evaluate the efficacy of the basal insulin intervention in preventing post-transplantation diabetes mellitus (PTDM). In such a cross-border trial we encountered significant data sharing barriers owing largely to disparate country-specific data security and privacy requirements, as well as cumbersome cross-border institutional IRB approval procedures,[2-4] among other logistic challenges. Holdups in data procurement have caused significant delays in the publication of clinical findings and consequently impeded the delivery of new therapeutics to patients. Such delays are undesirable and even unethical in some cases.

Among several solutions available in the literature, the federated learning paradigm, including the well-known meta-analysis, is of great popularity. A meta-estimation of treatment effect may be calculated by an inverse-variance weighted average of site-specific treatment effects obtained from individual data sources respectively, termed *the classical meta-analysis* in this article (for example, Reference 5). In general, federated learning analytics communicate only site-specific summary statistics with a semi-trusted central computing node, rather than the full subject-level data. In particular, the meta-estimation aims to pool site-specific estimates of treatment effects to produce an overall estimation as a final result.

Unfortunately, the data management protocol of the IPT-NODAT trial has limited the application of the classical meta-analysis in the data analysis. Two significant limitations include:

(i) Statistical inference (eg, confidence intervals) in the context of meta-analysis is highly sensitive to the quality and reliability of local results obtained from individual study sites. These local analyses often operate independently or with minimal across-site cooperation during intermediate steps. For example, when using a propensity score model [6] to address potential sampling bias in covariate distributions across study sites, the application of inverse probability weighting (IPW) is typically carried out at the local level, resulting in site-specific IPW computations. Even in the context of randomized trials, the balance of covariate distributions is not always warranted due to various challenges during the study implementation.[7] For example, in many clinical studies, additional study sites may be added or randomization schemes may be revised after interim analyses for ethical considerations, recruitment delays, patient dropouts, and other types of data attrition. These factors can exacerbate imbalances in covariate distributions, resulting in estimation bias and a loss of estimation efficiency. Furthermore, a meta-analysis does not enforce the control of common model specifications across participating sites, potentially compromising the interpretability of the pooled estimate, especially in a situation where a consistent set of confounding variables is lacking for correcting sampling bias.

(ii) The meta-analysis depends heavily on varying recruitment capacity across study sites in which small sample sizes or low sample variability at some study sites can impair the statistical power of the analysis, as summary statistics obtained from those sites may be unavailable or unreliable. Refer to Section 2 for supporting evidence in the IPT-NODAT study data.

Several federated learning methods have been developed to improve the classical meta-analysis approach. Jordan et al[8] developed a surrogate likelihood framework that communicates local estimates of gradients from a common initial estimate from one study site to update estimation and inference. This estimation method is sensitive to the quality of the initial values, such as the persistence of the approximation error in local Hessian from a single site to the global level. This framework is later extended by Duan et al[9,10] for distributed analyses of clinical datasets. A distributed empirical likelihood method designed for unbalanced datasets is recently proposed by Zhou et al.[11] Also see recent work on distributed average treatment effect estimation by Xiong et al[12] and Vo et al.[13] In addition, transfer learning methods have been adapted to leverage external source data to improve the average treatment effect estimation in a target site, allowing potential heterogeneity in covariate distributions; see for example Zeng et al[14] and Han et al.[15-17] In contrast to methods that prioritize statistical inference, there are also multiple lines of federated learning methods in the computer science literature that primarily focus on parameter estimation.[18,19] Most of these existing meta-analytics require a semi-trusted

center node under a divide-and-combine strategy in a parallelized operation paradigm that demands reliable local estimates and inferential quantities. Unfortunately, due to various reasons pointed out above, the demand for high-quality inputs from local sites is not easily satisfied in practice. For instance, in the IPT-NODAT trial (see Section 2), two of four hospitals fail to produce sensible local results.

This article focuses on the development of a more flexible and reliable meta-analysis methodology by overcoming the above-marked impediments to evaluate *differential average treatment effect (DATE)* through effective data-sharing management schemes, in which propensity score models are used to deal with unbalanced covariate distributions. DATE refers to a difference of marginal comparison of treatment effects between two treatment arms, which is also known as average treatment effect (ATE) in the literature of causal inference. Note that our goal is to estimate DATE, a more general estimand than ATE; in effect, DATE includes odds ratio and relative risk when applied to binary outcomes that are often collected and studied as the primary endpoint in clinical trials. Among multiple possible strategies concerning the calculation and utilization of propensity scores in the case of distributed data, we design four new procedures for the effectual communication of summary statistics across study sites, distinct from the federated learning architecture. Our proposed method is operated under the assumption that data collected across sites are compatible or exchangeable in that the propensity score model of treatment allocation is homogeneous across study sites. Consequently, we can seamlessly integrate the propensity score calculation and treatment effect estimation using a systematic cross-site collaboration. We term this new approach as *Collaborative Operation of Linked Analysis (*COLA*)*. In general, COLA provides an estimate of DATE; when applied to the counterfactual framework, under the fundamental identifiability assumptions, COLA can yield a causal inference for the average treatment effect.

The reason that the COLA method appears more flexible than the existing meta-analysis is that it adopts a sequential updating machinery[20] for cross-site communication, different from the currently popular parallelized operation. Our proposed COLA methodology can handle nonlinear estimands such as odds ratio in logistic regression using estimating equations approaches. Such statistical analytics avoid reliance on closed-form expressions of treatment effect estimates, which would otherwise be decomposed into site-specific statistics in a similar spirit to the federated learning machinery. The improved flexibility is achieved through different options of passing information in the COLA machinery to reach a desirable trade-off between information communication cost and numerical estimation efficiency. Figure 1 shows our proposed relays for summary data communications in COLA , which leads to a fully efficient IPW estimation of DATE in the sense that its convergence rate is at the order of the cumulative sample size. We show both theoretically and numerically that the COLA has no loss of statistical power in comparison to the oracle estimation obtained by the centralized analysis that merges data from all sites. The practical implication of preserved statistical power is that no loss of sample size means no need for additional patient recruitment to meet the pre-designed power for a clinical study. In contrast, this power protection won't hold in meta-type data integration approaches. In addition, To check the compatibility of data collected across sites amid data merging, following Reference 21, we utilize the generalized method of moments (GMM)[22] to test the homogeneity of the estimating equations. This diagnostic procedure allows us to examine if covariates sampled at different sites follow the same underlying population distribution and give rise to compatible propensity scores.

The organization of this article is as follows. We first introduce a motivating example of a multicenter clinical trial in Section 2. Section 3 begins with formulation and model assumptions for estimating DATE and then introduces the proposed COLA framework. Section 4 presents different options for passing information in the COLA machinery. Section 5 establishes the theoretical guarantees for our proposed methods. We illustrate our proposed methods with simulation studies in Section 6 and an application to the ITP-NODAT data example in Section 7. We make some concluding remarks in Section 8. The supporting information is given at the end of the references.

## 2 | MOTIVATING EXAMPLE: IPT-NODAT MULTICENTER TRIAL

We illustrate the proposed COLA methodology in a simple yet practically important setting of logistic regression with a binary outcome. This setting is motivated by the multicenter clinical study, IPT-NODAT. This cross-border clinical study is a randomized clinical trial conducted at four kidney transplant centers in Barcelona, Spain; Berlin, Germany; Graz, Austria; and Vienna, Austria. The trial investigates a common complication developed after kidney transplantation, post-transplant diabetes mellitus (PTDM), which is a unique form of diabetes and is associated with increased mortality and cardiovascular events. The goal of the study is to estimate DATE of a diabetics preventive treatment (ie, basal insulin intervention) vs a standard-of-care on preventing PTDM. A patient is diagnosed as a PTMD case if they receive antidiabetic therapy, have 2-h plasma glucose $\geq$ 200 mg/dL, or have Hemoglobin A1C (HbA1c) $\geq$ 6.5%. Two hundred and thirty-six

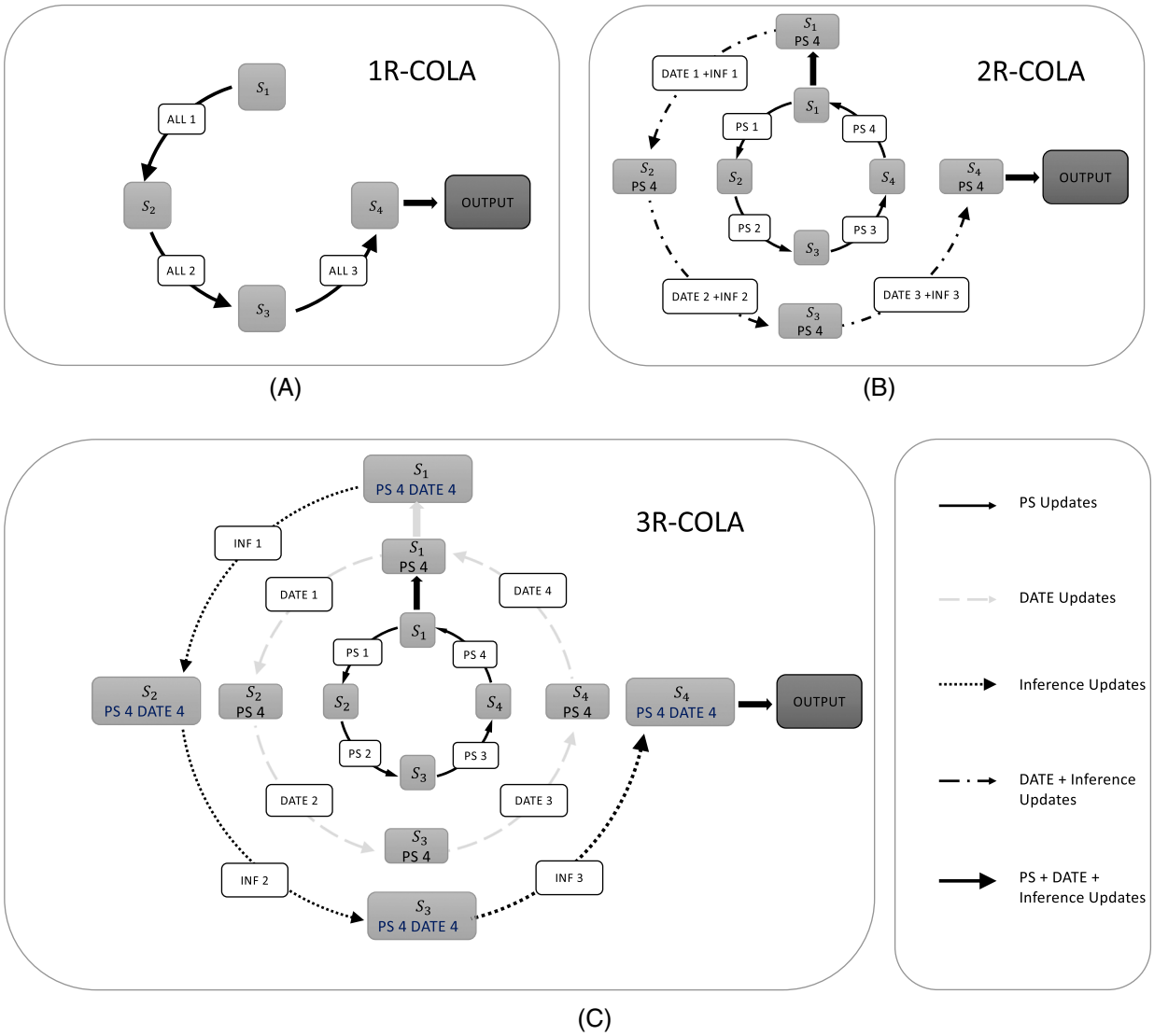


**FIGURE 1** A diagram showing 1R-COLA, 2R-COLA, and 3R-COLA procedures of the collaborative inference framework. Different types of arrows indicate the major updates involved in each round as shown in the legend. The subscripts 1 to 4 indicate the site numbers and the arrows start from Site 1 and end at Site 4. DATE is short for differential average treatment effect. Panel (A) shows 1R-COLA evolves one round operation that updates PS, DATE, and inferential quantities simultaneously. Panel (B) shows 2R-COLA involving two rounds of updates where the first round produces PS estimates, and the second round produces a DATE and its variance. Panel (C) shows 3R-COLA involving three-rounds of updates.

kidney-transplantation patients were recruited and randomized within each hospital to receive basal insulin injections or standard-of-care right after kidney transplantation. The endpoint of this trial was a binary outcome of the occurrence of diabetes (yes or no) at 12 months after the treatment.

The IPW method is preferable for analyzing this data due to its effectiveness in addressing the apparent imbalances present in certain covariate distributions of this study.[23] Meanwhile, some logistic challenges in data sharing have led to a significant delay in the creation of a centralized database due to the cross-border nature of this trial. Consequently, the dissemination of clinical findings from this trial has been greatly postponed, which otherwise could have delivered a great benefit to transplant patients, given that this trial has indeed shown a significant preventive effect in reducing the risk of PTDM.

The meta-analysis method is an obvious choice for expediting the statistical analysis prior to the availability of the centralized data. While meta-analysis bypasses the need for data sharing, it fails to take into account information from all sites.

This is because zero disease cases are observed from the treatment group in Barcelona and likewise, from the control group in Graz. Consequently, local DATE estimates are unavailable at these two hospitals. As a result, the meta-analysis would use only two (instead of four) local odds ratio estimates from Berlin (OR 1.1[0.29, 4.10]) and Vienna (OR 0.12[0.01, 1.12]). The resulting meta-estimate of the odds ratio is 0.62 with a 95% confidence interval of [0.2, 1.93], leading to a conclusion of no evidence of treatment effect. In contrast, when the centralized data is used in the analysis, the DATE is found to be statistically significant; see Section 7. Part of the reason for such discrepancy lies in the substantial data attrition in the meta-analysis, which is unfortunately underpowered compared to the centralized analysis using the full data combining four sites. Arguably, a technical gap of great impact in practice arises in the above analysis. To expedite the delivery of clinical findings without relying on centralized data management, it is of great interest to develop statistical methods that can provide the needed flexibility to handle rare outcomes and balance covariate distributions to maintain statistical power and produce an efficient and reliable estimation of DATE for basal insulin therapy vs the standard-of-care.

## 3 | METHODOLOGY

### 3.1 | Setup

Let us begin with some notations. Consider a random sample of $N$ individuals independently sampled from $K$ clinical sites, indexed by $j = 1, \ldots, K$, each site having a sample size of $n_j$. For each individual $i$, we observe an outcome $Y_i$ of interest, a binary treatment indicator $A_i \in \{0, 1\}$, and a $p$-element vector of baseline covariates $X_i$. Let $I_j = \{n_{j-1} + 1, \ldots, n_{j-1} + n_j\}$ be the index set for subjects from the $j$th site and we denote the $j$th site-specific data as $S_j = \{(Y_i, A_i, X_i) : i \in I_j\}$. We assume $\{(Y_i, A_i, X_i) : i \in \bigcup_{j=1}^{K} I_j\}$ are *independent and identically distributed* observations drawn from an underlying population of interest. Let $Y(a), a \in \{0, 1\}$, be a binary *intent-to-treat* outcome from treatment $a$. It is worth emphasizing that "intent-to-treat" bears a more general meaning than what is used in the missing data literature. Denote $\mu_1 = E\{Y(1)\}$ and $\mu_0 = E\{Y(0)\}$, respectively, the population-average outcome values by an active treatment ($a = 1$) or by a control treatment ($a = 0$). These outcome values are also referred to as potential outcomes in the Neyman-Rubin causal paradigm.[24,25] We are interested in the differential average treatment effect (DATE) expressed in the three estimands: the mean difference, $\Delta_D = \mu_1 - \mu_0$, the log risk ratio (logRR), $\Delta_{\text{RR}} = \log(\mu_1/\mu_0)$, and the log odds ratio (logOR), $\Delta_{\text{OR}} = \log[\{\mu_1/(1-\mu_1)\}/\{\mu_0/(1-\mu_0)\}]$.

*Remark* 1. DATE may be profiled by the baseline covariates $X$ by conditional means $\mu_1(X) = E\{Y(1)|X\}$ and $\mu_0(X) = E\{Y(0)|X\}$, both depending on patient's characteristics $X$. In this case, DATE may be referred to as adjusted DATE, conditional DATE or stratified DATE. In this article, we focus on the population-average DATE derived by averaging the conditional means over the sample space of $X$. By doing so, we obtain the population-average intent-to-treat effect where the mean outcome value of treatment $a$ is $\mu_a = E\{Y(a)\} = E_x[E\{Y(a)|X\}]$, where $E_x$ is the operation of expectation under the distribution of $X$. Following the inverse probability weighting (IPW) principle, we can approximate this $E_x$-operation using the law of large numbers by populating the observed patients who received the actual treatment $a$ via their propensity scores. In this way, we avoid estimating the joint distribution of $X$, which is in general very difficult.

To apply IPW to estimate DATE, we need an assumption that permits to "clone" observed patients to create a pseudo (or intent-to-treat) population via the propensity score method. Let $Y = I(A = 1)Y(1) + I(A = 0)Y(0)$ be the observed outcome.

**Assumption 1.** For $a \in \{0, 1\}$, we assume

(a) Consistency: $Y = Y(a)$ almost surely when $A = a$, namely the observed outcome is the same as the intent-to-treat outcome in a clinical study.
(b) Ignorability: $Y(a) \perp\!\!\!\perp A|X$, namely, treatment allocation is independent of intent-to-treat outcome.
(c) Positivity: $0 < pr(A = a|X) < 1$ for all $a$ almost surely, namely either treatment arm has a nonzero chance of allocation.

Under Assumption 1, it is easy to show that the average intent-to-treat outcome of arm $a$ is identified as $\mu_a = E[Y(a)] = E[I(A = a)Y/P(A = a|X)]$, $a = 0$ or 1. Thus, this $\mu_a$ can be estimated from data collected from a clinical study

through different weighting methods including the inverse probability of treatment weighting (IPW), G-computation, and augmented inverse propensity weighting (AIPW).[26]

## 3.2 | Inverse propensity weighting estimator

In this article, we illustrate our COLA method using IPW estimation in the proposed collaborative inference, although COLA also applies to other methods such as G-computation and AIPW. The propensity score $e(X)$ is the probability of being assigned to the treatment group conditional on the covariates, that is, $e(X) = pr(A = 1|X)$. A working model for $e(X)$ used in IPW method needs to be correctly specified for consistent estimation of the treatment effect. As such we introduce the following assumption.

**Assumption 2** (Propensity score model). A propensity score model $e(X;\gamma)$ with a finite-dimensional parameter $\gamma$ is correctly specified for $e(X)$ and parameter $\gamma$ is the same for all individuals.

Typically, the propensity score is modeled and estimated by the logistic regression, namely $e(X;\gamma) = \text{logit}(X^\top\gamma)$. We begin with a brief review of the classical estimation and inference method based on IPW in the setting where data from all sites are available and analyzed in a centralized fashion. We term this situation of centralized operation as the *oracle* setting in this article. The IPW method is a two-stage procedure. First, we estimate $\gamma$ as the solution to estimating equation $\sum_{i=1}^N \Psi^{\text{ps}}(A_i, X_i;\gamma) = 0$, denoted by $\hat{\gamma}$, where the kernel function is $\Psi^{\text{ps}}(A, X;\gamma) = X\{A - e(X;\gamma)\}$. Then, we fit a marginal structural model by solving $\sum_{i=1}^N \Psi^{\Delta}(A_i, X_i, Y_i;\hat{\gamma},\beta) = 0$, with $\Psi^{\Delta}(A, X, Y;\gamma,\beta) = (1, A)^\top\omega(A, X;\gamma)\{Y - g^{-1}(\beta_0 + \beta_A A)\}$, where $\omega(A, X;\gamma) = A/e(X;\gamma) + (1 - A)/\{1 - e(X;\gamma)\}$ is the inverse probability of treatment weight, $g(\cdot)$ is a user-specified link function, and the treatment coefficient $\beta_A$ is the treatment effect at a scale corresponding to the link function. For example, when the outcome is binary, one typically uses the logit link function such that $g^{-1}(x) = 1/(1 + e^{-x})$ and the slope parameter, that is, the coefficient for treatment $\beta_A = \Delta_{\text{OR}}$. One can also jointly estimate all parameters related to propensity score and treatment effect, denoted as $\theta = (\gamma, \beta)^\top$, by stacking $\Psi^{\text{ps}}$ and $\Psi^{\Delta}$ into a joint estimating function

$$\Psi(\theta) = \Psi(A, X, Y;\theta) = \begin{pmatrix} \Psi^{\text{ps}}(A, X;\gamma) \\ \Psi^{\Delta}(A, X, Y;\gamma,\beta) \end{pmatrix}. \tag{1}$$

Under Assumptions 1 and 2, the estimating function has mean zero when evaluated at the true value $\theta_0 = (\gamma_0, \beta_0)^\top$, that is, $E\{\Psi(\theta_0)\} = 0$ which will be needed for deriving the asymptotic results in Section 5. Hereafter, we assume both Assumptions 1 and 2 hold.

Let $\hat{\theta}^{\text{ora}} = (\hat{\gamma}^{\text{ora}}, \hat{\beta}^{\text{ora}})^\top$ denote the oracle estimator obtained from the centralized analysis which solves $\sum_{i=1}^N \Psi_i(\theta) = 0$, where $\Psi_i(\theta) = \Psi(A_i, X_i, Y_i;\theta)$. Since $\Psi(\theta)$ is an unbiased estimating function, under some mild regularity conditions, we have the asymptotic normality for $\hat{\theta}^{\text{ora}}$, namely $\sqrt{N}(\hat{\theta}^{\text{ora}} - \theta_0) \xrightarrow{d} N(0, J(\theta_0))$, where $J(\theta_0)$ is the inverse of the Godambe information matrix, also named as the sandwich covariance matrix.[27-31] Specifically, $J(\theta_0) = H(\theta_0)^{-1}V(\theta_0)\{H(\theta_0)^{-1}\}^\top$, where $V(\theta_0) = E_{\theta_0}\{\Psi(\theta_0)\Psi^\top(\theta_0)\}$ is the variability matrix and $H(\theta_0) = -E_{\theta_0}\{\partial\Psi(\theta_0)/\partial\theta^\top\}$ is the sensitivity matrix. We can estimate the variability and sensitivity matrices using their sample counterparts given by $V(\hat{\theta}^{\text{ora}}) = \sum_{i=1}^N \Psi_i(\theta)\Psi_i^\top(\theta)|_{\theta=\hat{\theta}^{\text{ora}}}$, and $H(\hat{\theta}^{\text{ora}}) = \sum_{i=1}^N -\partial\Psi_i(\theta)/\partial\theta^\top|_{\theta=\hat{\theta}^{\text{ora}}}$. The resulting oracle estimator $\hat{\theta}^{\text{ora}}$ and its variance obtained from the pooled data of size $N$ will serve as the gold standard to compare with our proposed distributed methods.

## 3.3 | Incremental treatment effect estimator

We consider a situation of practical importance where pooling data from multiple sites is prohibited at the current time or in the near future. To address this data-sharing challenge, we propose an estimation method, termed Collaborative Operation of Linked Analysis (COLA), which does not require sharing individual-level data but only certain summary statistics across institutes. Note that, different from most of the existing solutions, COLA is not derived in a parallel computing paradigm, but rather in a collaborative fashion analogous to a relay race. Specifically, given an order of study sites, an incremental estimator, denoted by $\hat{\theta}_k$, is sequentially updated over the first $k$ sites. We further define

$$H_j(\hat{\theta}_j) = \sum_{i \in I_j} -\frac{\partial \Psi_i(\theta)}{\partial \theta^\top}|_{\theta=\hat{\theta}_j},\ V_j(\hat{\theta}_j) = \sum_{i \in I_j} \Psi_i(\theta)\Psi_i^\top(\theta)|_{\theta=\hat{\theta}_j},$$

as the sensitivity matrix and variability matrix, respectively, evaluated at a local site $j \in \{1, \ldots, K\}$.

We explain the basic idea of our proposed method starting with the incremental estimator for site 1 and site 2. Given that the $n_1 + n_2$ subjects are independent and identically distributed drawn from an underlying population of interest, we can write the overall log-likelihood as $\sum_{i \in I_1} l_i(\theta) + \sum_{i \in I_2} l_i(\theta)$ where $l_i(\theta)$ is the log-likelihood for one individual. We denote the oracle estimator obtained by maximizing the overall likelihood as $\hat{\theta}_2^{\text{ora}} := \arg\max_{\theta \in \Theta} \sum_{i \in I_2} l_i(\theta) + \sum_{i \in I_1} l_i(\theta)$, where $\Theta$ is the parameter space of $\theta$. After taking a Taylor series expansion of the first piece $\sum_{i \in I_1} l_i(\theta)$ at $\hat{\theta}_1$, which is a local estimate that solves an estimating equation $\sum_{i \in I_1} \Psi_i(\theta) = 0$ based on the local data $S_1$ of site 1, the incremental estimator at site 2 is defined as

$$\hat{\theta}_2 = \arg\max_{\theta \in \Theta} \sum_{i \in I_2} l_i(\theta) + \frac{1}{2}(\hat{\theta}_1 - \theta)^T H_1(\hat{\theta}_1)(\hat{\theta}_1 - \theta).$$

The two estimators $\hat{\theta}_2^{\text{ora}}$ and $\hat{\theta}_2$ should be in close proximity to each other and their differences are driven by the reminder terms that are asymptotically ignored in the above expansion. Intuitively, $\hat{\theta}_2^{\text{ora}}$ maximizes an augmented log-likelihood in that the likelihood for data $S_2$ of site 2 is regularized by a constraint of the Mahalanobis distance of the target parameter $\theta$ from the initial estimate $\hat{\theta}_1$. To compute $\hat{\theta}_2$ and its inferential quantities, after obtaining $\{\hat{\theta}_1, H_1(\hat{\theta}_1), V_1(\hat{\theta}_1)\}$ from site 1, COLA passes the relevant summary statistics to site 2 where $\hat{\theta}_1$ is updated to $\hat{\theta}_2$ by solving the following estimating equation[20]:

$$\Psi_{n_2}(\hat{\theta}_2) + H_1(\hat{\theta}_1)(\hat{\theta}_1 - \hat{\theta}_2) = 0,$$

where $\Psi_{n_2}(\hat{\theta}_2) = \sum_{i \in I_2} \Psi_i(\hat{\theta}_2)$. Repeating this sequential updating, COLA can be carried out over a sequence of all sites to produce incremental estimators and inferential quantities. In particular, when updating from $\hat{\theta}_{k-1}$ to $\hat{\theta}_k$ at site $k$, we solve the following estimating equation:

$$\Psi_{n_k}(\hat{\theta}_k) + \sum_{j=1}^{k-1} H_j(\hat{\theta}_j)(\hat{\theta}_{k-1} - \hat{\theta}_k) = 0. \tag{2}$$

Equation (2) consists of two parts: the first term $\Psi_{n_k}(\hat{\theta}_k) = \sum_{i \in I_k} \Psi_i(\hat{\theta}_k)$ is based on individual-level data at site $k$, and the second term assembles cumulative summary statistics from all previous $k-1$ sites, which is formed by the optimally weighted differences between a current update $\hat{\theta}_k$ and a previous estimate $\hat{\theta}_{k-1}$. This regularization procedure encourages the current update to be shrunken towards the previous estimate, in which the optimal scaling factor of the shrinkage is determined by the total Hassian matrix $\sum_{j=1}^{k-1} H_j(\hat{\theta}_j)$ that reflects the quality of the previous estimation. The Newton-Raphson algorithm is applied to numerically find a solution $\hat{\theta}_k$. To initiate the root-finding process, site 1 can take either naive initial estimates, that is, a vector of zeros, or more informative initial estimates derived through the iteratively reweighted least squares scheme which is widely applied to find the maximum likelihood estimates of parameters in generalized linear models.

For statistical inference, we first sequentially compute the cumulative sensitivity and variability matrices over the sequence of all $K$ sites evaluated at a given point estimate. Then we compute the sandwich variance estimator at the last site, that is, site $K$. For example, one can update the sensitivity and variability matrices along with the incremental point estimates. That is, at site $k$, the sensitivity matrix is given by $\sum_{j=1}^{k} H_j(\hat{\theta}_j)$ and the variability matrix is given by $\sum_{j=1}^{k} V_j(\hat{\theta}_j)$, where each $\hat{\theta}_j, j = 1, \ldots, k$ is sequentially updated over the first $j$ sites.

### 3.4 | Model diagnostics for homogeneous parameters in the propensity score model

In order to confirm homogeneous parameters in the propensity score model assumed in our COLA method, following Reference 21, we introduced a hypothesis test for the compatibility of site-specific propensity score estimating equations

using GMM.[22] This allows confirming that the covariates are sampled from the same underlying distribution. The null hypothesis is that there exists a common parameter $\gamma$ at which $E_\gamma\{\Psi^{\text{ps}}_{n_k}(\gamma)\} = 0$ for $k = 1, \ldots, K$. Let $\widetilde{\gamma} = \arg\min_{\gamma\in\mathbb{R}^p} Q(\gamma)$ be the common estimator for the fully combined data, where

$$
\begin{aligned}
Q(\gamma) &= \begin{bmatrix} \Psi^{\text{ps}}_{n_1}(\gamma) \\ \cdots \\ \Psi^{\text{ps}}_{n_K}(\gamma) \end{bmatrix}^\top \begin{bmatrix} V^{\text{ps}}_{n_1}(\gamma) & 0 & 0 \\ 0 & \cdots & 0 \\ 0 & 0 & V^{\text{ps}}_{n_K}(\gamma) \end{bmatrix}^{-1} \begin{bmatrix} \Psi^{\text{ps}}_{n_1}(\gamma) \\ \cdots \\ \Psi^{\text{ps}}_{n_K}(\gamma) \end{bmatrix} \\
&= \Psi^{\text{ps}}_{n_1}(\gamma)^\top V^{-1}_{n_1}(\gamma)\Psi^{\text{ps}}_{n_1}(\gamma) + \cdots + \Psi^{\text{ps}}_{n_K}(\gamma)^\top V^{-1}_{n_K}(\gamma)\Psi^{\text{ps}}_{n_K}(\gamma).
\end{aligned}
$$

Wang et al[21] showed that under some regularity conditions, $Q(\widetilde{\gamma})$ converges in distribution to $\chi^2_{(K-1)p}$, where $p$ is the dimension of $\gamma$. For model diagnosis, we compare $Q(\widetilde{\gamma})$ with $\chi^2_{(K-1)p,\alpha}$, where $\alpha$ is the statistical significance level. In our implementation, we use the COLA estimator $\hat{\gamma}_K$ to replace $\widetilde{\gamma}$, as they are asymptotically equivalent according to Theorem 3 (see below); see the detail of implementation in Web Appendix C in the supporting information. If the diagnostic results indicate that some sites violate the homogeneity assumption, then special attention is required to proceed with the proposed procedure: see specific guidelines discussed in the Web Appendix I.

## 4 | IMPLEMENTATION

We propose three algorithms to implement COLA that produce asymptotically equivalent estimators. The nuance lies in the trade-off between communication efficiency and finite-sample numerical accuracy. To clarify the differences between the three algorithms, we summarize their outputs at each round of updates in Figure 2. Here we use a varying number of "†" in the superscript of each estimator to represent the number of rounds used in each implementation of COLA, as detailed in Figure 2. For example, $\hat{\gamma}^{\ddagger}_K$ denotes the propensity score model parameter estimated from a three-round algorithm with three †'s in the superscript.

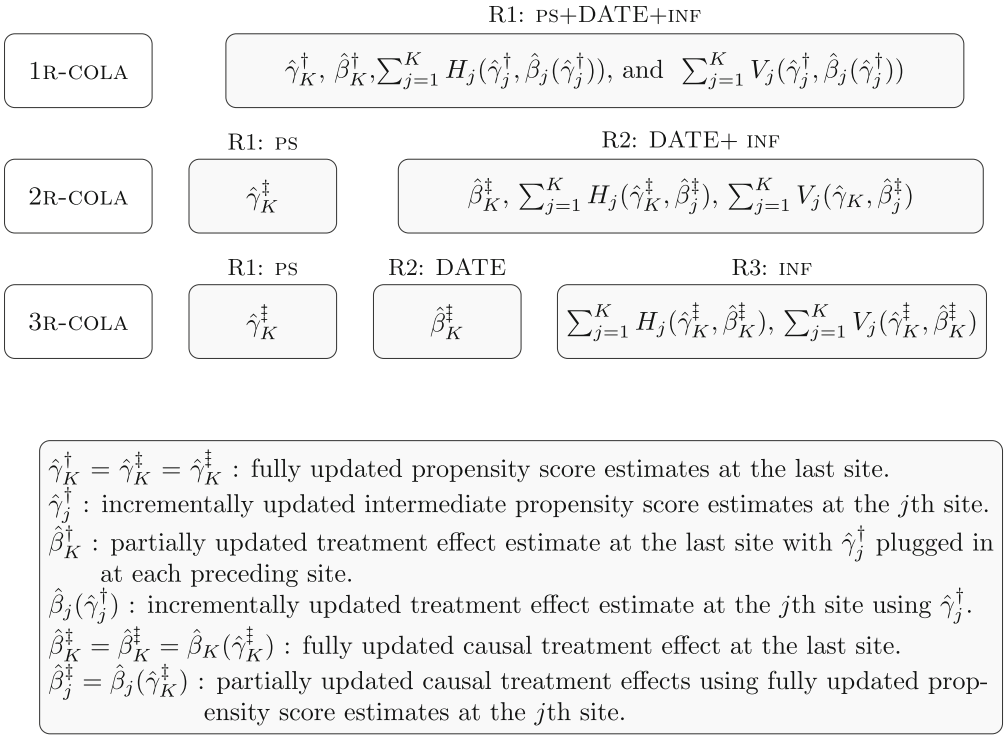


**FIGURE 2** Final outputs obtained after each round of update by four COLA methods.

### 4.1 | A one-round algorithm

We first introduce algorithm 1R-COLA which involves a single round of communication among participating sites, as shown in Figure 1A. It minimizes between-site communication at the price of possible reduced numerical stability and finite-sample performance. The one-round algorithm simultaneously updates the point estimate and variance-covariance matrix. That is, at site $k$, we compute $\hat{\gamma}_k^\dagger$ followed by $\hat{\beta}_k^\dagger(\hat{\gamma}_k^\dagger)$, and then update $\sum_{j=1}^k H_j(\hat{\gamma}_j^\dagger, \hat{\beta}_j^\dagger(\hat{\gamma}_j^\dagger))$, and $\sum_{j=1}^k V_j(\hat{\gamma}_j^\dagger, \hat{\beta}_j^\dagger(\hat{\gamma}_j^\dagger))$. Both point estimates and sandwich variances are different from 2R-COLA and 3R-COLA that are introduced below.

### 4.2 | A two-round algorithm

Given that a reliable estimate $\hat{\gamma}$ in the propensity score model is essential for proper estimation of treatment effect, the 1R-COLA algorithm above may be extended to a two-round algorithm, denoted by 2R-COLA illustrated in Figure 1B. This extension allows a full round of updates on the propensity score model before estimating the treatment effect and its inferential quantities.

Round 1: The first-round fits the propensity score model executing a full round of sequential update through all $K$ sites. We output the coefficient estimate $\hat{\gamma}_K^\ddagger$ at the last site, which is communicated back to all sites.
Round 2: While updating $\hat{\beta}_j^\ddagger = \hat{\beta}_j(\hat{\gamma}_K^\ddagger)$, the sensitivity and variability matrices, $\sum_{j=1}^k H_j(\hat{\gamma}_K^\ddagger, \hat{\beta}_j^\ddagger)$ and $\sum_{j=1}^k V_j(\hat{\gamma}_K^\ddagger, \hat{\beta}_j^\ddagger)$, are updated simultaneously using the current $\hat{\beta}_j^\ddagger$.

The pseudo-code for 2R-COLA is presented in the Appendix (Algorithm 1). An alternative two-round algorithm named 2R-COLA-INF estimates $\beta$ and $\gamma$ simultaneously in "Round 1" and only computes and updates variance-covariance matrices in "Round 2." The details for 2R-COLA-INF are given in Web Appendix D in the supporting information.

### 4.3 | A three-round algorithm

Our numerical experiences suggest that using a fully updated $\hat{\beta}_K^\ddagger$ in the calculation of sensitivity and variability matrices may gain some numerical stability than using the contemporaneous estimator $\hat{\beta}_j^\ddagger$ as shown in 2R-COLA. Then, we propose a three-round estimation algorithm, denoted by 3R-COLA, as shown in Figure 1C.

Round 1: The same "Round 1" of 2R-COLA is used to output $\hat{\gamma}_K^\ddagger$ which is the same as $\hat{\gamma}_K^\ddagger$.
Round 2: The second round sequentially updates the treatment effect $\hat{\beta}_j^\ddagger = \hat{\beta}_j(\hat{\gamma}_K^\ddagger)$ through all $K$ sites. This round outputs the estimated treatment effect, $\hat{\beta}_K^\ddagger = \hat{\beta}_K(\hat{\gamma}_K^\ddagger)$, which is communicated back to all sites.
Round 3: The third round estimates the asymptotic variance by sequentially updating the cumulative sums $\sum_{j=1}^K H_j(\hat{\gamma}_K^\ddagger, \hat{\beta}_K^\ddagger)$ and $\sum_{j=1}^K V_j(\hat{\gamma}_K^\ddagger, \hat{\beta}_K^\ddagger)$ overall $K$ sites, with $(\hat{\gamma}_K^\ddagger, \hat{\beta}_K^\ddagger)$ plugged in at each site.

2R-COLA and 3R-COLA produce the same point estimates for propensity scores model coefficients and the treatment effect, but different variance-covariance estimates. The finite-sample performance of statistical inference comparing 2R-COLA and 3R-COLA is investigated in the simulation study in Section 6.

## 5 | LARGE-SAMPLE PROPERTIES

Let $N_k$ be the cumulative sample size for the first $k$ sites, that is, $N_k = \sum_{j=1}^k n_j$. We discuss the large sample proprieties of our incremental estimators as $N_k = \sum_{j=1}^k n_k \to \infty$, instead of $\min_{j\in\{1,\ldots,k\}} n_j \to \infty$ in the parallel computing paradigm. This condition is satisfied when $n_k \to \infty$ at one of the sites, or when the number of sites $k \to \infty$. For simplicity, the former is discussed in detail in this article and we omitted "†" in the presentation. The asymptotic theory of 1R-COLA under $k \to \infty$ and fixed $n_j$'s may be shown in a similar way as in Reference 32 under extra conditions, and we leave detailed derivation to interested readers. Denote the $L^2$-norm of a vector $u$ by $\|u\|$. Let $N_\rho(\theta_0) = \{\theta : \|\theta - \theta_0\| \le \rho\}$, $\rho > 0$ be a

compact neighborhood of size $\rho$ around the true value $\theta_0$. In addition to Assumptions 1 and 2, we assume the following regularity conditions for the estimating function $\Psi$ given in Equation (1) to establish some key asymptotic properties.

**Assumption 3** (Regularity conditions). We assume the following on estimating function in Equation (1).

(a) The true value $\theta_0$ is the unique solution to $\lambda(\theta) = E\{\Psi(\theta)\} = 0$.
(b) The estimating function $\Psi(\theta)$ is continuously differentiable for all $\theta$ in the neighborhood $N_\rho(\theta_0)$.
(c) The sensitivity matrix $H(\theta)$ and the variability matrix $V(\theta)$ are positive definite for all $\theta \in N_\rho(\theta_0)$.
(d) The sensitivity matrix $H(\theta)$ is Lipschitz continuous for all $\theta \in N_\rho(\theta_0)$.

Condition 3a-c are mild regularity conditions needed for legitimate asymptotic behaviors of the COLA estimator $\hat{\theta}_k$ under the classical theory of estimating functions.[31,33] Condition 3d is usually satisfied for the generalized linear models.[34]

**Theorem 1.** *Under the regularity condition 3a-d, the cola estimator $\hat{\theta}_k$ is consistent for the true value $\theta_0$, that is, $\hat{\theta}_k \xrightarrow{p} \theta_0$, as $N_k \to \infty$.*

**Theorem 2.** *Under the regularity condition 3a-d, the cola estimator $\hat{\theta}_k$ is asymptotically normally distributed, that is, $\sqrt{N_k}(\hat{\theta}_k - \theta_0) \xrightarrow{d} N(0, J(\theta_0))$, as $N_k \to \infty$.*

**Theorem 3.** *Under the regularity condition 3a-d, the cola estimator $\hat{\theta}_k$ and the oracle estimator $\hat{\theta}^{\text{ora}}$ are asymptotically equivalent, in the sense that $\|\hat{\theta}_k - \hat{\theta}^{\text{ora}}\|^2 = o_p(N_k^{-1})$ as $N_k \to \infty$.*

The proofs of the above theorems are given in detail for 1R-COLA, 2R-COLA, and 3R-COLA respectively in the Web Appendix A. It follows from Theorem 2 that the proposed estimator takes advantage of the combined sample size; that is, the convergence rate of the COLA estimator $\hat{\theta}_k$ is $O_p(N_k^{-1/2})$, whereas the convergence rate of the local estimator at the $k$th site is $O_p(n_k^{-1/2})$. Note that the local convergence rate $O_p(n_k^{-1/2})$ is not equivalent to the cumulative convergence rate $O_p(N_k^{-1/2})$ unless $n_k/N_k = O(1)$ for all $k$. Reference 8 showed that this condition may be relaxed by allowing the sample size of the first data batch to be at the same order of $N_k$, under which the global Hessian may be replaced with the Hessian of the first site. However, in practice, determining which site's sample size matches the order of the total sample size is a challenging task; see detailed discussions of such challenge in Web Appendix B of the supporting information. Theorem 3 implies that the asymptotic difference between $\hat{\theta}_k$ and $\hat{\theta}^{\text{ora}}$ is $o_p(N_k^{-1/2})$, and thus they are stochastically equivalent in the sense that they have the same asymptotic normal distribution. This implies that the COLA estimator is fully efficient.

## 6 | SIMULATION EXPERIMENTS

We evaluate the finite-sample performance of the proposed collaborative inference method, comparing the above 3R-COLA, 2R-COLA, and 1R-COLA procedures with the classical meta-analysis (the inverse-variance weighted meta method[5]) and the oracle estimation (ie, the gold standard obtained by the centralized analysis). To mimic the motivating data example of the ITP-NODAT trial, we consider a binary outcome and estimate the marginal log odds ratio as DATE of interest. Additional simulations for continuous and count outcomes are included in the supporting information. We first generate the full data under an assumed model and then split the data into five subsets, one for a study site. We include three continuous variables $X_1, X_2$, and $X_3$ independently drawn from standard normal distribution $N(0, 1)$, and two independent binary covariates $X_4$ and $X_5$ from Bernoulli distribution with success probability of 0.5 and 0.6, respectively. We use $X$ to denote the vector of all five covariates. The treatment, $A$, follows Bernoulli distribution with $pr(A = 1|X) = \text{expit}(-1 + 0.3X_1 + 0.3X_2 + 0.5X_3 + 0.5X_4 + 0.3X_5)$, where $\text{expit}(x) = 1/(1 + e^{-x})$. The outcome, $Y$, is drawn from a Bernoulli distribution with $pr(Y = 1|A, X) = \text{expit}(-2.8 + 0.4A + 0.3X_1 + 0.5X_2 + 0.3X_3 + 0.3X_4 + 0.5X_5)$. The causal log odds ratio is estimated as 0.364 using the Monte-Carlo simulation of 1 000 000 random samples, and the proportion of cases ($Y = 1$) is approximately 30%. To simulate the scenarios of both unequal and equal proportions of cases across sites, we generate group indicators $I(j = 5)$ which follows the Bernoulli distribution with $pr\{I(j = 5)|Y\} = \text{expit}(a + bY)$, the probability that a sample belongs to the fifth site. The parameters $a$ and $b$ are predetermined such that we control the fifth site to have approximately 50 samples, of which 5% or 30% are cases. The sample size $n_5$ at the fifth site may not be exactly 50 because it is round to the next integer. We split the rest of the samples into sites 1 to 4 with sizes of 100, 80, 80,

and $100 - n_5$, respectively. That is, approximately we have site-specific sample sizes $n_1 = 100, n_2 = 80, n_3 = 80, n_4 = 50$, and $n_5 = 50$. We consider the following scenarios:

*scenario* 1. *Cases are equally distributed across all sites with 30% cases in each site, and the incremental estimator is sequentially updated over a pre-specified order of sites: 1, 2, 3, 4, 5.*

*scenario* 2. *Cases are unequally distributed among all sites, with the fifth site having a small proportion of cases (5%), while the overall proportion of cases is kept at 30%. The order of study sites remains the same.*

*scenario* 3. *Vary the order of study sites in sequential updates, while the proportion of cases at each site remains 30%.*

We evaluate the estimation and inference performances of different methods for estimating the DATE. Table 1 presents the simulation results under Scenarios 1 and 2 summarized over 30 000 replications. Scenario 1 is an ordinary setting where all sites have 30% cases and the sequential update starts with site 1 which has the largest sample size. As such, in scenario 1, the performances of our proposed three algorithms (1R-COLA, 2R-COLA, 3R-COLA) are similar to each other and also close to that of the oracle estimator. The numerical advantage of 3R-COLA over 2R-COLA is almost unnoticeable in terms of point estimates and variance, indicating they are both very close to the oracle estimator. The classical meta-analysis tends to be slightly unstable with a coverage probability (CP) of 91.3% and a few numerical failures.

In Scenario 2, a more challenging setting where the outcome is rare at site 5, the instability of meta-analysis becomes more evident. Specifically, meta-analysis suffers substantial numerical failures with 58.49% of the simulation replicates failing to reach convergence. In contrast, the rate of failures (<0.01%) of the proposed COLA methods is much smaller. The oracle estimation and the 3R-COLA algorithm produce very close CP and average absolute bias (ABIAS), which confirms the theoretical results in Section 5. In contrast, the CP of the meta-analysis estimation is 91.6%, much lower than the nominal level, and the ABIAS and empirical standard error (ESE) are larger for the competing methods.

In scenario 3, we test the robustness of our method to different orders of study sites in the sequential updates. As shown in Figure 3, because the fifth site has the smallest sample size, placing it as the starting site leads to the lowest CP and the highest numerical failure rates for all COLA methods, although the performances remain superior to meta-analysis. We also examine the robustness to different orders of study sites under Scenario 2 where site 5 is designed differently from the rest. When the distribution of cases is skewed, 3R-COLA shows the most robust performance to the ordering of sites and always achieves the "nominal" level coverage probability given that its inferential quantities are evaluated at the fully updated estimates. In contrast, 2R-COLA or 1R-COLA uses concurrent updates that appear to be more variable in intermittent steps. The details of the simulation results are shown in Web Appendix E.

**TABLE 1** Simulation results of estimating DATE under scenarios 1 and 2.

| Methods | Fails(%) | CP(%) | Abias $\times 10^{-3}$ | MSE $\times 10^{-3}$ | ESE $\times 10^{-3}$ |
|---|---|---|---|---|---|
| Scenario 1 | | | | | |
| *Oracle* | 0.00 | 94.7 | 217 | 264 | 273 |
| 3R-COLA | 0.00 | 94.7 | 216 | 262 | 271 |
| 2R-COLA | 0.05 | 94.8 | 216 | 262 | 271 |
| 1R-COLA | 0.05 | 94.1 | 225 | 266 | 283 |
| *Meta* | 3.86 | 91.3 | 238 | 258 | 298 |
| Scenario 2 | | | | | |
| *Oracle* | 0.00 | 94.5 | 214 | 262 | 270 |
| 3R-COLA | 0.00 | 94.5 | 213 | 261 | 268 |
| 2R-COLA | 0.01 | 93.0 | 213 | 244 | 268 |
| 1R-COLA | 0.01 | 92.4 | 222 | 249 | 282 |
| *Meta* | 58.49 | 91.6 | 237 | 261 | 297 |

Abbreviations: ABIAS, average absolute bias; CP, coverage probability; ESE, empirical standard error; FAILS, the percentage of non-convergence for incremental methods and the traditional meta-analysis method over 30 000 replications; MSE, median estimated standard error of the estimates.

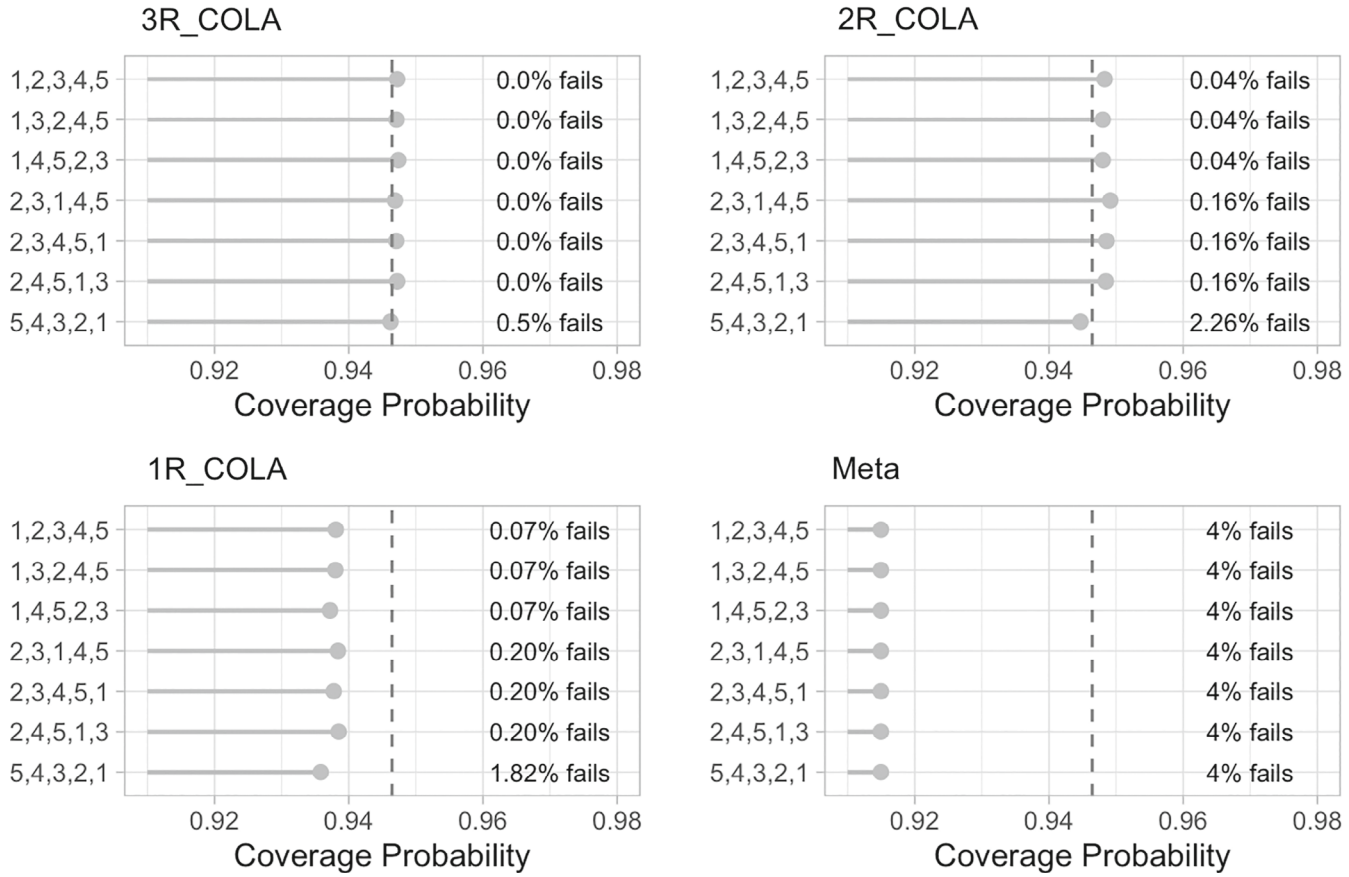


**FIGURE 3** Simulation results under scenario 3. Each panel presents coverage probabilities of a method under different orders of the five study sites as noted on the *y* axis. The dashed vertical line is placed at the coverage probability of the oracle method as a benchmark.

We conduct additional simulation studies in the supporting information under the following scenarios: (i) continuous and count outcomes (Web Appendix F), (ii) varying number of sites (Web Appendix G), (iii) a rare binary covariate at one of the sites which may lead to violation of positivity assumption (Web Appendix H). We observe similar performances and conclusions in Scenarios 1-3.

In summary, our simulation experiments lead to the following practical guidelines. In general, 3R-COLA is the top choice if the required communication cost is allowed as it provides higher accuracy than 2R-COLA and 1R-COLA. In addition, we suggest starting with the largest site to take advantage of large sample properties. If the outcome is rare or some covariates are highly skewed with potential violation of positivity at the largest site, then one can choose 3R-COLA which is less sensitive to such challenges than 2R-COLA and 1R-COLA. Alternatively, one can pick a new starting site with less severe distribution imbalances and sufficient sample size and proceed with either 2R-COLA or 1R-COLA, which requires less communication effort than 3R-COLA. Sometimes, data-sharing among smaller study sites might be less stringent in practice. In this case, one can merge a few smaller sites into a new site with combined data, which tends to improve performance. In addition, such a new site may serve as the starting site, when the largest site has less ideal quality, to conduct 2R-COLA or 1R-COLA algorithm.

We also evaluated the finite performance of the model diagnostics procedure proposed in Section 3.4. When the sample size is large, the procedure has shown a satisfactory control of Type I error (simulation results not shown). This is in agreement with the previous simulation studies in the literature; for example, References 21 and 32. When the sample size is small, the test appeared mildly anti-conservative with slightly inflated type I errors, resulting in protection for compatibility. In other words, this high-level vigilance makes the diagnostic method sensitive to heterogeneous data batches with the possibility of disqualifying some homogeneous data. In addition, to show that our proposed diagnostic procedure achieves reasonable power, we considered modifying the scenarios above in which one of the sites is set to have a distinct treatment-generating model from the rest of the sites in Web Appendix I of the supporting information.

## 7 | APPLICATION: IPT-NODAT MULTICENTER TRIAL

We apply the proposed collaborative inference method to analyze data from the ITP-NODAT trial. The primary goal of the ITP-NODAT trial is to estimate the DATE of basal insulin intervention in preventing overt PTDM at month 12 after the randomization in comparison to standard-of-care. Following the original analysis,[23] the population-average intent-to-treat effect of the basal insulin treatment is of interest. We include these covariates in the analysis: age, gender, family history of diabetes, whether the living donor or deceased donor, whether first-time transplantation or repeated transplantation, whether having polycystic kidney diseases, and whether having glomerular diseases. The proportion of missingness in the covariates is mild, thus we conduct a complete-case analysis excluding 42 dropouts and 8 participants following Schwaiger et al.[23]

We conduct COLA implementing the 2R-COLA algorithm without requiring subject-level data sharing. This analysis is particularly meaningful for the ITP-NODAT trial because pooling data from the four hospitals took three years to complete due to various cross-country data-sharing barriers. We also perform a centralized analysis of the pooled data as the gold standard for benchmarking, as well as the classical meta-analysis based on site-specific estimates for comparison.

The biggest challenge in the data analysis pertains to the unequal proportions of PTDM cases across hospitals, as shown in Web Figure 5 of the supporting information. There were zero PTDM cases recorded in the treatment group at the Barcelona hospital and zero PTDM cases in the control group at the Graz hospital, which prevented us from getting any site-specific results at Barcelona and Graz, leading to an exclusion of two out of four site-specific estimates. This data attrition is undesirable in classical meta-analysis.

Table 2 presents the estimated coefficients in the propensity score model obtained from the COLA method based on summary statistics and the centralized method based on the pooled data. The propensity score estimates obtained from the COLA method for the treatment and control groups overlap adequately as shown in Web Figure 5 of the supporting information. The model diagnostics procedure for the compatibility of site-specific propensity score estimating equations, although being anti-conservative, evidently confirmed the data combining with $Q(\hat{\gamma}_K) = 14.29$ ($P$-value $= 0.94$). Then we obtain an inverse probability weighted estimate of the odds ratio and its 95% confidence interval (CI). The estimated odds ratio is 0.37 (95% CI: 0.15, 0.93), which is very similar to the gold standard, 0.37 (95% CI: 0.15, 0.91). In contrast, the classical meta-analysis is based on site-specific estimates from two out of four study sites due to rare outcomes, and the meta-estimated odds ratio is 0.62 (95% CI: 0.20, 1.93) which underestimates the DATE by twice than the one from the centralized analysis.

It is evident from our analysis that basal insulin treatment reduces the risk of PTDM with an estimated odds ratio that is significantly less than one. This is in agreement with the findings reported in Reference 23, which performed a standard clinical trial analysis with no considerations of propensity score weighting. Little loss of statistical power occurred in our collaborative inference approach, while thoroughly overcoming data-sharing barriers and enjoying the maximal protection of data privacy. Had our method and analysis been available, these important clinical findings could have been published a few years earlier to add a new clinical treatment that benefits transplant patients. It is also worth noting that

**TABLE 2** Propensity score estimates for basal insulin treatment from combined data via centralized analysis and collaborative inference method.

| | **"Gold-standard" from combined data** | | | **Collaborative inference method** | | |
|---|---|---|---|---|---|---|
| | Estimates | Std. errors | $P$-value | Estimates | Std. errors | $P$-value |
| Gender | 0.13 | 0.30 | 0.72 | 0.13 | 0.30 | 0.67 |
| Age | −0.01 | 0.01 | 0.39 | −0.01 | 0.01 | 0.46 |
| Family diabetes history | 0.72 | 0.44 | 0.10 | 0.72 | 0.44 | 0.10 |
| First transplant | 0.03 | 0.46 | 0.97 | 0.02 | 0.46 | 0.95 |
| Glomerular disease | −0.33 | 0.32 | 0.31 | −0.33 | 0.32 | 0.32 |
| Polycystic kidney disease | −0.41 | 0.36 | 0.24 | −0.42 | 0.35 | 0.25 |
| Living doner | 0.86 | 0.45 | 0.05 | 0.87 | 0.45 | 0.05 |

*Note*: We use 2R-COLA algorithm to produce the results.

our proposed collaborative inference approach is not affected by imbalanced distributions of disease cases across study sites, a striking advantage over the classical meta-analysis method.

## 8 | CONCLUDING REMARKS

This article introduces a collaborative operation of linked analysis (COLA) framework that overcomes data-sharing barriers in the assessment of population average treatment effects. Although motivated by a DATE estimation problem, the proposed COLA by the means of estimating equations is general and applicable to a broad range of statistical analysis problems with data sharing challenges. We show the desirable asymptotic properties of the proposed distributed inference method. We also investigate the finite-sample performance through numerical experiments with four algorithms to implement COLA at different levels of communication costs. The simulation results show that little statistical efficiency is lost compared to the centralized method when estimating the DATE incrementally via 3R-COLA and 2R-COLA procedures. Even when the outcome is severely rare, 3R-COLA achieves similar results as the oracle method. Although we focus our attention on binary outcomes, our COLA framework enjoys the same properties and performance in the numerical illustrations for other outcome types under the generalized linear models as shown in Web Appendix F of the supporting information.

Meta-analytic types of causal inference methods in certain parallel computing diagrams can fail at two levels: local sites fail to converge and thus the pooled inference results fail to reflect the true parameters of the underlying population. Convergence failures occur when some sites do not have enough variability in outcome measurements. In practice, our COLA methods only require the first site to have enough data variability which makes it an appealing method for multi-center clinical trials that involve small study sites. To facilitate COLA in practice, we provide an R package for data analysis and an interactive information communication platform that allows each site to run the R program independently and upload and download the summary statistics via our platform (https://github.com/CollaborativeInference). The semi-manual process can be streamlined by the fast computation of the Lambda architecture[35] which is specifically designed for streaming data computing, adaptable for sequential computational algorithms such as our COLA method. Meanwhile, the emergence and development of federated learning which allows individual sites to collaboratively conduct data analysis while mitigating data privacy risks provides another useful arsenal to develop automated privacy-preserving software.[36-38] Furthermore, swarm learning which utilizes a decentralized blockchain-based machine-learning approach presents a promising platform to facilitate peer-to-peer networking while maintaining a high level of confidentiality through incorporating cutting-edge privacy-preserving architectures.[39] Without the need for a central coordinator, in a similar spirit to swarm learning, our COLA framework is decentralized with an emphasis on providing statistical inferences. Incorporating COLA to swarm learning architecture for enhanced data privacy and security is an interesting future work.

The current COLA framework is developed under a set of reasonable assumptions for identifiability and large-sample properties. Technically, the homogenous propensity score model assumption may be relaxed in the proposed COLA and the resulting COLA estimate of the treatment effect may remain asymptotically unbiased. This relaxation relies on the condition that all local sample sizes are sufficiently large so that the site-specific propensity score model produces reliable consistent PS estimates to the corresponding true chance of treatment allocation. For reasons discussed above, local propensity score models may fail to converge. Thus, we omit this extension for merely a theoretical interest. To verify the homogenous propensity score assumption, we propose a diagnostic test in Section 3.4 for data quality control purposes. When there is strong evidence that one or a few of the sites are flagged out for heterogeneity, one may simply exclude these sites and the analysis will be conducted on a smaller sample size. A way to overcome this potential challenge is to inflate the target sample size in the study design in order to absorb possible data attrition. During the development of our method, we considered a logistic parametrization for the propensity score model, as it is the typical choice in practical applications. However, other nonparametric approaches, such as tree-based or neural-net-based models, may potentially yield better results. Future research may investigate collaborative approaches for estimating parameters in these models and properly accounting for the uncertainty in treatment effect estimation such as incorporating the targeted maximum likelihood estimation.[40] Our method also has the potential for extension to settings where there are high-dimensional covariates available to model the propensity score. In such cases, model section procedure should be incorporated and post-variable selection inference should be carefully dealt with Reference 41. We also assumed the correct specification of the propensity score model which may be relaxed by employing a doubly robust method such as AIPW. Additional rounds of communication for estimation of nuisance parameters might be needed and we leave the details of implementation to

interested readers. Additionally, future work can focus on reducing communication burdens between sites by allowing a varying control of data privacy in different parallel problems. In recent causal inference method development, a line of research is primarily aimed at combining multiple datasets collected by different designs from potentially heterogeneous populations.[42-45] Most data fusion methods estimate causal treatment effects by incorporating patient-level data from auxiliary data sources into the main data source without consideration of data privacy issues. We plan to take advantage of the privacy-preserving nature of COLA and extend it to data fusion problems. Another interesting potential extension of our method is to transport our COLA estimation which targets the population underlying the current multi-center clinical trials to a new population.[46,15]

## ACKNOWLEDGEMENTS

The authors thank Manfred Hecking and Amelie Kurnikowski for providing ITP-NODAT trial data and their constructive comments on the data application. The authors thank Lan Luo for her comments on some technical details. Additionally, we thank the anonymous reviewers, Associate Editor, and Editor for their constructive feedback and insightful comments, which have greatly improved this article. Xu's research was partially funded by NIH R01GM139926, and Song's research was partially funded by NSF DMS2113564, NIH U24ES028502, and NIH R01ES033656.

## DATA AVAILABILITY STATEMENT

R code for simulating and conducting COLA for distributed clinical trial data is available online in the Supporting Information section. The real-world data from the IPT-NODAT trial are subject to European Union General Data Protection Regulation (EU GDPR) and are currently not available for public sharing.

## ORCID

*Xu Shi* https://orcid.org/0000-0001-8566-9552
*Peter X.-K. Song* https://orcid.org/0000-0001-7881-7182

## REFERENCES

1. Hernán MA, Brumback BA, Robins JM. Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Stat Med*. 2002;21(12):1689-1709.
2. Carter BL, Ardery G. Avoiding pitfalls with implementation of randomized controlled multicenter trials: strategies to achieve milestones. *J Am Heart Assoc*. 2016;5(12):e004432.
3. Mello MM, Francer JK, Wilenzick M, Teden P, Bierer BE, Barnes M. Preparing for responsible sharing of clinical trial data. *N Engl J Med*. 2013;369(17):1651-1658.
4. Coates EC, Mann-Salinas EA, Caldwell NW, Chung KK. Challenges associated with managing a multicenter clinical trial in severe burns. *J Burn Care Res*. 2020;41(3):681-689.
5. Cochran WG. The combination of estimates from different experiments. *Biometrics*. 1954;10(1):101-129.
6. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.
7. Morgan KL, Rubin DB. Rerandomization to improve covariate balance in experiments. *Ann Stat*. 2012;40(2):1263-1282. doi:10.1214/12-AOS1008
8. Jordan MI, Lee JD, Yang Y. Communication-efficient distributed statistical inference. *J Am Stat Assoc*. 2019;114(526):668-681. doi:10.1080/01621459.2018.1429274
9. Duan R, Boland MR, Moore JH, Chen Y. ODAL: a one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. *Pac Symp Biocomput*. 2018;24:30-41.
10. Duan R, Ning Y, Chen Y. Heterogeneity-aware and communication-efficient distributed statistical inference. *Biometrika*. 2022;109(1):67-83.
11. Zhou L, She X, Song PXK. Distributed empirical likelihood approach to integrating unbalanced datasets. *Stat Sin*. 2023;33:2209-2231.
12. Xiong R, Koenecke A, Powell M, Shen Z, Vogelstein JT, Athey S. Federated causal inference in heterogeneous observational data. *Stat Med*. 2023;42(24):4418-4439.
13. Vo TV, Hoang TN, Lee Y, Leong TY. Federated estimation of causal effects from observational data. *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence.* PMLR; 2022;180:2024-2034.
14. Zeng Z, Kennedy EH, Bodnar LM, Naimi AI. Efficient generalization and transportation. arXiv preprint arXiv:2302.00092, 2023.
15. Han L, Hou J, Cho K, Duan R, Cai T. Federated adaptive causal estimation (FACE) of target treatment effects. arXiv preprint arXiv:2112.09313, 2021.
16. Han L, Li Y, Niknam BA, Zubizarreta JR. Privacy-preserving, communication-efficient, and target-flexible hospital quality measurement. arXiv preprint arXiv:2203.00768, 2022.
17. Han L, Shen Z, Zubizarreta J. Multiply robust federated estimation of targeted average treatment effects. *Proceedings of Advances in Neural Information Processing Systems.* 2023;36.

18. Konečnỳ J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D. Federated learning: strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492, 2016.
19. McMahan B, Moore E, Ramage D, Hampson S, Arcas B. Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20 th International Conference on Artificial Intelligence and Statistics*. New York: PMLR; 2017:1273-1282.
20. Luo L, Song PXK. Renewable estimation and incremental inference in generalized linear models with streaming data sets. *J R Stat Soc Series B Stat Methodology*. 2020;82(1):69-97.
21. Wang F, Wang L, Song PXK. Quadratic inference function approach to merging longitudinal studies: validation and joint estimation. *Biometrika*. 2012;99(3):755-762.
22. Hansen LP. Large sample properties of generalized method of moments estimators. *Econometrica*. 1982;50:1029-1054.
23. Schwaiger E, Krenn S, Kurnikowski A, et al. Early postoperative basal insulin therapy versus standard of care for the prevention of diabetes mellitus after kidney transplantation: a multicenter randomized trial. *J Am Soc Nephrol*. 2021;32(8):2083-2098. doi:10.1681/ASN.2021010127
24. Neyman JS. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. (translated and edited by DM Dabrowska and TP Speed, statistical science (1990), 5, 465-480). *Ann Agric Sci*. 1923;10:1-51.
25. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. *J Am Stat Assoc*. 2005;100(469):322-331.
26. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89(427):846-866.
27. Godambe VP. An optimum property of regular maximum likelihood estimation. *Ann Math Stat*. 1960;31(4):1208-1211.
28. Godambe VP. *Estimating Functions*. Oxford, UK: Oxford University Press; 1991.
29. Stefanski LA, Boos DD. The calculus of M-estimation. *Am Stat*. 2002;56(1):29-38.
30. Freedman DA. On the so-called "Huber sandwich estimator" and "robust standard errors". *Am Stat*. 2006;60(4):299-302.
31. Song PXK. *Correlated Data Analysis: Modeling, Analytics, and Applications*. Berlin: Springer Science & Business Media; 2007.
32. Luo L, Zhou L, Song PXK. Real-time regression analysis of streaming clustered data with possible abnormal data batches. *J Am Stat Assoc*. 2022;118:2029-2044.
33. Tsiatis AA. *Semiparametric Theory and Missing Data*. New York: Springer; 2006.
34. McCullagh P, Nelder JA. *Generalized Linear Models*. London: Routledge; 2019.
35. Warren J, Marz N. *Big Data: Principles and Best Practices of Scalable Realtime Data Systems*. New York: Simon and Schuster; 2015.
36. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. *IEEE Signal Process Mag*. 2020;37(3):50-60.
37. Kairouz P, McMahan HB, Avent B, et al. Advances and open problems in federated learning. *Found Trends Mach Learn*. 2021;14(1–2):1-210.
38. Li L, Fan Y, Tse M, Lin KY. A review of applications in federated learning. *Comput Ind Eng*. 2020;149:106854.
39. Warnat-Herresthal S, Schultze H, Shastry KL, et al. Swarm learning for decentralized and confidential clinical machine learning. *Nature*. 2021;594(7862):265-270.
40. van der Laan MJ, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Vol 4. New York: Springer; 2011.
41. Wang F, Zhou L, Tang L, Song PX. Method of contraction-expansion (MOCE) for simultaneous inference in linear models. *J Mach Learn Res*. 2021;22(1):8639-8670.
42. Yang S, Ding P. Combining multiple observational data sources to estimate causal effects. *J Am Stat Assoc*. 2020;115(531):1540-1554. doi:10.1080/01621459.2019.1609973
43. Wang C, Lu N, Chen WC, et al. Propensity score-integrated composite likelihood approach for incorporating real-world evidence in single-arm clinical studies. *J Biopharm Stat*. 2020;30(3):495-507. doi:10.1080/10543406.2019.1684309
44. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *Proc Natl Acad Sci*. 2016;113(27):7345-7352. doi:10.1073/pnas.1510507113
45. Shi X, Pan Z, Miao W. Data integration in causal inference. *WIREs Computational Statistics*. 2023;15(1):e1581.
46. Dahabreh IJ, Petito LC, Robertson SE, Hernán MA, Steingrimsson JA. Toward causally interpretable meta-analysis: transporting inferences from multiple randomized trials to a new target population. *Epidemiology*. 2020;31(3):334-344.


## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Hu M, Shi X, Song PX-K. Collaborative inference for treatment effect with distributed data-sharing management in multicenter studies. *Statistics in Medicine*. 2024;43(11):2263-2279. doi: 10.1002/sim.10068

## APPENDIX A. PSEUDO CODE FOR 2R-COLA

For the convenience of programming, we consider site index $j \in \{0, 1, 2, \ldots, K\}$, where $j = 0$ does not correspond to any actual site and $N_0 = 0$. We use a superscript of "c" to denote a summary value of the control group and a superscript of "t" to denote the same summary value of the treatment group. With some abuse of notation, We use $\widehat{H}_{N_j}$ and $\widehat{V}_{N_j}$ to denote the cumulative sensitivity and variability matrices over the first $j$ sites. Similarly, we use $\widehat{H}^{\text{ps}}_{N_j}$ to denote the cumulative sensitivity matrix for the PS model at site $j$.

**Algorithm 1.** Two-round 2R-COLA implementation of collaborative inference algorithm

**Initialize:** $\hat{\gamma}_0 = 0_p, H_0 = 0_{p\times p}, \hat{\omega}^c_0 = 0, \hat{\omega}^t_0 = 0, \hat{v}^c_0 = 0, \hat{v}^t_0 = 0; \widehat{H}^{\text{ps}}_{N_0} = 0_{p\times p}, \widehat{H}_{N_0} = 0_{(p+2)\times(p+2)}, \widehat{V}_{N_0} = 0_{(p+2)\times(p+2)};$
**Round 1:for** $j \in \{1, 2, 3, \ldots, K\}$
**do**

**Input:** $A_j, X_j, \hat{\gamma}_{j-1},$ and $\widehat{H}^{\text{ps}}_{N_{j-1}}$

**Output:** $\hat{\gamma}_j$ and $\widehat{H}^{\text{ps}}_{N_j}$

**Run iteration** $r$ **until convergence to obtain** $\hat{\gamma}_j$ **:**

$$\hat{\gamma}_j^{(r)} - \hat{\gamma}_j^{(r-1)} = \left\{ \widehat{H}^{\text{ps}}_{N_{j-1}} + H^{\text{ps}}_j(\hat{\gamma}_j^{(r-1)}) \right\}^{-1} \left\{ \widehat{H}^{\text{ps}}_{N_{j-1}}(\hat{\gamma}_{j-1} - \hat{\gamma}_j^{(r-1)}) + \Psi^{\text{ps}}_{n_j}(\hat{\gamma}_j^{(r-1)}) \right\}$$

**Update:** $\widehat{H}^{\text{ps}}_{N_j} = \widehat{H}^{\text{ps}}_{N_{j-1}} + H^{\text{ps}}_j(\hat{\gamma}_j)$

**end**

**Output:** $\hat{\gamma}_K \sim\sim$ the global estimate of the PS model parameter, also denoted by $\hat{\gamma}^{\ddagger}_K$.

**Round 2:for** $j \in \{1, 2, 3, \ldots, K\}$

**do**

**Input:** $A_j, X_j, Y_j, \hat{\gamma}_K, \hat{\omega}^c_{j-1}, \hat{\omega}^t_{j-1}, \hat{v}^c_{j-1}, \hat{v}^t_{j-1}, \widehat{H}_{N_{j-1}},$ and $\widehat{V}_{N_{j-1}}$

**Output:** $\hat{\omega}^c_j, \hat{\omega}^t_j, \hat{v}^c_k, \hat{v}^t_j, \widehat{H}_{N_j}$, and $\widehat{V}_{N_j}$ $\hat{\omega}^c_j = \hat{\omega}^c_{j-1} + \sum_{i=1}^{n_j} \frac{1-A_{ji}}{\hat{e}(X_{ji};\hat{\gamma}_K)}, \quad \hat{\omega}^t_j = \hat{\omega}^t_{j-1} + \sum_{i=1}^{n_j} \frac{A_{ji}}{\hat{e}(X_{ji};\hat{\gamma}_K)}$

$\hat{v}^c_j = \hat{v}^c_{j-1} + \sum_{i=1}^{n_j} \frac{(1-A_{ji})Y_{ji}}{\hat{e}(X_{ji};\hat{\gamma}_K)}, \quad \hat{v}^t_j = \hat{v}^t_{j-1} + \sum_{i=1}^{n_j} \frac{A_{ji}Y_{ji}}{\hat{e}(X_{ji};\hat{\gamma}_K)}$

$\hat{\mu}^c_j = \hat{v}^c_j/\hat{\omega}^c_j, \qquad \hat{\mu}^t_j = \hat{v}^t_j/\hat{\omega}^t_j$

$\hat{\beta}_{Aj} = \log\{\hat{\mu}^t_j/(1-\hat{\mu}^t_j\} - \log\{\hat{\mu}^c_j/(1-\hat{\mu}^c_j)\}, \hat{\beta}_{0j} = \log\{\hat{\mu}^c_j/(1-\hat{\mu}^c_j)\}$

$\hat{\theta}_j = (\hat{\gamma}_K, \hat{\beta}_{0j}, \hat{\beta}_{Aj})^\top, \widehat{H}_{N_j} = \widehat{H}_{N_{j-1}} + H_j(\hat{\theta}_j), \widehat{V}_{N_j} = \widehat{V}_{N_{j-1}} + V_j(\hat{\theta}_j)$

**end**

**Final Output:** $\hat{\theta}_K$ and $\widehat{J}_{N_j} = \widehat{H}^{-1}_{N_j}\widehat{V}_{N_j}(\widehat{H}^{-1}_{N_j})^\top$