Journal of Machine Learning Research 26 (2025) 1-50 Submitted 9/23; Published 1/25

DisC^{2}o-HD: Distributed causal inference with covariates shift for analyzing real-world high-dimensional data

Jiayi Tong* JIAYI.TONG@PENNMEDICINE.UPENN.EDU
Department of Biostatistics, Epidemiology and Informatics
University of Pennsylvania
Philadelphia, PA 19104, USA

Jie Hu* JIE.HU@PENNMEDICINE.UPENN.EDU
Department of Biostatistics, Epidemiology and Informatics
University of Pennsylvania
Philadelphia, PA 19104, USA

George Hripcsak GH13@CUMC.COLUMBIA.EDU
Department of Biomedical Informatics
Columbia University
New York, NY 10027, USA

Yang Ning YN265@CORNELL.EDU
Department of Statistics and Data Sciences
Cornell University
Ithaca, NY 14853, USA

Yong Chen[†] YCHEN123@PENNMEDICINE.UPENN.EDU
Department of Biostatistics, Epidemiology and Informatics
University of Pennsylvania
Philadelphia, PA 19104, USA

**: The first two authors contribute equally to the paper.*
[†]*: Corresponding author*

Editor: David Sontag

Abstract

High-dimensional healthcare data, such as electronic health records (EHR) data and claims data, present two primary challenges due to the large number of variables and the need to consolidate data from multiple clinical sites. The third key challenge is the potential existence of heterogeneity in terms of covariate shift. In this paper, we propose a distributed learning algorithm accounting for covariate shift to estimate the average treatment effect (ATE) for high-dimensional data, named DisC^{2}o-HD. Leveraging the surrogate likelihood method, our method calibrates the estimates of the propensity score and outcome models to approximately attain the desired covariate balancing property, while accounting for the covariate shift across multiple clinical sites. We show that our distributed covariate balancing propensity score estimator can approximate the pooled estimator, which is obtained by pooling the data from multiple sites together. The proposed estimator remains consistent if either the propensity score model or the outcome regression model is correctly specified. The semiparametric efficiency bound is achieved when both the propensity score and the outcome models are correctly specified. We conduct simulation studies to demonstrate the performance of the proposed algorithm; additionally, we conduct an empirical study to present the readiness of implementation and validity.

Keywords: Causal Inference; Distribution Shift; Federated Learning; High-dimensional Data; Real-World Data

1. Introduction

Causal inference, which aims to elucidate the cause-effect relationships underlying the observed phenomena, usually relies on carefully designed experiments to establish causality (Hernán and Robins, 2010). However, in many domains, conducting controlled experiments may be unfeasible, leaving researchers to look for alternative methods. The increasing amount of real-world data (RWD) that captured the patients' clinical information offer a valuable opportunity for the researchers to investigate the causal relationships on a larger scale. By providing resourceful and rich observational data, the RWD shed light on building complex healthcare systems, inform evidence-based decision making, and drive advancements across diverse fields in addition to public health such as social sciences, economics, and beyond.

In the past few decades, the distributed research networks (DRNs) have been built to facilitate large-scale observational studies, covering large sample sizes and diverse populations, for example, the Observational Health Data Sciences and Informatics (OHDSI) consortium (Hripcsak et al., 2015), an international network of researchers and observation health databases, and the Patient-Centered Clinical Research Network (PCORnet) (Fleurence et al., 2014; Collins et al., 2014), which covers groups of diverse healthcare institutions and CRNs across the U.S. These research networks are highly valuable for clinical research by improving statistical power and enhancing the generalizability of the findings (Friedman et al., 2010). The growth of research networks has made it possible to analyze rare events and improve the accuracy of statistical models.

However, when utilizing large-scale RWD collected from CRNs for causal inference, there are three critical challenges to address. The first challenge revolves around the difficulty of sharing patient-level data, often due to privacy concerns and varying policy regulations in biomedical research (Behlen and Johnson, 1999). Sharing individual patient-level data can be time-consuming, logistically challenging, or infeasible in practice. The second major challenge arises when interest lies in conducting comparative effectiveness research via causal inference using RWD, where high-dimensional covariates collected in RWD are used to control the impact of confounders. Last but not least, the existence of population heterogeneity, also known as distribution shift or covariate shift, is also a key challenge to consider in practice. The differences in the underlying population could be caused by factors such as geographical variability in disease patterns, variations in patient characteristics, and regional differences in practice patterns. For example, there are studies using multiple electronic health records (EHR) datasets from Mayo Clinics and Vanderbilt University Medical Center (VUMC) to investigate the causal effects of candidate non-cancer drugs to be used for the treatment of cancer for drug repurposing (Xu et al., 2015; Wu et al., 2019). These studies successfully identified potential candidates for antineoplastic repurposing. A notable observation in these studies is that patient characteristics, including factors such as racial distribution and medication usage (such as insulin utilization), exhibit variations across the different sites. When conducting multi-site analyses in which patient-level data cannot be shared, it is essential to employ statistical methods that account for covariate shifts. Ignoring these differences can lead to biased estimates of causal relationships, an increased risk of overfitting, compromised generalizability of the findings, and potentially ineffective decision-making.

To address the first challenge in data sharing, a divide-and-conquer procedure is commonly used (Zhang et al., 2013; Lee et al., 2017; Battey et al., 2018). In particular, Battey et al. (2018) is one of the earliest innovations on distributed hypothesis testing for divide-and-concur estimator with high-dimensional data. After calculating and sharing the local estimators from the local patient-level data at each data site to the lead site or coordinating center, the final estimator is obtained by taking the average over the local estimators. Though sharing the estimators across sites mitigates the need in sharing patient-level data, the theoretical and empirical performance of this simple average method is suboptimal, especially when dealing with a large number of clinical sites and rare disease setting in multi-site studies (Duan et al., 2020b). In the past few years, an enhanced distributed learning, known as the surrogate likelihood approach, was proposed for association studies and prediction tasks (Wang et al., 2017; Jordan et al., 2018; Duan et al., 2018, 2020a,b, 2022). By requiring summary statistics from collaborating sites, the method is communication-efficient and privacy-preserving. In real-world settings, communication costs present a significant challenge, particularly in collaborative studies where transferring summary-level statistics demands considerable human labor. In response, it is essential to develop a communication-efficient distributed learning algorithm that minimizes communication rounds across sites.

In the context of addressing the second challenge posed by high-dimensional settings using RWD, considerable efforts have been dedicated to estimating the average treatment effect (ATE) in recent years. A number of notable methods have emerged, each of which presents innovative strategies. Belloni et al. (2014), Farrell (2015), Belloni et al. (2017), and Chernozhukov et al. (2018), have proposed a two-step approach. In this approach, they advocate first estimating the propensity score through penalized maximum likelihood and subsequently utilizing the efficient score function to estimate the ATE. Athey et al. (2018) introduced a different perspective by proposing approximate residual balancing. Notably, this approach eliminates the need for a propensity score model, while maintaining a requirement for linearity in covariates for the outcome model. This method was shown to be semiparametrically efficient and the balancing weights converge to the inverse propensity score but with a slower rate under suitable regularity conditions (Hirshberg and Wager, 2021). Bradic et al. (2019) contributed an estimator that excels in situations of rate/sparsity double robustness. The key advantage of this estimator is its root-n consistency even when either the propensity score model or the outcome model lacks sparsity, as long as the other model exhibits sufficient sparsity. Additionally, Tan (2020a) and Ning et al. (2020) proposed the high-dimensional covariate balancing propensity score method, which provides doubly robust confidence intervals for ATE involving high-dimensional covariates. These methods yield root-n consistent and asymptotically normal estimators, contingent on the accurate specification of either the propensity score model or the outcome model.

However, it is important to note that these methods are not specifically designed to accommodate scenarios where data are distributed across multiple clinical sites within research networks. Recent studies in this direction have led to the proposal of distributed learning algorithms for ATE estimation in causal inference (Xiong et al., 2021; Han et al., 2021), where the propensity score model and outcome model were used for causal effect estimation. Nevertheless, these methods cannot be applied directly to analyze high-dimensional data, as lasso-type estimators may not be aggregated directly in multisite settings (Battey et al., 2018).

In this paper, we propose a solution called DisC^{2}o-HD to simultaneously address all three challenges: data sharing, distributed causal inference for high-dimensional data, and covariate shift. Our method is specifically designed for analyzing real-world high-dimensional data and incorporates three key features:

- Firstly, DisC^{2}o-HD leverages the surrogate method to estimate the propensity score model and outcome model. Our proposed method can be implemented within a few rounds of communication, requiring only a single round of communication for participating sites to transfer summary-level statistics when fitting each model; all estimation iterations are conducted within a single local site or a designated lead site. This efficiency feature makes the proposed method more applicable to practical settings, enabling the generation of real-world data-based evidence.

- Secondly, our method effectively handles high-dimensional data to estimate average treatment effects (ATE). By properly integrating the rich information contained in high-dimensional data into our inferential procedure, we obtain more reliable estimates of causal effects. Furthermore, our method demonstrates robustness to model misspecifications. Even if either the propensity score model or the outcome model is incorrectly specified, our distributed algorithm produces results comparable to the gold standard method that relies on pooling patient-level data.

- Thirdly, DisC^2o-HD accounts for heterogeneous populations (e.g., differences in population or systematic variations) across multiple sites. It is capable of accommodating variations among different sites, ensuring the applicability of our method to diverse patient cohorts. Its robustness in handling the distributional differences makes it highly applicable in real-world situations.

Overall, DisC^{2}o-HD addresses the challenges of data sharing, distributed causal inference for high-dimensional data, and covariate shift. It offers a comprehensive and reliable approach for analyzing real-world high-dimensional data while protecting privacy and accommodating population heterogeneity across multiple sites.

We use the following notation. For $v = (v_1, \ldots, v_p) \in \mathbb{R}^p$ and $1 \leqslant q \leqslant \infty$, $\|v\|_0 = |\operatorname{supp}(v)|$, where $\operatorname{supp}(v) = \{j : v_j \neq 0\}$, and $\|v\|_q = \left(\sum_{i=1}^d |v_i|^q\right)^{1/q}$. The Orlicz norm associated with a Young's modulus ψ of X is defined by $\|X\|_\psi = \inf\{C > 0 : \mathbb{E}[\psi(|X|/C)] \leqslant 1\}$. $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ represent the minimal and maximal eigenvalues of A if we have a symmetric matrix A. For two positive sequences a_n and b_n, $a_n \asymp b_n$ if there exist $C, C' > 0$ such that $C \leqslant a_n/b_n \leqslant C'$ holds. For $\psi_1 = e^{x^2} - 1$, a random variable X is sub-Gaussian, if $\|X\|_{\psi_1} < \infty$. Denote $a \vee b = \max(a, b)$.

2. Methodology

Given the context of analyzing multisite data from research networks, we assume that individual patient data (IPD) are stored at K sites in a distributed manner and IPD cannot be shared across sites. Without loss of generality and for notation simplicity, we assume that each site has equal sample size n. We denote N as the total sample size of the pooled data from all K clinical sites (i.e., $N = Kn$). For the i-th patient from the k-th site, we observe (T_{ki}, Y_{ki}, X_{ki}), where $T_{ki} \in \{0, 1\}$ is the binary treatment assignment, Y_{ki} is the outcome variable, and X_{ki} is a p-dimensional vector of pre-treatment covariates.

We formulate the problem by considering the existence of covariate shifts (i.e., heterogeneous covariate distributions) across the sites. The term "covariate shift" refers to the variation in the distribution of the covariates across distinct sites. To provide a concrete illustration, suppose we have a single covariate, denoted as X, which is accessible across all sites. This covariate X follows a normal distribution at each site. The "covariate shift" scenario of X means the distribution of X at site A might have different mean and variance values compared to the distribution of X at another site. The definition of distribution shift is illustrated visually in Figure 1 and defined as follows:

Definition 1 *Let's denote by P and Q two probability measures associated with random variables X and X' respectively, defined on the same probability space (Ω, F), and suppose they admit density functions f and g. If there exists a set A such that for all x in A, $f(x) \neq g(x)$ and the Lebesgue measure of A (denoted by $\lambda(A)$) is not zero, then the density functions f and g (and thus the distributions P and Q) are different. This can be written formally as:*

$$\exists A \in F, \lambda(A) > 0 \text{ such that } \forall x \in A, f(x) \neq g(x)$$

This essentially means that there is a non-negligible set of outcomes where the two random variables X and X' have different densities.

The potential outcome under treatment and the potential outcome under control for the k-th site are denoted as $Y_{ki}(1)$ and $Y_{ki}(0)$, respectively. The observed outcome Y_{ki} is denoted as $Y_{ki} = T_{ki}Y_{ki}(1) + (1 - T_{ki})Y_{ki}(0)$. The parameter of interest by taking the covariate shift into account is the average treatment effect (ATE) over the K populations, defined as

$$\Delta^* = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[Y_{ki}(1) - Y_{ki}(0)\right] = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[Y_{ki}(1)\right] - \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[Y_{ki}(0)\right].$$

The terms $\tau_1^* := \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[Y_{ki}(1)\right]$ and $\tau_0^* := \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[Y_{ki}(0)\right]$ can be obtained in similar way by replacing the treatment assignment. Therefore, in the following subsections including derivations, equations, and algorithms, we will focus on the estimation of τ_1^*, which is the expected outcome for the treated cohort.

At k-th site, we consider the following logistic propensity score (PS) model and a linear outcome model (OM):

$$\mathbb{P}\left(T_{ki} = 1 \mid X_{ki}\right) = \pi\left(X_{ki}^T\theta\right), \tag{1}$$

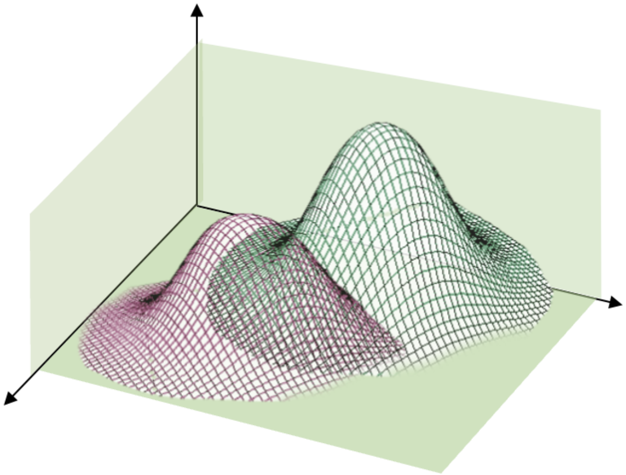

Figure 1: Illustration of distribution shift in 3-dimension plot, where mean values of the densities are the same, but the values of covariance matrices are different.

$$\mathbb{E}\left[Y_{ki}(1) \mid X_{ki}\right] = X_{ki}^T \beta, \tag{2}$$

where θ is a p-dimensional unknown vector that is homogeneous across all sites, $\pi(z) = 1/(1+\exp(-z))$, and β is a p-dimensional unknown vector. We assume that the PS model is homogeneous across all sites, and the same applies to the OM model. In this paper, we allow the models (1) and (2) to be misspecified. In the following sections, we will first present the algorithm and theoretical results by assuming both models are correctly specified. Then, we will further present the asymptotic distribution of the proposed estimator when either model is mis-specified in Section 3.4.

2.1 Background: Pooled method

If all of the patient-level data from K sites can be pooled together, two potential challenges exist: high-dimensional data and covariate shift. To address the first challenge, a variety of methods have been developed, as reviewed in the Introduction. To fix the idea, in this work we focus on the high-dimension covariate balancing method proposed by Ning et al. (2020). The first step is to estimate the propensity score via the following ℓ_1-penalized estimator

$$\hat{\theta}_{\text{pooled}} = \underset{\theta \in \mathbb{R}^p}{\arg\min} Q_N(\theta) + \lambda_{\text{pooled}}^{'} \|\theta\|_1, \tag{3}$$

where $Q_N(\theta) = \frac{1}{K}\sum_{k=1}^K Q_k(\theta)$ with

$$Q_k(\theta) = \frac{1}{n}\sum_{i=1}^n \left\{(1-T_{ki}) X_{ki}^T \theta + T_{ki} / \exp\left(X_{ki}^T \theta\right)\right\} \tag{4}$$

with $Q_k(\theta)$ being a generalized quasilikelihood function from the kth site, which is similar to the quasi-likelihood function for generalized linear models (Wedderburn, 1974) and was also used in Tan (2020b) and Tan (2020a) for fitting propensity score using high-dimensional data. In our study, the quasilikelihood function was chosen due to its alignment with the robust covariate balancing property exhibited by the corresponding quasi-score function (Ning et al., 2020). To establish the doubly robust confidence interval for ATE when analyzing high-dimensional data, the hyperparameter $\lambda_{\text{pooled}}^{'}$ in the (3) satisfied the following KKT condition

$$\left\| \frac{1}{Kn} \sum_{k=1}^{K} \sum_{i=1}^{n} \left\{ \frac{T_{ki}}{\pi\left(X_{ki}^{T} \hat{\theta}_{\text{pooled}}\right)} - 1 \right\} X_{ki} \right\|_{\infty} \leqslant \lambda_{\text{pooled}}^{'} .$$

This inequality implies that the maximum difference between the weighted average of X_{ki} in the treatment group and the population average of X_{ki} (e.g., $\frac{1}{Kn} \sum_{k=1}^{K} \sum_{i=1}^{n} X_{ki}$) is at most $\lambda_{\text{pooled}}^{'}$. Thus, the estimated propensity score $\pi(X_{ki}^{T} \hat{\theta}_{\text{pooled}}$) can approximately balance the covariates X_{ki} (Tan, 2020a; Ning et al., 2020).

After $\hat{\theta}_{\text{pooled}}$ is obtained, we then estimate the parameter through the following global loss function of the outcome model:

$$\hat{\beta}_{\text{pooled}} = \underset{\theta \in \mathbb{R}^{p}}{\arg\min} L_{N}(\beta, \hat{\theta}_{\text{pooled}}) + \lambda_{\text{pooled}}^{''} \|\beta\|_{1}, \tag{5}$$

where

$$L_{N}\left(\beta, \theta\right) = \frac{1}{Kn} \sum_{k=1}^{K} \sum_{i=1}^{N} \left\{ \frac{T_{ki}}{\exp\left(X_{ki}^{T} \theta\right)} \left(Y_{ki} - X_{ki}^{T} \beta\right)^{2} \right\} \tag{6}$$

is a weighted least square loss designed to achieve the desired doubly robust property. Finally, we obtain the AIPW estimate of $\tau_{1}^{*} = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\left[Y_{ki}(1)\right]$

$$\hat{\tau}_{1,\text{pooled}} = \frac{1}{Kn} \sum_{k=1}^{K} \sum_{i=1}^{n} \left\{ X_{ki}^{T} \hat{\beta}_{k} + \frac{T_{ki}}{\pi(X_{ki}^{T} \hat{\theta}_{k})} \left(Y_{ki} - X_{ki}^{T} \hat{\beta}_{k}\right) \right\}.$$

With the similar procedure, we estimate $\hat{\tau}_{0,\text{pooled}}$. Finally, we have:

$$\hat{\Delta}_{1,\text{pooled}} = \hat{\tau}_{1,\text{pooled}} - \hat{\tau}_{0,\text{pooled}}.$$

Regarding covariate shift, a notable aspect is that its presence does not affect the estimation of ATE in the pooled method. In other words, the existence of covariate shift does not introduce additional challenges when applying the introduced method for analyzing pooled data, as we assume the conditional distributions of Y are the same across all sites. Therefore, this method is treated as the gold standard method and will be compared with our method in simulation studies and data application.

2.2 Naive Method: Simple average method

Although $\hat{\theta}_{\text{pooled}}$ achieves the covariate balance property when analyzing the pooled data, in a distributed setting, where patient-level data cannot be shared across sites, the pooled estimator is not directly applicable. In such scenarios, an additional challenge arises: the decentralization of data. To tackle the complexities posed by high-dimensional and decentralized data analysis, a straightforward method is to aggregate the local estimates, which is a divide-and-conquer procedure. Specifically, each site fit the propensity score model and the outcome model:

$$\hat{\theta}_k = \underset{\theta \in \mathbb{R}^p}{\arg\min} Q_j(\theta) + \lambda' \|\theta\|_1, \tag{7}$$

$$\hat{\beta}_k = \underset{\theta \in \mathbb{R}^p}{\arg\min} L_k\left(\beta, \hat{\theta}_k\right) + \lambda'' \|\beta\|_1, \tag{8}$$

where λ' and λ'' are the regularization parameters, and

$$L_k(\beta, \theta) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{T_{ki}}{\exp(X_{ki}^T \theta)} \left(Y_{ki} - X_{ki}^T \beta\right)^2 \right\}.$$

Within each site, we obtain the local AIPW estimate of τ_1^*

$$\hat{\tau}_{1,k} = \frac{1}{n} \sum_{i=1}^{n} \left\{ X_{ki}^T \hat{\beta}_k + \frac{T_{ki}}{\pi(X_{ki}^T \hat{\theta}_k)} \left(Y_{ki} - X_{ki}^T \hat{\beta}_k\right) \right\}.$$

Finally, we aggregate the local AIPW estimators

$$\hat{\tau}_{1,\text{simple average}} = \frac{1}{K} \sum_{k=1}^{K} \hat{\tau}_{1,k}.$$

With the similar procedure, we estimate $\hat{\tau}_{0,\text{simple average}}$, and define $\hat{\Delta}_{1,\text{simple average}} = \hat{\tau}_{1,\text{simple average}} - \hat{\tau}_{0,\text{simple average}}$. Although this method requires only a single round of communications to combine the local estimates, which have a convergence rate of $1/\sqrt{n}$, and the relatively small local sample sizes (i.e., n) could potentially lead to biased local estimates, which in turn could affect the overall accuracy of the ATE.

2.3 Proposed method: DisC^{2}o-HD

In this paper, we present our method, DisC^{2}o-HD, uniquely designed to collectively tackle all three challenges mentioned earlier – high-dimensional data, decentralized data, and covariate shift. The resolution of all these challenges is essential to bring federated causal learning into practical utility. In particular, motivated by the Taylor expansion of the likelihood function (Jordan et al., 2018), we proposed to construct a robust surrogate function of $Q_N(\theta)$ shown in the pooled method:

$$\tilde{Q}(\theta, \bar{\theta}) := Q_1(\theta) + \left(\nabla Q_N(\bar{\theta}) - \nabla Q_1(\bar{\theta})\right)^T \theta + \frac{1}{2}(\theta - \bar{\theta})^T \left(\nabla^2 Q_N(\bar{\theta}) - \nabla^2 Q_1(\bar{\theta})\right)(\theta - \bar{\theta}), \tag{9}$$

where $\bar{\theta}$ is an initial estimator of θ, for example, a meta-analysis estimator which required one more round of communication or a local estimator from the lead site, $Q_1(\theta)$, $\nabla Q_1(\bar{\theta})$, and $\nabla^2 Q_1(\bar{\theta})$ are calculated within the 1-st site, also known as the lead site, assuming that we have full access to the patient-level data. For the rest of the sites from site 2 to K, they only need to communicate $\nabla Q_k(\bar{\theta})$, a p-dimensional vector, and $\nabla^2 Q_k(\bar{\theta})$, a $p \times p$ matrix. We note that unlike the surrogate likelihood proposed by Jordan et al. (2018), $\tilde{Q}(\theta, \bar{\theta})$ requires communicating the Hessian matrix $\nabla^2 Q_k(\bar{\theta})$, which is essential to account for the covariate shift. A key procedure in Jordan et al. (2018) involves substituting the global Hessian matrix $\nabla^2 Q_N(\theta)$, with a local Hessian matrix (e.g., Hessian matrix for Site 1), $\nabla^2 Q_1(\theta)$. This substitution relies on the assumption:

$$\|\nabla^2 Q_1(\theta) - \nabla^2 Q_N(\theta)\| < \delta = o(1)$$

This indicates that the local datasets should be homogeneous from site to site. However, in settings characterized by heterogeneous data distribution, such as those involving covariate shifts — which are of particular interest to our study — this assumption may not hold. In other words, with the presence of covariate shift:

$$\|\nabla^2 Q_1(\theta) - \nabla^2 Q_N(\theta)\| = O(1)$$

Faced with such a scenario, if we do not account for the covariate shift, an additional constant term would be induced in the error bound of the surrogate estimators. More details are provided in the Supplementary Appendix 3.7. Therefore, we proposed using the average of all Hessian matrices from collaborating sites, rather than replacing the global Hessian matrix solely with the local one. This method ensures that:

$$\left\| \frac{1}{K} \sum_{k=1}^{K} \nabla^2 Q_k(\theta) - \nabla^2 Q_N(\theta) \right\| = 0 < \delta = o(1)$$

As a result, by collecting the Hessian matrices from the collaborating sites, the proposed method can address the issue of covariate shifts across different sites. This incorporation of Hessians shows that the presence of covariate shift does not affect the estimation of the ATE when using high-dimensional data. Then, we obtain the penalized surrogate propensity score estimator through

$$\tilde{\theta} = \underset{\theta \in \mathbb{R}^p}{\arg\min} \tilde{Q}\left(\theta, \bar{\theta}\right) + \lambda_{\text{PS}} \|\theta\|_1,$$

where λ_{PS} is a regularization parameter.

After we obtain the estimator, $\tilde{\theta}$, from the propensity score function, we need to further fit the outcome model in a distributed manner. Similarly, we construct the surrogate loss function:

$$\begin{aligned} \tilde{L}(\beta, \bar{\beta}, \tilde{\theta}) = L_1(\beta) &+ \left(\nabla L_N(\bar{\beta}, \tilde{\theta}) - \nabla L_1(\bar{\beta}, \tilde{\theta})\right)^T \beta \\ &+ \frac{1}{2}(\beta - \bar{\beta})^T \left(\nabla^2 L_N(\bar{\beta}, \tilde{\theta}) - \nabla^2 L_1(\bar{\beta}, \tilde{\theta})\right)(\beta - \bar{\beta}), \end{aligned} \tag{10}$$

where $\bar{\beta}$ is an initial outcome model estimator. To construct $\tilde{L}(\beta, \bar{\beta}, \tilde{\theta})$ in (10), sites other than the lead site only need to contribute $\nabla L_k(\bar{\beta}, \tilde{\theta})$ and $\nabla^2 L_k(\bar{\beta}, \tilde{\theta})$. Then, we compute the penalized surrogate outcome model estimator

$$\tilde{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \tilde{L}\left(\beta, \bar{\beta}, \tilde{\theta}\right) + \lambda_{\text{OM}} \|\beta\|_1,$$

where λ_{OM} is a regularization parameter.

In order to enhance the theoretical analysis, we utilize the sample splitting technique within our algorithms, a widely employed approach in semiparametric models and causal inference (Chernozhukov et al., 2018; Newey and Robins, 2018; van der Laan et al., 2011). Specifically, splitting data enables us to achieve independence between $\tilde{\theta}$, $\tilde{\beta}$, and X_{ki} in the final AIPW estimator. Consequently, the classical Bernstein Inequality for sub-exponential sums can be applied separately to bound each of them. Given the dilemma of sharing patient-level data, splitting patient-level data across sites into several folds requires additional rounds of communications. Furthermore, splitting patient-level data across sites results in a smaller sample size for obtaining the initial estimators of $\bar{\theta}$ and $\bar{\beta}$, which in turn can affect the convergence rate of the final AIPW estimator. Instead, we split the K sites into three sets K_1, K_2, and K_3 with roughly equal size. This splitting strategy is shown to outperform the one by splitting patient-level data. For further discussion on the performance of splitting data versus not splitting it, as well as the comparison between splitting patient-level data and splitting sites, please refer to Section 4 in the supplementary material. In this section, we conduct additional numerical simulations to compare the performance between splitting and not splitting data, as well as between splitting patient-level data (split n) and splitting sites (split K).

To obtain the final average treatment effect (ATE), the following three steps are required.

- Step 1 involves conducting high-dimensional covariate balancing propensity score estimation using the aforementioned surrogate approach. The estimation is conducted in K_1, K_2, and K_3, respectively.
- Step 2 entails fitting the outcome model in a distributed manner, employing a similar surrogate likelihood function approach. The estimation is conducted in K_1, K_2, and K_3, respectively.
- Finally, in Step 3, we calculate the augmented inverse propensity weighted (AIPW) estimators from different splits and aggregate them to obtain the final ATE estimator.

For each step, we provide explicit algorithms, namely Algorithms 1, 2, and 3, respectively. In Algorithms 1 and 2, we show the algorithm in K_1 as an example. Same procedure is conducted for K_2 and K_3. The asymptotic variance of the final estimator $\tilde{\tau}_1$ in Algorithm 3, can be estimated through:

$$\widehat{V} = \frac{1}{Kn}\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{T_{ki}}{\pi\left(X_{ki}^T\tilde{\theta}\right)^2}\left(Y_{ki} - X_{ki}^T\tilde{\beta}\right)^2 + \frac{1}{Kn}\sum_{k=1}^{K}\sum_{i=1}^{n}\left(X_{ki}^T\tilde{\beta} - \tilde{\tau}_1\right)^2,$$

where $\tilde{\theta} = \left(\tilde{\theta}_{K_1} + \tilde{\theta}_{K_2} + \tilde{\theta}_{K_3}\right)/3, \tilde{\beta} = \left(\tilde{\beta}_{K_1} + \tilde{\beta}_{K_2} + \tilde{\beta}_{K_3}\right)/3$. The variance $\hat{V}$ can be computed in a distributed manner. In the following section, we present the theoretical justification of the proposed algorithm.

Remark 2 *A key advantage of our proposed method is its ability to remain unaffected by the presence of covariate shift when estimating the ATE using high-dimensional data. In particular, we construct the surrogates of the propensity score model and outcome model with* *second order. These second-order surrogate models make use of the gradients and Hessians of the global objective functions, and approximate the higher-order derivatives of the global objective function by the counterparts of the objective function at the leading site, while the first-order surrogate likelihood method (Wang et al., 2017; Jordan et al., 2018) only keeps the gradient of the objective function and approximates all higher-order derivatives by the alternatives at the leading site.*

As a comparison, the corresponding surrogate functions constructed from the idea in Jordan et al. (2018) are

$$\tilde{Q}^{original}(\theta, \bar{\theta}) = Q_1(\theta) + \left(\nabla Q_N(\bar{\theta}) - \nabla Q_1(\bar{\theta})\right)^T \theta, \tag{11}$$

$$\tilde{L}^{original}(\beta, \bar{\beta}, \tilde{\theta}) = L_1(\beta) + \left(\nabla L_N(\bar{\beta}, \tilde{\theta}) - \nabla L_1(\bar{\beta}, \tilde{\theta})\right)^T \beta. \tag{12}$$

We refer the method that only necessitates the collection of first gradients across multiple sites to as the original surrogate method hereafter. We can similarly estimate ATE by plugging the penalized estimators using the original surrogate method. In our simulation studies and data application, we will compare our proposed method, denoted as DisC^{2}o-HD-2, with this original surrogate method, denoted as DisC^{2}o-HD-1.

Algorithm 1 Distributed high-dimensional propensity score estimation on K_1

```
Require: $\{T_{ki}, Y_{ki}, X_{ki}\}$ for $i = 1, \ldots, n$, and $k \in K_1$.
    At site $k = 1$, calculate the initial propensity score estimator:
        $\bar{\theta}_{K_1} = \arg\min_{\theta \in \mathbb{R}^p} Q_1(\theta) + \lambda_{\text{PS,initial}} \|\theta\|_1,$
where $Q_1(\theta)$ is defined in Equation (4) and $\lambda_{\text{PS,initial}}$ is the initial regularization parameter.
    Broadcast $\bar{\theta}_{K_1}$ to all collaborating sites in $K_1$.
for site $k \in K_1$ do
    Compute the first gradient $\nabla Q_k(\bar{\theta}_{K_1})$ and second gradient $\nabla^2 Q_k(\bar{\theta}_{K_1})$. Broadcast these values to the leading site (i.e., $k = 1$).
end for
    Construct the surrogate loss $\tilde{Q}\left(\theta, \bar{\theta}_{K_1}\right)$ in Equation (9), where
        $\nabla Q_N\left(\bar{\theta}_{K_1}\right) = \frac{1}{|K_1|} \sum_{k \in K_1} \nabla Q_k\left(\bar{\theta}_{K_1}\right)$ and $\nabla^2 Q_N\left(\bar{\theta}_{K_1}\right) = \frac{1}{|K_1|} \sum_{k \in K_1} \nabla^2 Q_k\left(\bar{\theta}_{K_1}\right).$
Then, compute the penalized surrogate propensity score estimator
        $\tilde{\theta}_{K_1} = \arg\min_{\theta \in \mathbb{R}P} \tilde{Q}\left(\theta, \bar{\theta}_{K_1}\right) + \lambda_{\text{PS}} \|\theta\|_1,$
where $\lambda_{\text{PS}}$ is a regularization parameter.
```

Algorithm 2 Distributed high-dimensional outcome model on K_1

```
Require: $\{T_{ki}, Y_{ki}, X_{ki}\}$ for $i = 1, \ldots, n$, and $k \in K_1$.
    At site $k = 1$, calculate the initial outcome model estimator:
        $\bar{\beta}_{K_1} = \arg\min_{\theta \in \mathbb{R}^p} L_1\left(\beta, \tilde{\theta}_{K_2}\right) + \lambda_{\text{OM, initial}} \|\beta\|_1,$
    Broadcast $\bar{\beta}_{K_1}$ to all collaborating sites in $K_1$.
for site $k \in K_1$ do
    Compute the first gradient $\nabla L_k(\bar{\beta}_{K_1}, \tilde{\theta}_{K_2})$ and second gradient $\nabla^2 L_k(\bar{\beta}_{K_1}, \tilde{\theta}_{K_2})$.
    Broadcast these values to the leading site (i.e., $k = 1$)
end for
    Construct the surrogate loss as defined in Equation (10). Then, compute the penalized
    surrogate outcome model estimator
        $\tilde{\beta}_{K_1} = \arg\min_{\beta \in \mathbb{R}^p} \tilde{L}\left(\beta, \bar{\beta}_{K_1}, \tilde{\theta}_{K_2}\right) + \lambda_{\text{OM}} \|\beta\|_1,$
    where $\lambda_{\text{OM}}$ is a regularization parameter.
```

Algorithm 3 Calculation of AIPW estimators and final ATE

```
Require: $\{T_{ki}, Y_{ki}, X_{ki}\}$ for $i = 1, \ldots, n$, and $k = 1, \ldots, K$.
    Calculate $\tilde{\theta}_{K_1}$ by Algorithm 1 on $K_1$ and broadcast $\tilde{\theta}_{K_1}$ to all sites in $K_2$ and $K_3$.
    Calculate $\tilde{\beta}_{K_2}$ by Algorithm 2 on $K_2$ with $\tilde{\theta}_{K_1}$ and broadcast $\tilde{\beta}_{K_2}$ to all sites in $K_3$.
for site $k \in K_3$ do
    Calculate the AIPW estimator of $\tau_1^* = \mathbb{E}\left[Y_{ki}(1)\right]$
        $\tilde{\tau}_{1,k} = \frac{1}{n} \sum_{i=1}^{n} \left\{ X_{ki}^T \tilde{\beta}_{K_2} + \frac{T_{ki}}{\pi\left(X_{ki}^T \tilde{\theta}_{K_1}\right)} \left(Y_{ki} - X_{ki}^T \tilde{\beta}_{K_2}\right) \right\}.$
end for
    Aggregate the local AIPW estimators in $K_3$
        $\tilde{\tau}_{1,K_3} = \frac{1}{|K_3|} \sum_{k \in K_3} \tilde{\tau}_{1,k}.$
    Calculate $\tilde{\tau}_{1,K_1}$ and $\tilde{\tau}_{1,K_2}$.
    Calculate the final estimator of $\tau_1^*$:
        $\tilde{\tau}_1 = \left(\tilde{\tau}_{1,K_1} + \tilde{\tau}_{1,K_2} + \tilde{\tau}_{1,K_3}\right)/3$
```

3. Theoretical Results

3.1 Assumptions

In this section, we present and discuss the assumptions under which our theoretical results are proved.

Assumption 1 (Unconfoundedness) *The treatment assignment is unconfounded, i.e.,* $\{Y_{ki}(0), Y_{ki}(1)\} \perp\!\!\!\perp T_{ki}|X_{ki}$.

Assumption 2 (Overlap) *There exists a constant* $c_0 > 0$ *such that* $c_0 \leq \mathbb{P}(T_{ki} = 1|X_{ki}) \leq 1 - c_0$.

Assumption 7 requires that there is no unmeasured confounder. Assumption 8 implies that every sample has a positive probability to receive the treatment or belong to the control group. When Assumptions 7 and 8 are satisfied, the treatment assignment is considered as strongly ignorable (Rosenbaum and Rubin, 1983). The above two assumptions are standard in the causal inference literature.

Assumption 3 (Design) *The minimal and maximal eigenvalues of* $\mathbb{E}[X_{ki}X_{ki}^T]$ *are contained in a bounded interval that does not contain zero.*

Assumption 3 requires the design matrix is well conditioned. The same eigenvalue condition has been used to analyze high-dimensional lasso and causal inference problems (Van de Geer et al., 2014; Ning and Liu, 2017; Bradic et al., 2019). This assumption is utilized in Lemma 10, as detailed in the Supplementary Appendix Section 3.5 on the Restricted Strong Convexity (RSC) condition. It is a necessary condition for both the propensity score model and the outcome model in our proposed method. If site-specific covariates are present – for instance, a unique site indicator for each site – this assumption is violated, rendering the theoretical framework we have established for high-dimensional data inapplicable.

3.2 Restricted strong convexity(RSC) conditions

Assumption 4 (Model) X_{ki} *has a bounded sub-Gaussian norm. Moreover,* $\varepsilon_{ki}^* = Y_{ki}(1) - X_{ki}^T\beta^*$ *also has a bounded sub-Gaussian norm.*

Assumption 10 is a mild regularity condition on the tail of error term ε_{ki}^* and design X_{ki}. This assumption controls the behavior of the error term and enables us to use various concentration inequalities in high-dimensional statistics.

Assumption 5 (Sparsity) *Let* $s_1 = \|\theta^*\|_0$, *and* $s_2 = \|\beta^*\|_0$. *Assume that*

$$\frac{\sqrt{s_2(s_1 \vee s_2)}\log(p \vee Kn)}{\sqrt{Kn}} + \frac{s_1\sqrt{s_1 s_2 \log(p \vee Kn)\log^4(p \vee n)}}{n} = o(1)$$

as $s_1, s_2, p, K, n \to \infty$.

Assumption 11 imposes conditions on how fast the model sparsity s_1 and s_2, the covariate dimension p and the number of sites K can grow with the local sample size n. When $s_1 \asymp s_2 \asymp s$, up to some logarithmic factors, the condition reduces to $\frac{s}{\sqrt{Kn}} + \frac{s^2}{n} = o(1)$. In addition, when K is fixed, it further reduces to $s/\sqrt{n} = o(1)$, which is identical to the existing sparsity conditions for high-dimensional treatment effect estimation (Tan, 2020a; Ning et al., 2020). Finally, we comment that the assumption may still hold even if s_1 is large but s_2 is small (more precisely, $s_1 s_2$ is small), which is known as the sparsity double robustness property.

Assumption 6 (Variance) *We assume that there exists some constant* $c_1 > 0$ *such that* $\mathbb{E}(\varepsilon_{ki}^{*2}|X_{ki}) \geq c_1, \mathbb{E}(X_{ki}^T\beta^*)^4 = O(s_2^2)$.

Assumption 12 is a mild condition on the noise and design. The first assumption guarantees the nondegeneracy of the asymptotic variance, while the second part is used in the Lyapunov condition in CLT.

3.3 Asymptotic distribution when both models are correctly specified

In this section, we present our main results with respect to the propensity score model, potential outcome model and the proposed estimator $\tilde{\tau}_1$, respectively. For simplicity, we use C_L and M to denote some generic constants, whose values may differ from line to line.

For propensity score estimator $\tilde{\theta}_K$ obtained from Algorithm 1, we summarize its error bound in the following Proposition 3.

Proposition 3 *Under assumptions 1-6, with* $\lambda_{PS} \asymp \sqrt{\frac{\log(p\vee Kn)}{Kn}} + \frac{s_1\log^2(p\vee n)}{n}$, *we have:*

$$\left\|\tilde{\theta}_K - \theta^*\right\|_2 \leq C_L\left(\sqrt{\frac{s_1\log(p\vee Kn)}{Kn}} + \frac{s_1^{3/2}\log^2(p\vee n)}{n}\right),$$

$$\frac{3}{Kn}\sum_{(k,i)\in K}\left\{X_{ki}^T(\tilde{\theta}_K - \theta^*)\right\}^2 \leq C_L\left(\frac{s_1\log(p\vee Kn)}{Kn} + \frac{s_1^3\log^4(p\vee n)}{n^2}\right),$$

$$\frac{3}{Kn}\sum_{(k,i)\in K^c}\left\{X_{ki}^T(\tilde{\theta}_K - \theta^*)\right\}^2 \leq C_L\left(\frac{s_1\log(p\vee Kn)}{Kn} + \frac{s_1^3\log^4(p\vee n)}{n^2}\right).$$

holds with probability at least $1 - \frac{M}{(p\vee n)^8} - \frac{M}{n^8}$, *where* $K = K_1$, K_2, *or* K_3 *and* K^c *represents another set of servers.*

It can be seen that the error bound of $\tilde{\theta}_K$ consist of two terms. The first term is the classical lasso error bound when using the pooled data, signifying the optimal convergence rate achievable through this method. The second term can be represent by $s_1^{1/2}\log(p\vee n)\left\|\bar{\theta}_K - \theta^*\right\|_2^2$, which comes from the convergence rate of the initial local estimator $\bar{\theta}_K$. Consequently, it's generally smaller than that of the local estimator $\bar{\theta}_K$, which is $\sqrt{\frac{s_1\log(p\vee n)}{n}}$. While for the potential outcome estimator $\tilde{\beta}_K$, its error bound is provided as follows.

Proposition 4 *Under assumptions 1-6, with $\lambda_{OM} \asymp \sqrt{\frac{\log(p \vee Kn)}{Kn}}$, we have:*

$$\left\|\tilde{\beta}_K - \beta^*\right\|_2 \leq C_L \left(\sqrt{\frac{s_2 \log(p \vee Kn)}{Kn}}\right),$$

$$\frac{3}{Kn} \sum_{(i,j) \in K} \left\{X_{ki}^T(\tilde{\beta}_K - \beta^*)\right\}^2 \leq C_L \left(\frac{s_2 \log(p \vee Kn)}{Kn}\right),$$

$$\frac{3}{Kn} \sum_{(i,j) \in K^c} \left\{X_{ki}^T(\tilde{\beta}_K - \beta^*)\right\}^2 \leq C_L \left(\frac{s_2 \log(p \vee Kn)}{Kn}\right).$$

holds with probability at least $1 - \frac{M}{(p \vee n)^8} - \frac{M}{n^8}$, where $K = K_1$, K_2, or K_3 and K^c represents another set of servers.

We can see that the rate of $\widetilde{\beta}_K$ is the same as that of the pooled estimator, which is obtained by directed applying the penalized loss function to the pooled data with the sample size $Kn/3$. This is due to that the objective function (6) is in a weighted least squares form. On one hand, the error in $\widetilde{\theta}_K$ leads to a constant scaling of the weights, impacting only the constant scale but not the rate of convergence. On the other hand, a linear model has the lossless feature when using the surrogate likelihood method. This means that the distributed method can obtain exactly the same results as when the data are pooled together for analysis. Therefore, the convergence rate will not be affected by the initial local estimator $\bar{\beta}_K$. Specifically, the convergence rate of the Lasso estimator $\widetilde{\beta}_K$ mainly depends on the infinity norm of the gradient, $\|\nabla \tilde{L}(\beta^*, \bar{\beta}_K, \tilde{\theta}_{K^c})\|_\infty$. Given that $\tilde{L}(\beta^*, \tilde{\theta}_{K^c})$ represents a weighted least squares loss, by some calculation, it is easy to verify that $\nabla \tilde{L}(\beta^*, \bar{\beta}_K, \tilde{\theta}_{K^c}) = \nabla L_N(\beta^*, \tilde{\theta}_{K^c})$, thus the convergence rate of the initial estimator $\bar{\beta}_K$ does not affect the error bound of $\|\nabla L_N(\beta^*, \tilde{\theta}_{K^c})\|_\infty$. Consequently, the error bound of $\widetilde{\beta}_K$ shares the same order as that of the pooled estimator. Then, we present our main result on the bound of the proposed ATE estimator $\tilde{\tau}_1$ in the following two theorems.

Theorem 5 *Under assumptions 1-6, we have*

$$|\tilde{\tau}_1 - \hat{\tau}_1^*| \leq C_L \left(\frac{\sqrt{s_2(s_1 \vee s_2)}\log(p \vee Kn)}{Kn} + \frac{\sqrt{s_2(s_1 \vee s_2)\log(p \vee Kn)\log^4(p \vee n)}}{n\sqrt{Kn}}\right)$$

with probability at least $1 - \frac{M}{(p \vee n)^8} - \frac{M}{n^8}$, where

$$\hat{\tau}_1^* = \frac{1}{Kn} \sum_{k=1}^{K} \sum_{i=1}^{n} \left\{X_{ki}^T \beta^* + \frac{T_{ki}}{\pi\left(X_{ki}^T \theta^*\right)}\left(Y_{ki} - X_{ki}^T \beta^*\right)\right\}$$

is the asymptotic linear representation of the pooled AIPW estimator.

The first term in Theorem 12 represents the intrinsic error due to the estimation of nuisance parameters, which remains even if we are able to construct the pooled estimator by combining the data from multiple sites. Specifically, the order of the first term agrees with the findings of existing studies on the application of the hdCBPS estimator to pooled data with a sample size of Kn (Ning et al., 2020). The second term in Theorem 12 represents the cost incurred due to distributed learning. When $K = o(n)$ holds, our distributed estimator is equivalent to the global estimator in terms of the error bound. Furthermore, in Theorem 13, we present our main result on the Berry-Esseen bound for the proposed estimator $\tilde{\tau}_1$.

Theorem 6 *Under assumptions 1-6, we have*

$$\sup_{x\in\mathbb{R}}\left|\Pr\left(\frac{\sqrt{Kn}(\tilde{\tau}_1-\tau_1^*)}{\sqrt{\widehat{V}}}\le x\right)-\Phi(x)\right|$$
$$\le \frac{M}{(p\vee n)^8}+\frac{M}{n^8}+C_L\left(\frac{\sqrt{s_2(s_1\vee s_2)}\log(p\vee Kn)}{\sqrt{Kn}}+\frac{s_1\sqrt{s_1s_2\log(p\vee Kn)\log^4(p\vee n)}}{n}\right),$$

where C_L is a sufficiently large constant, and M depends on C_L.

Theorem 13 implies the asymptotic normality of $\tilde{\tau}_1$, from which we can construct valid confidence intervals and hypothesis tests for τ_1^*. Hahn (1998) proposed the semiparametric asymptotic variance bound for estimating τ_1^*, that is

$$V^* := \mathbb{E}\left[\frac{1}{\pi_{ki}^*}\mathbb{E}\left[\varepsilon_{ki}^2 \mid X_{ki}\right]+\left(X_{ki}^T\beta^*-\tau_1^*\right)^2\right],$$

where π_{ki}^* is the true value of the propensity score. We then show in Proposition 14 that the variance estimator defined in Theorem 13 consistently estimates V^*. Consequently, the proposed estimator $\tilde{\tau}_1$ achieves the semiparametric efficiency bound.

Proposition 7 (Consistency of variance estimator) *The variance estimator satisfies*

$$|\widehat{V}-V^*|\le C_L\left(\sqrt{\frac{(s_1\vee s_2)\log(p\vee Kn)}{Kn}}+\frac{s_1^{3/2}\log^2(p\vee n)}{n}\right),$$

with probability at least $1-\frac{M}{(p\vee n)^8}-\frac{M}{n^8}$.

Throughout the various error bounds mentioned above, we find that $\frac{\sqrt{s_2(s_1\vee s_2)}\log(p\vee Kn)}{\sqrt{Kn}}+\frac{s_1\sqrt{s_1s_2\log(p\vee Kn)\log^4(p\vee n)}}{n}$ dominates all other terms above. Therefore, in Assumption 5, we typically assume that $\frac{\sqrt{s_2(s_1\vee s_2)}\log(p\vee Kn)}{\sqrt{Kn}}+\frac{s_1\sqrt{s_1s_2\log(p\vee Kn)\log^4(p\vee n)}}{n}=o(1)$, which implies that all the other terms are also $o(1)$. Let $s = s_1\vee s_2$. Up to some logarithmic factors, this assumption can be reduced to $\frac{s}{\sqrt{Kn}}+\frac{s^2}{n}=o(1)$. This further simplifies to $s/\sqrt{n}=o(1)$ regardless of whether K is fixed or approaching infinity, which is identical to the existing sparsity conditions for high-dimensional treatment effect estimation Tan (2020a); Ning et al. (2020).

3.4 Asymptotic distribution when the models are misspecified

In this section, we first examine the robustness of the proposed estimator when the propensity score model is misspecified while the outcome model is correctly specified. In this setting, we assume that the true propensity score does not conform to the assumed parametric class, i.e., $\mathbb{P}\left(T_{ki}=1 \mid X_{ki}\right) \notin \left\{\pi\left(X_{ki}^{T}\theta\right) : \theta \in \mathbb{R}^{p}\right\}$. We define the estimand obtained through Algorithm 1 as follows:

$$\theta^{o} = \underset{\theta \in \mathbb{R}^{d}}{\arg\min} \mathbb{E}\left[Q_k(\theta)\right].$$

In the following, we demonstrate that the proposed estimator $\tilde{\tau}_1$ in Algorithm 3 is asymptotically equivalent to $\hat{\tau}^{o}_{1,\text{PS}}$ defined as follows

$$\hat{\tau}^{o}_{1,\text{PS}} = \frac{1}{Kn}\sum_{k=1}^{K}\sum_{i=1}^{n}\left\{X_{ki}^{T}\beta^{*} + \frac{T_{ki}}{\pi\left(X_{ki}^{T}\theta^{o}\right)}\left(Y_{ki} - X_{ki}^{T}\beta^{*}\right)\right\},$$

where β^{*} is the true value of β in the outcome model.

Proposition 8 *Under Assumptions 1-6, with θ^{*} replaced by θ^{o}, the proposed estimator satisfies*

$$|\tilde{\tau}_1 - \hat{\tau}^{o}_{1,PS}| \leq C_L\left(\frac{\sqrt{s_2(s_1 \vee s_2)}\log(p \vee Kn)}{Kn} + \frac{s_1\sqrt{s_1 s_2 \log(p \vee Kn)\log^{4}(p \vee n)}}{n\sqrt{Kn}}\right)$$

with probability at least $1 - \frac{M}{n^8}$, where C_L is a sufficiently large constant and M is another constant depending on C_L.

Since $\hat{\tau}^{o}_{1,\text{PS}}$ is asymptotically normal with mean τ_1^{*}, we can establish the asymptotic normality of the proposed estimator $\tilde{\tau}_1$. Consequently, the resulting confidence intervals remain valid even under the misspecified propensity score model. This provides justification for the robustness of the confidence intervals (Tan, 2020a; Ning et al., 2020).

We then can examine the robustness of the proposed estimator when the propensity score model is correctly specified while the outcome model is misspecified. In this scenario, we assume that the true potential outcome is nonlinear, meaning that $\mathbb{E}\left[Y_{ki}(1) \mid X_{ki}\right] \notin \left\{X_{ki}\beta : \beta \in \mathbb{R}^{p}\right\}$. Next, we define the estimand obtained through Algorithm 2 as follows:

$$\beta^{o} = \underset{\beta \in \mathbb{R}^{d}}{\arg\min} \mathbb{E}\left[L_k\left(\beta, \theta^{*}\right)\right],$$

where θ^{*} is the true value in the propensity score model. Define $\hat{\tau}^{o}_{1,\text{OM}}$ as

$$\hat{\tau}^{o}_{1,\text{OM}} = \frac{1}{Kn}\sum_{k=1}^{K}\sum_{i=1}^{n}\left\{X_{ki}^{T}\beta^{o} + \frac{T_{ki}}{\pi\left(X_{ki}^{T}\theta^{*}\right)}\left(Y_{ki} - X_{ki}^{T}\beta^{o}\right)\right\}.$$

Proposition 9 *Under Assumptions 1-6, with β^* replaced by β^o, the proposed estimator satisfies*

$$|\tilde{\tau}_1 - \hat{\tau}^o_{1,OM}| \leq C_L \left(\frac{\sqrt{s_2(s_1 \vee s_2)} \log(p \vee Kn)}{Kn} + \frac{s_1\sqrt{s_1 s_2 \log(p \vee Kn) \log^4(p \vee n)}}{n\sqrt{Kn}} \right)$$

with probability at least $1 - \frac{M}{n^8}$, where C_L is a sufficiently large constant and M is another constant depending on C_L.

Similarly, we demonstrate that the proposed estimator $\tilde{\tau}_1$ in Algorithm 3 is asymptotically equivalent to $\hat{\tau}^o_{1,\mathrm{OM}}$. This equivalence implies the robustness of confidence intervals even in the presence of misspecified outcome models.

4. Simulation studies

In this section, we examine the performance of the proposed DisC^{2}o-HD-2 estimator by comparing them with the pooled estimator, the local average method, and the DisC^{2}o-HD-1 estimator. Without loss of generality, for $k = 1, \ldots, K$ and $i = 1, \ldots, n$, the treatment T_{ki} are generated from a logistic regression with $\pi_{ki} = \mathrm{expit}(-0.5+0.5X_{ki1}+0.3X_{ki2}-0.3X_{ki3}+0.3X_{ki4}-0.3X_{ki5})$, the potential outcomes satisfy $Y_{ki}(1) = 2+0.3X_{ki1}+0.2X_{ki2}-0.2X_{ki3}+0.2X_{ki4}-0.2X_{ki5}+\epsilon_{ki1}$ and $Y_{ki}(0) = 1+0.3X_{ki1}+0.2X_{ki2}-0.2X_{ki3}+0.2X_{ki4}-0.2X_{ki5}+\epsilon_{ki0}$, where ϵ_{ki1} and ϵ_{ki0} are i.i.d from $N(0,1)$, while the p-dimensional covariates are generated from $\mathbf{X}_{ki} \sim N(0, \mathbf{\Sigma}_k)$. We consider the following seven scenarios.

(I) **Homogeneous covariates with $p < n$:** We consider the dimension with $p = 100$ and the sample size in each site is fixed at $n = 200$, while the covariance matrix $\mathbf{\Sigma}_k$ is set to be $\Sigma_{k;st} = 0.5^{|s-t|}$ for $k = 1, \ldots, K$. In this case, the simple size is larger than the dimension, and the distribution of covariates is homogeneous across sites.

(II) **Heterogeneous covariates (i.e., covariate shift) with $p < n$:** We consider the dimension with $p = 100$ and the sample size in each site is fixed at $n = 200$, while the covariance matrix $\mathbf{\Sigma}_k$ is set to be $\Sigma_{k;st} = \rho_k^{|s-t|}$, where $\rho_k \sim \mathrm{Uniform}(0.2, 0.8)$ for $k = 1, \ldots, K$. In this case, the simple size is larger than the dimension, and there is a shift in the distribution of covariates across sites.

(III) **Homogeneous covariates with $p > n$:** We consider the dimension with $p = 500$ and the sample size in each site is fixed at $n = 200$, while the covariance matrix $\mathbf{\Sigma}_k$ is set to be $\Sigma_{k;st} = 0.5^{|s-t|}$ for $k = 1, \ldots, K$. In this case, the simple size is smaller than the dimension, and the distribution of covariates is homogeneous across sites.

(IV) **Heterogeneous covariates with $p > n$:** We consider the dimension with $p = 500$ and the sample size in each site is fixed at $n = 200$, while the covariance matrix $\mathbf{\Sigma}_k$ is set to be $\Sigma_{k;st} = \rho_k^{|s-t|}$, where $\rho_k \sim \mathrm{Uniform}(0.2, 0.8)$ for $k = 1, \ldots, K$. In this case, the simple size is smaller than the dimension, and there is a shift in the distribution of covariates across sites.

(V) **Misspecified propensity score model with** $p > n$: We further consider the transformed covariates $\mathbf{X}_{ki,mis} = (X_{ki,1}, X_{ki,2}, X_{ki,3}^3, \exp(X_{ki,4}), X_{ki,5}(1 + \exp(X_{ki,6}))^{-2}, X_{ki,7}, \dots, X_{ki,p})$. In this case, the treatment T_{ki} is generated from the logistic regression in (IV) with $\mathbf{X}_{ki}$ replaced with the transformed covariates $\mathbf{X}_{ki,mis}$, while the potential outcomes are generated in the same way as in (IV).

(VI) **Misspecified outcome model with** $p > n$: We consider the same transformed covariates as in (V). In this case, the potential outcomes y_{ki} is generated from the linear regression in (IV) with $\mathbf{X}_{ki}$ replaced with the transformed covariates $\mathbf{X}_{ki,mis}$, while the treatments are generated in the same way as in (IV).

(VII) **Misspecified propensity score and outcome models with** $p > n$: We consider the same transformed covariates as in (V). In this case, both the treatment and potential outcomes are generated from the models in (IV)with $\mathbf{X}_{ki}$ replaced with the transformed covariates $\mathbf{X}_{ki,mis}$.

In each scenario, we repeat the simulation 100 times and vary the number of sites K in $\{10, 20, 30, 40, 50, 60\}$ to mimic research networks with moderate to large size, respectively. The regularization parameters in both algorithms are selected via cross-validation.

We compare the proposed DisC^{2}o-HD-2 approach and other approaches in terms of the root-mean-squared error (RMSE), absolute value of bias, variance under all scenarios. The comparison results of correctly specified cases are present in Figure 7 and 3. The four figures show that the proposed DisC^{2}o-HD-2 approach tends to have smaller bias and variance and hence have significantly smaller RMSE than that of other approaches except the pooled estimator in all scenarios. In addition, as the number of sites increases, the RMSE, bias, and variance of the proposed method decrease accordingly. However, the local average method and DisC^{2}o-HD-1 are less robust as the number of sites increases. In addition, the proposed DisC^{2}o-HD-2 approach is robust to the model misspecification and usually outperforms other methods as shown in Figure 6.

In summary, under the distributed setting, the proposed DisC^{2}o-HD-2 approach demonstrates superior performance and closely approximates the pooled results as the number of sites (K) increases. The proposed approach also exhibits more robust performance under model misspecifications.

5. Data application

A number of studies have been conducted to investigate the long-term consequences of SARS-CoV-2, the virus responsible for COVID-19. Post-acute sequelae of SARS-CoV-2, hereafter referred to as PASC, can manifest as various health issues affecting multiple organ systems, appearing four weeks or more after infection. The World Health Organization has defined post-COVID-19 conditions as those occurring three months after the initial infection, lasting a minimum of two months, and lacking an alternative diagnosis. However, there is limited information regarding the impact of the vaccine on PASC in diverse pediatric populations, particularly children in the United States.

In this section, we assess the proposed methods by utilizing a simulated data generated from summary statistics derived from electronic health record (EHR) data in existing studies

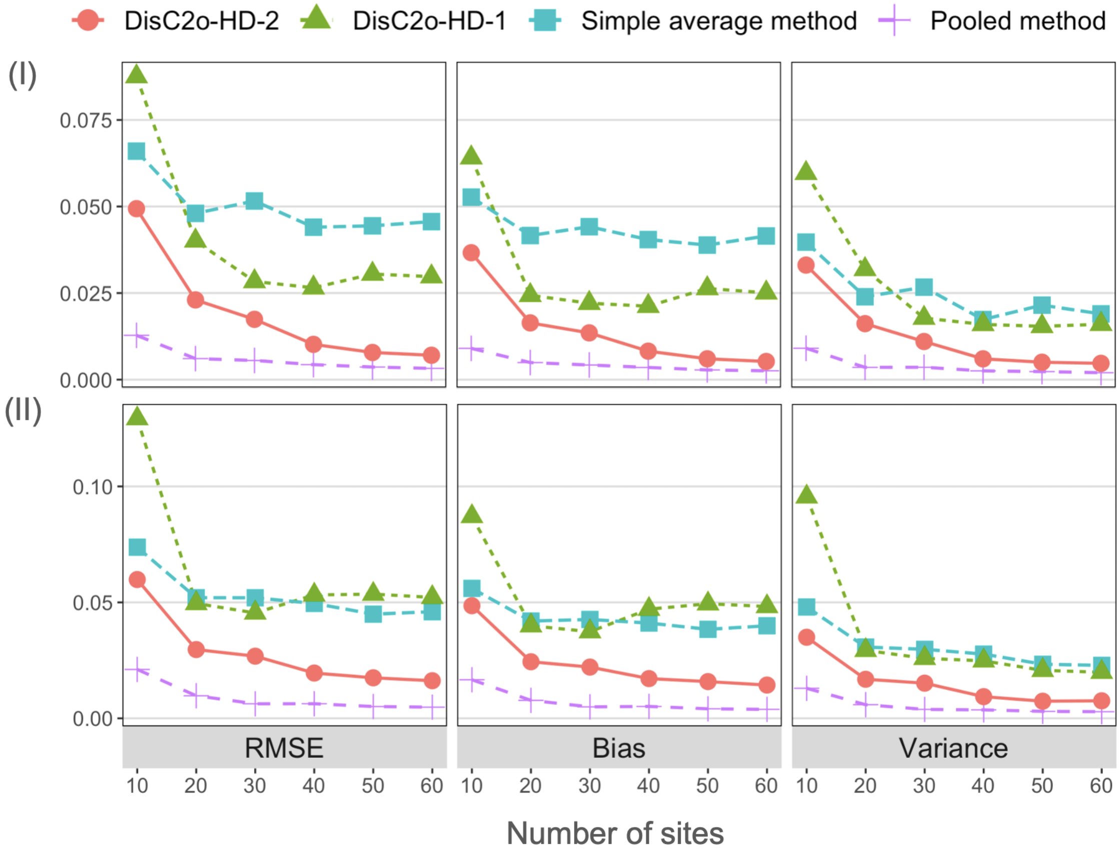

Figure 2: Simulation results for the lower-dimensional settings (i.e., $p < n$). Upper panel: comparison results of different methods under scenario (I) – homogeneous covariates with $p < n$; lower panel: comparison results of different methods under scenario (II) – heterogeneous covariates with $p < n$

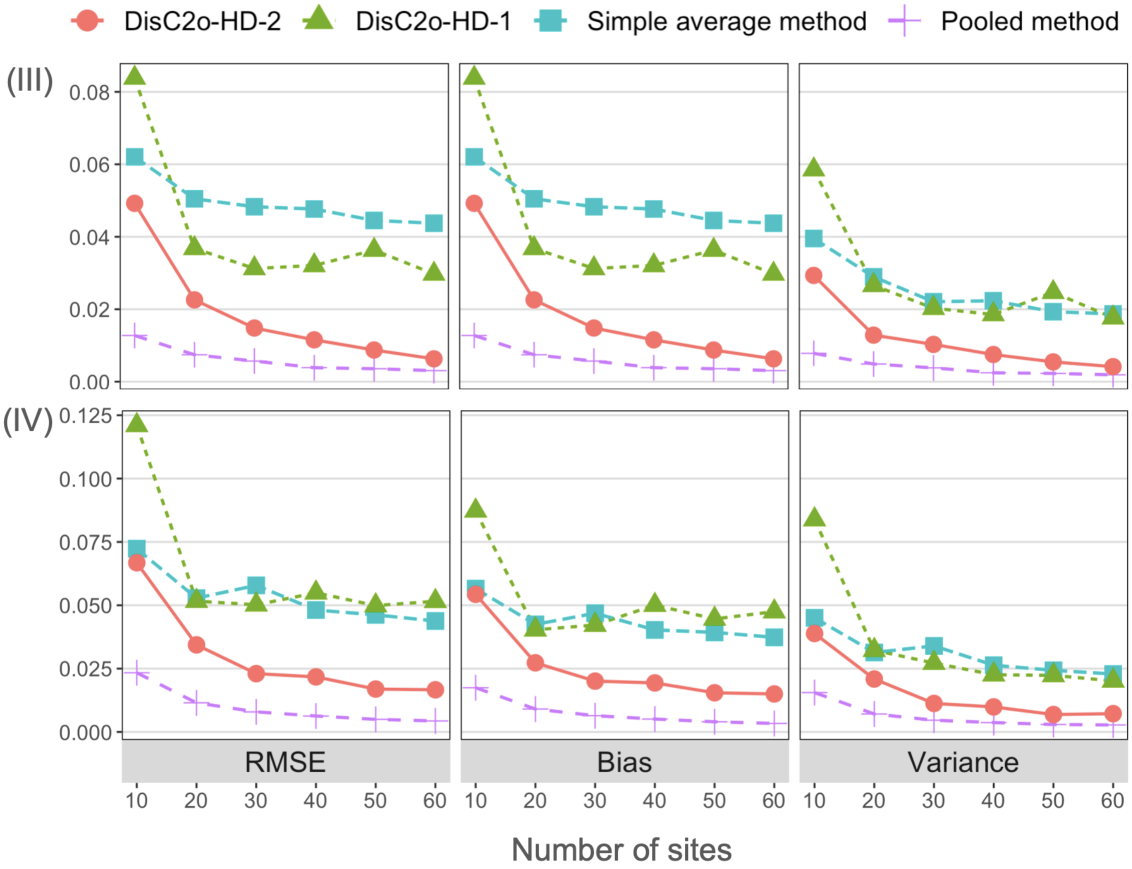

Figure 3: Simulation results for the high-dimensional settings (i.e., $p > n$). Upper panel: comparison results of different methods under scenario (III) – homogeneous covariates with $p > n$; lower panel: comparison results of different methods under scenario (IV) – heterogeneous covariates with $p > n$

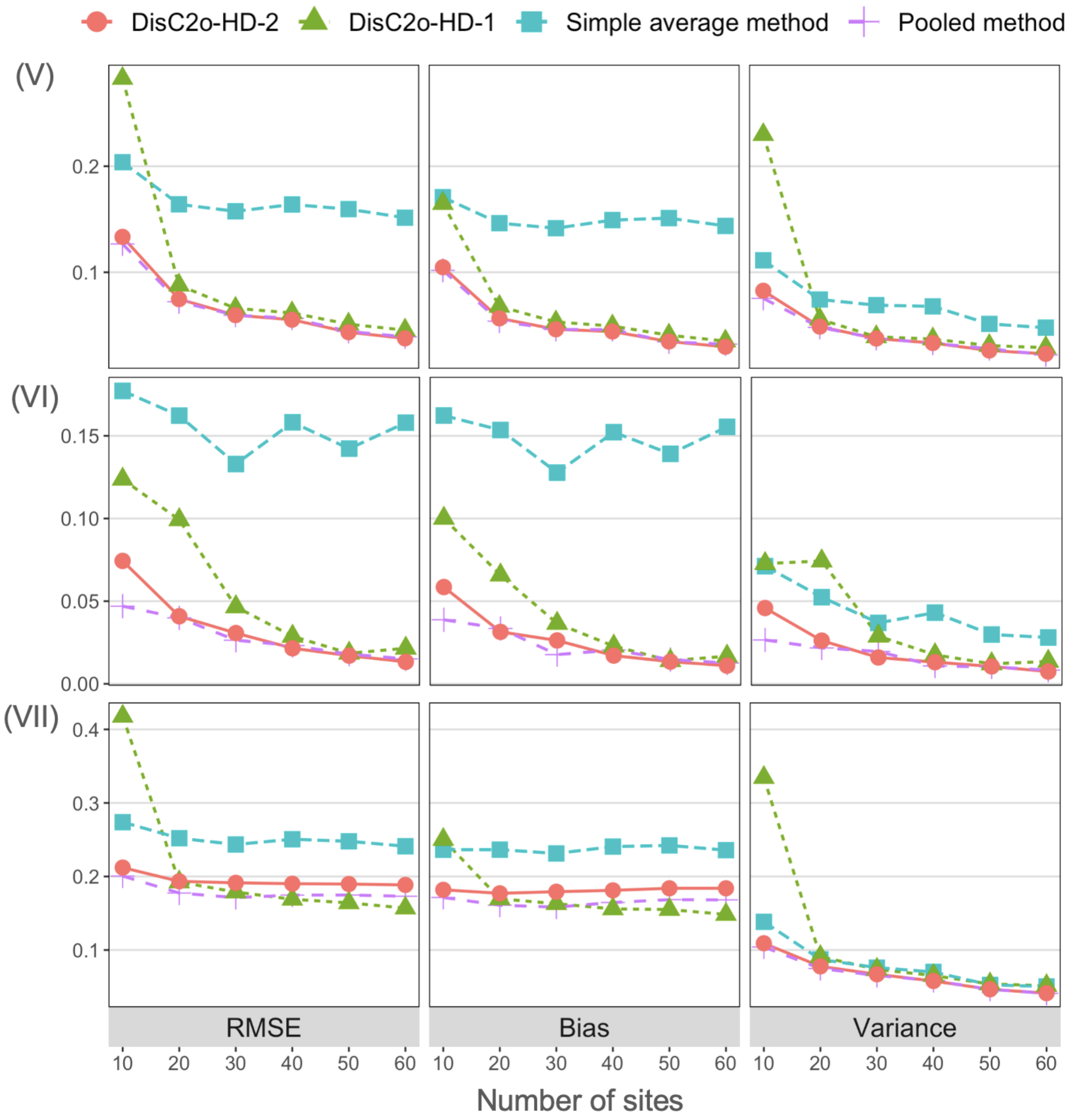

Figure 4: Simulation results for the high-dimensional settings with model misspecification. Upper panel: comparison results of different methods under scenario (V) – misspecified propensity score model with $p > n$; middle panel: comparison results of different methods under scenario (VI) – misspecified outcome model with $p > n$; lower panel: comparison results of different methods under scenario (VI) – misspecified propensity score model and outcome model with $p > n$

on the impact of vaccination on PASC in children during the Omicron period (Wu et al., 2024; Thaweethai et al., 2023). These EHR data included a comprehensive collection of routinely gathered clinical information, including patient demographics, medications, coded procedures, coded diagnoses, medical history, allergies, lab results, microbiology, blood bank data, pathology, vital signs, surgery, anesthesia, and more.

The eligibility criteria for participants in this study included being between the ages of 5 and 11 at the beginning of the study period, with no previous COVID-19 vaccination or documented SARS-CoV-2 infection. Additionally, participants were required to have a prior encounter within 18 months before entering the cohort to ensure an ongoing interaction with the healthcare system. The intervention under investigation was vaccination, specifically comparing those who received any type of COVID-19 vaccine with those who did not receive any. The outcome of interest was the count of PASC features observed within 28 to 179 days following the initial SARS-CoV-2 test date.

To mimic the patients' medical records in the EHR data, we simulated a set of confounders using the summary statistics reported in the existing studies. These confounders includes demographic variables such as age and gender, race, obesity status, Pediatric Medical Complexity Algorithm (PMCA) score (Simon et al., 2018), the number of visits to the emergency department in the 18 months leading up to 7 days before cohort entry, the number of inpatient visits during the same time frame, the number of outpatient visits within the specified period, the count of unique medications prescribed within the 18 months prior to 7 days before cohort entry, and the presence of diagnoses related to 205 chronic condition clusters within the same timeframe. In total, we included 248 confounders in our analysis.

Our final simulated dataset includes six clinical sites and consists of 1,158 individuals, with each site contributing approximately 193 patients to the analysis. We have complete access to all simulated 1,158 individuals, enabling us to apply both pooled analysis and the proposed method to the dataset. Figure 5 illustrates the results of the data analysis, comparing the pooled method, simple average method, DisC^{2}o-HD-1, and the proposed method. Among these methods, the proposed method (highlighted in red) yields the ATE estimation of -0.26 (95% CI: [-0.44, -0.08]) closest to that of the pooled method estimation of -0.24 (95% CI: [-0.38, -0.14]) (highlighted in purple), which is considered as the gold standard. It is worth noting that the proposed method exhibits a loss in efficiency, resulting in a wider confidence interval for the estimate. The results of this study by examining the impact of COVID-19 vaccination on children aged 5 to 11 showed consistent findings with previous studies conducted on adults (Wynberg et al., 2022). However, further investigations are warranted to better understand the effects of vaccination on children in this age group.

6. Conclusion

Overall, this paper presents a novel approach to address the challenges of high-dimensional healthcare data analysis, offering a distributed learning algorithm that effectively accounts for covariate shift and enables accurate estimation of the average treatment effect. The proposed method shows promise for improving healthcare research and decision-making by leveraging large-scale data from multiple clinical sites. The implementation of our proposed method requires a uniform set of covariates across all participating sites to ensure the validity of statistical analyses. However, we recognize that this requirement could limit its applicability in scenarios with structural missingness. Therefore, it is crucial for future extensions to address the presence of structural missingness in multi-site studies when analyzing high-dimensional data. Additionally, we plan to extend this method for other types of outcomes and address the potential issue of small sample sizes of the collaborating clinical sites. It is also important to note that although we assume the coefficients are homogeneous across sites in our models, these coefficients could indeed vary across different sites, particularly when analyzing complex, real-world, multi-site data. Some efforts have been made in the field of distributed learning. For example, Duan et al. (2022) proposed using the density ratio tilting method to accommodate differences in coefficients. We look forward to extending our current framework to more comprehensively understand and account for the potential variability of coefficients in the model. In addition to calculating the ATE at the population level, it is essential to assess the site-specific ATE. This analysis can yield significant insights into the factors driving heterogeneity across clinical sites.

Acknowledgments and Disclosure of Funding

We would like to thank the editor, David Sontag, for his invaluable guidance and feedback throughout the review process. We also thank anonymous reviewers for their constructive comments and suggestions, which have significantly enhanced the quality of this work. Additionally, we are grateful to Yiwen Lu for her valuable contributions to the visualization of the figures presented in this paper.

This work was supported in part by National Institutes of Health (U01TR003709, U24MH136069, RF1AG077820, 1R01LM014344, 1R01AG077820, R01LM012607, R01AI130460, R01AG073435, R56AG074604, R01LM013519, R01DK128237, R56AG069880, R21AI167418, R21EY034179). This work was supported partially through Patient-Centered Outcomes Research Institute (PCORI) Project Program Awards (ME-2019C3-18315 and ME-2018C3-14899). All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee.

Appendix A. Proofs and Technical Details

Recall that the loss functions for DisC^{2}o-HD estimator is defined as:

$$\tilde{Q}(\theta,\bar{\theta}) = Q_1(\theta) + \left(\nabla Q_N(\bar{\theta}) - \nabla Q_1(\bar{\theta})\right)^T \theta + \frac{1}{2}(\theta-\bar{\theta})^T \left(\nabla^2 Q_N(\bar{\theta}) - \nabla^2 Q_1(\bar{\theta})\right)(\theta-\bar{\theta}),$$

$$\begin{aligned}\tilde{L}(\beta,\bar{\beta},\tilde{\theta}_{K^c}) = L_1(\beta) &+ \Big(\nabla L_N(\bar{\beta},\tilde{\theta}_{K^c}) - \nabla L_1(\bar{\beta},\tilde{\theta}_{K^c})\Big)^T \theta \\ &+ \frac{1}{2}(\beta-\bar{\beta})^T \Big(\nabla^2 L_N(\bar{\beta},\tilde{\theta}_{K^c}) - \nabla^2 L_1(\bar{\beta},\tilde{\theta}_{K^c})\Big)(\theta-\bar{\theta}),\end{aligned}$$

respectively. Where $\bar{\theta}$ and $\bar{\beta}$ are proper initial estimator.

We first present and discuss the assumptions under which our theoretical results are proved.

Assumption 7 (Unconfoundedness) *The treatment assignment is unconfounded, i.e.,* $\{Y_{ki}(0), Y_{ki}(1)\} \perp\!\!\!\perp T_{ki} \mid X_{ki}$.

Assumption 8 (Overlap) *There exists a constant* $c_0 > 0$ *such that* $c_0 \leq \mathbb{P}(T_{ki} = 1 | X_{ki}) \leq 1 - c_0$.

Assumption 7 requires that there is no unmeasured confounder. Assumption 8 implies that every sample has a positive probability to receive the treatment or belong to the control group. When Assumptions 7 and 8 are satisfied, the treatment assignment is considered as strongly ignorable (Rosenbaum and Rubin, 1983). The above two assumptions are standard in the causal inference literature.

Assumption 9 (Design) *The minimal and maximal eigenvalues of* $\mathbb{E}[X_{ki} X_{ki}^T]$ *are contained in a bounded interval that does not contain zero.*

Assumption 3 requires the design matrix is well conditioned. The same eigenvalue condition has been used to analyze high-dimensional lasso and causal inference problems (Van de Geer et al., 2014; Ning and Liu, 2017; Bradic et al., 2019).

Assumption 10 (Model) X_{ki} *has mean* 0 *and a bounded sub-Gaussian norm. Moreover,* $\varepsilon^*_{ki} = Y_{ki}(1) - X_{ki}^T \beta^*$ *also has a bounded sub-Gaussian norm.*

Assumption 10 is a mild regularity condition on the tail of error term ε^*_{ki} and design X_{ki}. This assumption controls the behavior of the error term and enables us to use various concentration inequalities in high-dimensional statistics. Since the mean shift of X_{ki} does not influence the procedure of the proof, without loss of generality, we further assume that X_{ki} has a mean of 0 for the convenience of the proof, even though it may vary across sites.

Assumption 11 (Sparsity) *Let* $s_1 = \|\theta^*\|_0$, *and* $s_2 = \|\beta^*\|_0$. *Assume that*

$$\frac{\sqrt{s_2(s_1 \vee s_2)} \log(p \vee Kn)}{\sqrt{Kn}} + \frac{s_1 \sqrt{s_1 s_2 \log(p \vee Kn) \log^4(p \vee n)}}{n} = o(1)$$

as $s_1, s_2, p, m, n \to \infty$.

Assumption 11 imposes conditions on how fast the model sparsity s_1 and s_2, the covariate dimension p and the number of sites K can grow with the local sample size n. When $s_1 \asymp s_2 \asymp s$, upto some logarithmic factors, the condition reduces to $\frac{s}{\sqrt{Kn}} + \frac{s^2}{n} = o(1)$. In addition, when K is fixed, it further reduces to $s/\sqrt{n} = o(1)$, which is identical to the existing sparsity conditions for high-dimensional treatment effect estimation (Tan, 2020a; Ning et al., 2020). Finally, we comment that the assumption may still hold even if s_1 is large but s_2 is small (more precisely, $s_1 s_2$ is small). This is known as the sparsity double robustness property, recently proposed by Bradic et al. (2019).

Assumption 12 (Variance) *We assume that there exists some constant* $c_1 > 0$ *such that* $\mathbb{E}(\varepsilon^{*2}_{ki} | X_{ki}) \geq c_1$, $\mathbb{E}\left(X_{ki}^T \beta^*\right)^4 = O(s_2^2)$.

Assumption 12 is a mild condition on the noise and design. The first assumption guarantees the nondegeneracy of the asymptotic variance, while the second part is used in the Lyapunov condition in CLT.

Throughout this proof, we use the following notation. For $v = (v_1, \ldots, v_p) \in \mathbb{R}^p$ and $1 \le q \le \infty$, we define $\|v\|_q = \left(\sum_{i=1}^d |v_i|^q\right)^{1/q}$, $\|v\|_0 = |\text{supp}(v)|$, in which $\text{supp}(v) = \{j : v_K \neq 0\}$. The Orlicz norm associated with a Young's modulus ψ of X is defined by $\|X\|_\psi = \inf\{C > 0 : \mathbb{E}[\psi(|X|/C)] \le 1\}$. If a matrix A is symmetric, then $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ represent the minimal and maximal eigenvalues of A. For two positive sequences a_n and b_n, we write $a_n \asymp b_n$ if there exist $C, C' > 0$ such that $C \le a_n/b_n \le C'$ holds. We denote $\psi_1 = e^{x^2} - 1$ and if a random variable X is sub-Gaussian, then $\|X\|_{\psi_1} < \infty$. Denote $a \vee b = \max(a, b)$.

Throughout the proof below, we denote J to be a set of the servers, while K^c to be another set of servers. As an example, when J is the K_1, then K^c can be either K_2 or K_3.

A.1 Main result

Proposition 10 *Under assumption 1-6, with $\lambda_{ps} \asymp \sqrt{\frac{\log(p\vee Kn)}{Kn}} + \frac{s_1 \log^2(p\vee n)}{n}$, for DisC^{2}o-HD estimator $\tilde{\theta}_K$, we have:*

$$\left\|\tilde{\theta}_K - \theta^*\right\|_2 \le C_L \left(\sqrt{\frac{s_1 \log(p \vee Kn)}{Kn}} + \frac{s_1^{3/2} \log^2(p \vee n)}{n}\right),$$

$$\frac{3}{Kn} \sum_{(k,i)\in K} \left\{X_{ki}^T(\tilde{\theta}_K - \theta^*)\right\}^2 \le C_L \left(\frac{s_1 \log(p \vee Kn)}{Kn} + \frac{s_1^3 \log^4(p \vee n)}{n^2}\right),$$

$$\frac{3}{Kn} \sum_{(k,i)\in K^c} \left\{X_{ki}^T(\tilde{\theta}_K - \theta^*)\right\}^2 \le C_L \left(\frac{s_1 \log(p \vee Kn)}{Kn} + \frac{s_1^3 \log^4(p \vee n)}{n^2}\right).$$

holds with probability at least $1 - \frac{M}{(p\vee n)^8} - \frac{M}{n^8}$, where $K = K_1$, K_2, or K_3 and K^c represents another set of servers.

Proof First Claim:

Consider the events defined as:

$$\mathcal{E}_1 = \left\{\|\bar{\theta}_K - \theta^*\|_2 \le C_1 \sqrt{\frac{s_1 \log(p \vee n)}{n}}\right\},$$

$$\mathcal{E}_2 = \left\{\|\nabla Q_N(\theta^*)\|_\infty \le C_1 \sqrt{\frac{\log(p \vee Kn)}{Kn}}\right\},$$

$$\mathcal{E}_3 = \left\{\max_K \|X_{i1}\|_\infty \le C \log(p \vee n)\right\},$$

$$\mathcal{E}_4 = \left\{\frac{1}{n} \sum_{(i,1)\in J} \left\{X_{i1}^T(\bar{\theta}_K - \theta^*)\right\}^2 \le C_L \left\|\bar{\theta}_K - \theta^*\right\|_2^2\right\},$$

$$\mathcal{E}_5 = \left\{\frac{3}{Kn} \sum_{(k,i)\in K} \left\{X_{ki}^T(\bar{\theta}_K - \theta^*)\right\}^2 \le C_L \left\|\bar{\theta}_K - \theta^*\right\|_2^2\right\}.$$

For event $\mathcal{E}_1$, it is a result of lemma 17. While for $\mathcal{E}_2, \mathcal{E}_3$, we can apply lemma 21 and union bound. For $\mathcal{E}_4, \mathcal{E}_5$, they are the result from lemma 25. Plugging in the results above, it can be obtained that: $\mathbb{P}\left(\bigcap_{i=1}^5 \mathcal{E}_i\right) \geq 1 - \frac{M}{(p\vee n)^8} - \frac{M}{n^8}$.

Constrain $\|\delta\|_2 \leq C_1\sqrt{\frac{s_1 \log(p\vee n)}{n}}$, we can see that for the loss function of our DisC^{2}o-HD estimator satisfies

$$\begin{aligned}
&\tilde{Q}(\theta^* + \delta) - \tilde{Q}(\theta^*) - \nabla\tilde{Q}(\theta^*)\delta \\
&= Q_1(\theta^* + \delta) - Q_1(\theta^*) - \nabla Q_1(\theta^*)\delta + \frac{1}{2}\delta^T\left(\nabla^2 Q_N(\bar{\theta}_K) - \nabla^2 Q_1(\bar{\theta}_K)\right)\delta \\
&= Q_1(\theta^* + \delta) - Q_1(\theta^*) - \nabla Q_1(\theta^*)\delta - \frac{1}{2}\delta^T\nabla^2 Q_1(\theta^*)\delta + \frac{1}{2}\delta^T\nabla^2 Q_N(\bar{\theta}_K)\delta \\
&+ \frac{1}{2}\delta^T\left(\nabla^2 Q_1(\theta^*) - \nabla^2 Q_1(\bar{\theta}_K)\right)\delta \\
&= o(\|\delta\|_2^2) + \frac{1}{2}\delta^T\nabla^2 Q_N(\bar{\theta}_K)\delta + \frac{1}{2}\delta^T\left(\nabla^2 Q_1(\theta^*) - \nabla^2 Q_1(\bar{\theta}_K)\right)\delta \\
&\geq \mu\|\delta\|_2^2 - \mu'\frac{\log p}{n}\|\delta\|_1^2 + \frac{1}{2}\delta^T\left(\nabla^2 Q_1(\theta^*) - \nabla^2 Q_1(\bar{\theta}_K)\right)\delta + o(\|\delta\|_2^2),
\end{aligned}$$

where the last inequality is a direct result of lemma 19 and algebra. In addition,

$$\left|\frac{1}{2}\delta^T\left(\nabla^2 Q_1(\theta^*) - \nabla^2 Q_1(\bar{\theta}_K)\right)\delta\right| = \left|\frac{1}{n}\sum_{(i,1)\in J}\exp\left(-X_{i1}^T(\theta^* + t(\bar{\theta}_K - \theta^*))\right)X_{i1}^T(\bar{\theta}_K - \theta^*)\left(X_{i1}^T\delta\right)^2\right| = o(\|\delta\|_2^2).$$

Then, we have

$$\tilde{Q}(\theta^* + \delta) - \tilde{Q}(\theta^*) - \nabla\tilde{Q}(\theta^*)\delta \geq \mu\|\delta\|_2^2 - \mu'\frac{\log p}{n}\|\delta\|_1^2 + o(\|\delta\|_2^2) \geq \frac{\mu}{2}\|\delta\|_2^2 - \mu'\frac{\log p}{n}\|\delta\|_1^2.$$

Thus, $\tilde{Q}$ also satisfies the RSC condition given in Negahban et al. (2012). Then, by directly applying Corollary 1 of Negahban et al. (2012), we have:

$$\left\|\tilde{\theta}_K - \theta^*\right\|_2 \leq \frac{3\sqrt{s_1}\lambda_{\text{ps}}}{C},$$

for every $\lambda_{\text{ps}} \geq \left\|\nabla\tilde{Q}(\theta^*)\right\|_\infty$.

We can see that

$$\begin{aligned}
\nabla\tilde{Q}(\theta^*) &= \nabla Q_1(\theta^*) + (\nabla Q_N(\bar{\theta}_K) - \nabla Q_1(\bar{\theta}_K)) + (\nabla^2 Q_N(\bar{\theta}_K) - \nabla^2 Q_1(\bar{\theta}_K))(\theta^* - \bar{\theta}_K) \\
&= \nabla Q_1(\theta^*) - \nabla Q_1(\bar{\theta}_K) + \nabla Q_N(\bar{\theta}_K) - \nabla Q_N(\theta^*) \\
&+ \nabla Q_N(\theta^*) + (\nabla^2 Q_N(\bar{\theta}_K) - \nabla^2 Q_1(\bar{\theta}_K))(\theta^* - \bar{\theta}_K) \\
&= \nabla Q_N(\theta^*) + \left(\nabla^2 Q_1(\bar{\theta}_K) - \nabla^2 Q_1(\theta^* + t_1(\bar{\theta}_K - \theta^*))\right)(\bar{\theta}_K - \theta^*) \\
&+ \left(\nabla^2 Q_N(\theta^* + t_2(\bar{\theta}_K - \theta^*)) - \nabla^2 Q_N(\bar{\theta}_K)\right)(\bar{\theta}_K - \theta^*),
\end{aligned}$$

where $t_1, t_2 \in [0,1]$. Under $\mathcal{E}_4 \cap \mathcal{E}_5$, we can see:

$$\begin{aligned}
&\left\| \left(\nabla^2 Q_1(\bar{\theta}_K) - \nabla^2 Q_1(\theta^* + t_1(\bar{\theta}_K - \theta^*))\right)(\bar{\theta}_K - \theta^*) \right\|_\infty \\
&\le t_1 \left\| \frac{1}{n} \sum_{(i,1)\in J} \exp\left(-X_{i1}^T(\theta^* + s(\bar{\theta}_K - \theta^*))\right) X_{i1} \left(X_{i1}^T(\bar{\theta}_K - \theta^*)\right)^2 \right\|_\infty \\
&\le M \left\|X_{i1}\right\|_\infty \left| \frac{1}{n} \sum_{(1,j)\in J} \left\{X_{i1}^T(\bar{\theta}_K - \theta^*)\right\}^2 \right| \\
&\le MC \log(p \vee n) \left\|\bar{\theta}_K - \theta^*\right\|_2^2.
\end{aligned}$$

Notice that we can deal with the last term in a similar manner. Plugging it back in the equation we have, it shall be observed that:

$$\left\|\nabla \tilde{Q}(\theta^*)\right\|_\infty \le \left\|\nabla Q_N(\theta^*)\right\|_\infty + 2M' \log(p \vee n) \left\|\bar{\theta}_K - \theta^*\right\|_2^2.$$

Under event $\bigcap_{i=0}^6 \mathcal{E}_i$, we can see that:

$$\left\|\nabla \tilde{Q}(\theta^*)\right\|_\infty \le C_1 \sqrt{\frac{\log(p \vee Kn)}{Kn}} + C_2 \frac{s_1 \log^2(p \vee n)}{n}.$$

Then, with properly chosen λ_{ps}, under $\bigcap_{i=0}^6 \mathcal{E}_i$, we can see that:

$$\left\|\tilde{\theta}_K - \theta^*\right\|_2 \le C_L \left(\sqrt{\frac{s_1 \log(p \vee Kn)}{Kn}} + \frac{s_1^{3/2} \log^2(p \vee n)}{n} \right).$$

Second Claim:

Let $S = \{i : \theta^*_{Ji} \neq 0\}$. By definition, we can see that:

$$\tilde{Q}(\tilde{\theta}_K) + \lambda_{\mathrm{ps}} \left\|\tilde{\theta}_K\right\|_1 \le \tilde{Q}(\theta^*) + \lambda_{\mathrm{ps}} \left\|\theta^*\right\|_1, \tag{13}$$

$$\tilde{Q}(\tilde{\theta}_K) - \tilde{Q}(\theta^*) - \nabla \tilde{Q}(\theta^*)(\tilde{\theta}_K - \theta^*) + \left\|\tilde{\theta}_{JS^c}\right\|_1 \le -\nabla \tilde{Q}(\theta^*)(\tilde{\theta}_K - \theta^*) + \lambda_{\mathrm{ps}} \left(\left\|(\theta^* - \tilde{\theta}_K)_S\right\|_1 \right).$$

On the left hand side, we can see that

$$\begin{aligned}
&\tilde{Q}(\tilde{\theta}_K) - \tilde{Q}(\theta^*) - \nabla \tilde{Q}(\theta^*)(\tilde{\theta}_K - \theta^*) \\
&= \left(\nabla \tilde{Q}(\theta^* + t(\tilde{\theta}_K - \theta^*)) - \nabla \tilde{Q}(\theta^*)\right)(\tilde{\theta}_K - \theta^*) \\
&= t(\tilde{\theta}_K - \theta^*)^T \left(\nabla^2 \tilde{Q}(\theta^* + t'(\tilde{\theta}_K - \theta^*))\right)(\tilde{\theta}_K - \theta^*). \\
&= t(\tilde{\theta}_K - \theta^*)^T \left(\nabla^2 Q_1(\theta^* + t'(\tilde{\theta}_K - \theta^*)) + \nabla^2 Q_N(\bar{\theta}_K) - \nabla^2 Q_1(\bar{\theta}_K)\right)(\tilde{\theta}_K - \theta^*). \\
&= t \frac{3}{Kn} \sum_{(k,i)\in K} \left\{ \exp(-X_{ki}^T \bar{\theta}_K) \left(X_{ki}^T(\tilde{\theta}_K - \theta^*)\right)^2 + O\left(\left(X_{ki}^T(\tilde{\theta}_K - \theta^*)\right)^2 \left(X_{ki}^T(\bar{\theta}_K - \theta^*)\right) \right) \right\},
\end{aligned} \tag{14}$$

where $t, t' \in [0,1]$.

Furthermore, under assumption 7, the left hand side satisfies:

$$\tilde{Q}(\tilde{\theta}_K) - \tilde{Q}(\theta^*) - \nabla\tilde{Q}(\theta^*)(\tilde{\theta}_K - \theta^*) \geq C\frac{3}{Kn}\sum_{(k,i)\in K}\left\{X_{ki}^T(\tilde{\theta}_K - \theta^*)\right\}^2.$$

On the right hand side, under $\sum_{i=0}^{6}\mathcal{E}_i$, we have:

$$\begin{aligned}\tilde{Q}(\theta^*)(\theta^* - \tilde{\theta}_K) + \lambda_{\text{ps}}\left(\left\|(\theta^* - \tilde{\theta}_K)_S\right\|_1\right) &\leq \left\|\nabla\tilde{Q}(\theta^*)\right\|_\infty\left\|\theta^* - \tilde{\theta}_K\right\|_1 + \lambda_{\text{ps}}\left\|\theta^* - \tilde{\theta}_K\right\|_1 \\ &\leq C_L\left(\frac{s_1\log(p\vee Kn)}{Kn} + \frac{s_1^3\log^4(p\vee n)}{n^2}\right).\end{aligned}$$

Plugging in the result we have on left hand side, the desired result is thus obtained.

Third Claim:

By equation (14), we can see that under $\bigcap_{i=0}^{6}\mathcal{E}_i$,

$$\tilde{Q}(\tilde{\theta}_K) - \tilde{Q}(\theta^*) \geq \nabla\tilde{Q}(\theta^*)(\tilde{\theta}_K - \theta^*).$$

Since

$$|\nabla\tilde{Q}(\theta^*)(\tilde{\theta}_K - \theta^*)| \leq \left\|\nabla\tilde{Q}(\theta^*)\right\|_\infty\|\tilde{\theta}_K - \theta^*\|_1,$$

we can plug these facts in (13). Thus,

$$-\left(\left\|\nabla\tilde{Q}(\theta^*)\right\|_\infty\|\tilde{\theta}_K - \theta^*\|_1\right) + \lambda_{\text{ps}}\left\|\tilde{\theta}_K\right\|_1 \leq \lambda_{\text{ps}}\|\theta^*\|_1.$$

Denote $\frac{\lambda_{\text{ps}}}{\|\nabla\tilde{Q}(\theta^*)\|_\infty} = c$, we have $\left\|\nabla\tilde{Q}(\theta^*)\right\|_\infty = c\lambda_{\text{ps}}$. Since $\|\tilde{\theta}_K\|_1 = \left\|\theta^* + (\tilde{\theta}_K - \theta^*)_{S^c} + (\tilde{\theta}_K - \theta^*)_S\right\| = \left\|\theta^* + (\tilde{\theta}_K - \theta^*)_{S^c}\right\|_1 + \left\|(\tilde{\theta}_K - \theta^*)_S\right\|_1$, we can see that:

$$-c\lambda_{\text{ps}}\|\tilde{\theta}_K - \theta^*\|_1 + \lambda_{\text{ps}}\left\|(\tilde{\theta}_K - \theta^*)_S\right\|_1 \leq \lambda_{\text{ps}}\left(\|\theta^*\|_1 - \left\|\theta^* + (\tilde{\theta}_K - \theta^*)_{S^c}\right\|_1\right) \leq \lambda_{\text{ps}}\left\|\left(\tilde{\theta}_K - \theta^*\right)_S\right\|_1.$$

That is,

$$\left\|\left(\tilde{\theta}_K - \theta^*\right)_{S^c}\right\|_1 \leq \frac{1+c}{1-c}\left\|\left(\tilde{\theta}_K - \theta^*\right)_S\right\|_1.$$

Thus, under $\bigcap_{i=0}^{6}\mathcal{E}_i$,

$$\frac{3}{Kn}\sum_{(i,j)\in K^c}\left\{X_{ki}^T(\tilde{\theta}_K - \theta^*)\right\}^2 \leq C_L\left(\frac{s_1\log(p\vee Kn)}{Kn} + \frac{s_1^3\log^4(p\vee n)}{n^2}\right).$$

Since $\lambda_{\text{ps}} \geq \left\|\nabla\tilde{Q}(\theta^*)\right\|_\infty$, $c \leq 1$. Thus, by taking $\lambda_{\text{ps}} = 2\left\|\nabla\tilde{Q}(\theta^*)\right\|_\infty$, we can apply lemma 6 of Bradic et al. (2019), and the desired result is obtained. ■

Notice that the same result should still hold if we interchange J with K^c. Thus, this result is equivalent to proposition 10.

Proposition 11 *Under assumption 1-6, with $\lambda_{om} \asymp \sqrt{\frac{\log(p\vee Kn)}{Kn}}$, for DisC^{2}o-HD estimator $\tilde{\beta}_K$, we have:*

$$\left\|\tilde{\beta}_K-\beta^*\right\|_2 \le C_L\left(\sqrt{\frac{s_2\log(p\vee Kn)}{Kn}}\right),$$
$$\frac{3}{Kn}\sum_{(i,j)\in K}\left\{X_{ki}^T(\tilde{\beta}_K-\beta^*)\right\}^2 \le C_L\left(\frac{s_2\log(p\vee Kn)}{Kn}\right),$$
$$\frac{3}{Kn}\sum_{(i,j)\in K^c}\left\{X_{ki}^T(\tilde{\beta}_K-\beta^*)\right\}^2 \le C_L\left(\frac{s_2\log(p\vee Kn)}{Kn}\right).$$

holds with probability at least $1-\frac{M}{(p\vee n)^8}-\frac{M}{n^8}$, where $K=K_1$, K_2, or K_3 and K^c represents another set of servers.

Proof **First Claim:**

Consider the events defined as:

$$\mathcal{E}_1=\left\{\left\|\nabla^2L_N(\beta^*,\tilde{\theta}_{K^c})-\nabla^2L_1(\beta^*,\tilde{\theta}_{K^c})\right\|_\infty \le C_1\sqrt{\frac{\log(p\vee n)}{n}}\right\},$$
$$\mathcal{E}_2=\left\{\left\|\nabla L_N(\beta^*,\tilde{\theta}_{K^c})\right\|_\infty \le C_1\sqrt{\frac{\log(p\vee Kn)}{Kn}}\right\}.$$

For $\mathcal{E}_1,\mathcal{E}_2$, we can apply lemma 21 and union bound, where we can consider $\tilde{\theta}_{K^c}$ to be fixed. Combining the results above, we can see that $\mathbb{P}\left(\bigcap_{i=1}^2\mathcal{E}_i\right)\ge 1-\frac{M}{(p\vee n)^8}$.

We can see that for the loss function of our DisC^{2}o-HD estimator, we have,

$$\begin{aligned}
&\tilde{L}(\beta^*+\delta,\tilde{\theta}_{K^c})-\tilde{L}(\beta^*,\tilde{\theta}_{K^c})-\nabla\tilde{L}(\beta^*,\tilde{\theta}_{K^c})^T\delta\\
&=L_1(\beta^*+\delta,\tilde{\theta}_{K^c})-L_1(\beta^*,\tilde{\theta}_{K^c})-\nabla L_1(\beta^*,\tilde{\theta}_{K^c})^T\delta+\frac{1}{2}\delta^T\left(\nabla^2L_N(\bar{\beta}_K,\tilde{\theta}_{K^c})-\nabla^2L_1(\bar{\beta}_K,\tilde{\theta}_{K^c})\right)\delta.\\
&=\frac{1}{2}\delta^T\nabla^2L_1(\bar{\beta}_K,\tilde{\theta}_{K^c})\delta+\frac{1}{2}\delta^T\left(\nabla^2L_N(\bar{\beta}_K,\tilde{\theta}_{K^c})-\nabla^2L_1(\bar{\beta}_K,\tilde{\theta}_{K^c})\right)\delta\\
&=\frac{1}{2}\delta^T\nabla^2L_N(\bar{\beta}_K,\tilde{\theta}_{K^c})\delta\\
&\ge C\|\delta\|_2^2,
\end{aligned}$$

where the last inequality is a direct result of lemma 18 and algebra. Thus, $\tilde{L}$ also satisfies the RSC condition given in Negahban et al. (2012). Then, by directly applying Corollary 1 of Negahban et al. (2012), we have:

$$\left\|\tilde{\beta}_K-\beta^*\right\|_2 \le \frac{3\sqrt{s_2}\lambda_{\text{om}}}{C},$$

for every $\lambda_{\text{om}} \ge \left\|\nabla\tilde{L}(\beta^*,\tilde{\theta}_{K^c})\right\|_\infty$.

Following the same technique as proposition 10, we can show that

$$\begin{aligned}
&\nabla\tilde{L}(\beta^*,\tilde{\theta}_{K^c})\\
&=\nabla L_1(\beta^*,\tilde{\theta}_{K^c})+(\nabla L_N(\bar{\beta}_K,\tilde{\theta}_{K^c})-\nabla L_1(\bar{\beta}_K,\tilde{\theta}_{K^c}))+\left(\nabla^2 L_N(\bar{\beta}_K,\tilde{\theta}_{K^c})-\nabla^2 L_1(\bar{\beta}_K,\tilde{\theta}_{K^c})\right)(\beta^*-\bar{\beta}_K)\\
&=\nabla L_1(\beta^*,\tilde{\theta}_{K^c})-\nabla L_1(\bar{\beta}_K,\tilde{\theta}_{K^c})+\nabla L_N(\bar{\beta}_K,\tilde{\theta}_{K^c})-\nabla L_N(\beta^*,\tilde{\theta}_{K^c})+\nabla L_N(\beta^*,\tilde{\theta}_{K^c})\\
&+\left(\nabla^2 L_N(\bar{\beta}_K,\tilde{\theta}_{K^c})-\nabla^2 L_1(\bar{\beta}_K,\tilde{\theta}_{K^c})\right)(\beta^*-\bar{\beta}_K)\\
&=\left(\nabla^2 L_N(\beta^*,\tilde{\theta}_{K^c})-\nabla^2 L_N(\bar{\beta}_K,\tilde{\theta}_{K^c})\right)(\bar{\beta}_K-\beta^*)-\left(\nabla^2 L_1(\beta^*,\tilde{\theta}_{K^c})-\nabla^2 L_1(\bar{\beta}_K,\tilde{\theta}_{K^c})\right)(\bar{\beta}_K-\beta^*)\\
&+\nabla L_N(\beta^*,\tilde{\theta}_{K^c})\\
&=\nabla L_N(\beta^*,\tilde{\theta}_{K^c}).
\end{aligned}$$

Then, we can see that under $\bigcap_{i=1}^2 \mathcal{E}_i$

$$\left\|\nabla\tilde{L}(\beta^*,\tilde{\theta}_{K^c})\right\|_\infty\le\left\|\nabla L_N(\beta^*,\tilde{\theta}_{K^c})\right\|_\infty.$$

Under event $\bigcap_{i=1}^2 \mathcal{E}_i$, we can see that:

$$\left\|\nabla\tilde{L}(\beta^*,\tilde{\theta}_{K^c})\right\|_\infty\le C_1\sqrt{\frac{\log(p\vee Kn)}{Kn}}.$$

Then, with properly chosen λ_{om}, under $\bigcap_{i=1}^2 \mathcal{E}_i$, we can see that:

$$\left\|\tilde{\beta}_K-\beta^*\right\|_2\le C_L\sqrt{\frac{s_2\log(p\vee Kn)}{Kn}}.$$

Second and Third Claim:
The proof is an analog of proposition 10. ■

Notice that the same result still holds if we interchange J with K^c. Thus, this result is equivalent to lemma 11.

Theorem 12 *Under assumption 1-6, the distributed estimator for DisC^{2}o-HD method satisfies*

$$|\tilde{\tau}_1-\hat{\tau}_1^*|\le C_L\left(\frac{\sqrt{s_2(s_1\vee s_2)}\log(p\vee Kn)}{Kn}+\frac{\sqrt{s_2(s_1\vee s_2)\log(p\vee Kn)\log^4(p\vee n)}}{n\sqrt{Kn}}\right)$$

with probability at least $1-\frac{M}{(p\vee n)^8}-\frac{M}{n^8}$, where C_L is a sufficiently large constant and M is another constant depending on C_L.

Proof
We focus on $\left|\tilde{\tau}_{1,K_1}-\hat{\tau}^*_{1,K_1}\right|$ first. While $\left|\tilde{\tau}_{1,K_2}-\hat{\tau}^*_{1,K_2}\right|$ and $\left|\tilde{\tau}_{1,K_3}-\hat{\tau}^*_{1,K_3}\right|$ can be dealt with in a similar manner.

Consider the events defined as:

$$
\begin{aligned}
\mathcal{E}_0 &= \left\{ \|\tilde{\theta}_{K^c} - \theta^*\|_2 \le C_L \left(\sqrt{\frac{s_1 \log(p \vee Kn)}{Kn}} + \frac{s_1^{3/2} \log^2(p \vee n)}{n} \right) \right\}, \\
\mathcal{E}_1 &= \left\{ \|\tilde{\beta}_{K^c} - \beta^*\|_2 \le C_L \sqrt{\frac{s_2 \log(p \vee Kn)}{Kn}} \right\}, \\
\mathcal{E}_2 &= \left\{ \frac{3}{Kn} \sum_{(k,i)\in K} \left\{ X_{ki}^T(\tilde{\theta}_{K^c} - \theta^*) \right\}^2 \le C_L \left(\frac{s_1 \log(p \vee Kn)}{Kn} + \frac{s_1^3 \log^4(p \vee n)}{n^2} \right), \right. \\
&\qquad \left. \frac{3}{Kn} \sum_{(k,i)\in K} \left\{ X_{ki}^T(\tilde{\beta}_{K^c} - \beta^*) \right\}^2 \le C_L \frac{s_2 \log(p \vee Kn)}{Kn} \right\}, \\
\mathcal{E}_3 &= \left\{ \left\| \frac{3}{Kn} \sum_{(k,i)\in K} \left\{ \left(\frac{T_{ki}}{\pi(X_{ki}^T\theta^*)} - 1 \right) X_{ki}^T \right\} \right\|_\infty \le C_L \sqrt{\frac{\log(p \vee Kn)}{Kn}} \right\}, \\
\mathcal{E}_4 &= \left\{ \left\| \frac{3}{Kn} \sum_{(k,i)\in K} X_{ki}\varepsilon_{ki}^* \right\|_\infty \le C_L \sqrt{\frac{\log(p \vee Kn)}{Kn}} \right\}.
\end{aligned}
$$

By proposition 10 and 11 we may realize that $\mathcal{E}_0, \mathcal{E}_1, \mathcal{E}_2$ will hold with probability at least $1 - \frac{M}{(p\vee n)^8} - \frac{M}{n^8}$ for both one-step and DisC^{2}o-HD estimator. We can see that $\frac{T_{ki}}{\pi(X_{ki}^T\theta^*)} - 1$ is bounded by the strong ignorability assumption, with zero expected value. By lemma 21 and union bound, we can see event $\mathcal{E}_3, \mathcal{E}_4$ will hold with probability at least $1 - \frac{M}{(p\vee Kn)^8}$. Thus, $\bigcap_{i=0}^4 \mathcal{E}_i$ will hold with probability at least $1 - \frac{M}{(p\vee n)^8} - \frac{M}{n^8}$.

By rearranging the terms, we have:

$$
\begin{aligned}
\hat{\tau}_{1,J}^* - \tilde{\tau}_{1,J} &= \frac{3}{Kn} \sum_{(k,i)\in K} \left\{ X_{ki}^T\beta^* + \frac{T_{ki}}{\pi(X_{ki}^T\theta^*)} \left(Y_{ki} - X_{ki}^T\beta^* \right) \right\} \\
&\quad - \frac{3}{Kn} \sum_{(k,i)\in K} \left\{ \left(X_{ki}^T\tilde{\beta}_{K^c} + \frac{T_{ki}}{\pi(X_{ki}^T\tilde{\theta}_{K^c})} \left(Y_{ki} - X_{ki}^T\tilde{\beta}_{K^c} \right) \right) \right\} \\
&= \Delta_1 + \Delta_2 + \Delta_3.
\end{aligned}
$$

where

$$
\begin{aligned}
\Delta_1 &= \frac{3}{Kn} \sum_{(k,i)\in K} \left(\frac{T_{ki}}{\pi\left(X_{ki}^T\tilde{\theta}_{K^c}\right)} - \frac{T_{ki}}{\pi\left(X_{ki}^T\theta^*\right)} \right) X_{ki}^T \left(\tilde{\beta}_{K^c} - \beta^* \right), \\
\Delta_2 &= \frac{3}{Kn} \sum_{(k,i)\in K} \left(\frac{T_{ki}}{\pi\left(X_{ki}^T\theta^*\right)} - 1 \right) X_{ki}^T \left(\tilde{\beta}_{K^c} - \beta^* \right), \\
\Delta_3 &= \frac{3}{Kn} \sum_{(k,i)\in K} \left(\frac{T_{ki}}{\pi(X_{ki}^T\theta^*)} - \frac{T_{ki}}{\pi(X_{ki}^T\tilde{\theta}_{K^c})} \right) \left(Y_{ki} - X_{ki}^T\beta^* \right).
\end{aligned}
$$

Consider Δ_1, we may see that

$$|\Delta_1| = \left| \frac{3}{Kn} \sum_{(k,i)\in K} T_{ki} \left\{ \left(\exp\left(-X_{ki}^T \tilde{\theta}_{K^c}\right) - \exp\left(-X_{ki}^T \theta^*\right) \right) X_{ki}^T \left(\tilde{\beta}_{K^c} - \beta^* \right) \right\} \right|$$
$$\leq C' \left(\frac{3}{Kn} \sum_{(k,i)\in K} \left\{ X_{ki}^T \left(\tilde{\beta}_{K^c} - \beta^* \right) \right\}^2 \right)^{1/2} \left(\frac{3}{Kn} \sum_{(k,i)\in K} \left\{ X_{ki}^T \left(\tilde{\theta}_{K^c} - \theta^* \right) \right\}^2 \right)^{1/2},$$

where we applied the mean value theorem and Cauchy inequality in the second line. Under $\mathcal{E}_0 \bigcap \mathcal{E}_1$, we have :

$$|\Delta_1| \leq C \left(\frac{\sqrt{s_1 s_2} \log(p \vee Kn)}{Kn} + \frac{s_1 \sqrt{s_1 s_2 \log(p \vee Kn) \log^4(p \vee n)}}{n\sqrt{Kn}} \right).$$

While for Δ_2, we have:

$$\Delta_2 = \frac{3}{Kn} \sum_{(k,i)\in K} \left\{ \left(\frac{T_{ki}}{\pi(X_{ki}^T \theta^*)} - 1 \right) X_{ki}^T \left(\tilde{\beta}_{K^c} - \beta^* \right) \right\}$$
$$\leq \left\| \frac{3}{Kn} \sum_{(k,i)\in K} \left\{ \left(\frac{T_{ki}}{\pi(X_{ki}^T \theta^*)} - 1 \right) X_{ki}^T \right\} \right\|_\infty \left\| \tilde{\beta}_{K^c} - \beta^* \right\|_1.$$
$$\leq \left\| \frac{3}{Kn} \sum_{(k,i)\in K} \left\{ \left(\frac{T_{ki}}{\pi(X_{ki}^T \theta^*)} - 1 \right) X_{ki}^T \right\} \right\|_\infty \sqrt{s_2} \left\| \tilde{\beta}_{K^c} - \beta^* \right\|_2$$

Thus, under $\mathcal{E}_1 \bigcap \mathcal{E}_3$,

$$|\Delta_2| \leq C_L \frac{s_2 \log(p \vee Kn)}{Kn}.$$

While for Δ_3, we can take advantage of the sample splitting method, which would give us:

$$\mathbb{E} \left[\frac{3}{Kn} \sum_{(k,i)\in K} \left\{ \left(\frac{T_{ki}}{\pi(X_{ki}^T \tilde{\theta}_{K^c})} - \frac{T_{ki}}{\pi(X_{ki}^T \theta^*)} \right) \right\} \varepsilon_{ki}^* \right]$$
$$= \mathbb{E} \left[\mathbb{E} \left[\frac{3}{Kn} \sum_{(k,i)\in K} \left(\frac{T_{ki}}{\pi(X_{ki}^T \tilde{\theta}_{K^c})} - \frac{T_{ki}}{\pi(X_{ki}^T \theta^*)} \right) \varepsilon_{ki}^* \middle| \{X_{ki}, Y_{ki}, T_{ki}\}_{(i,j)\in K^c} \bigcap \{X_{ki}\}_{(k,i)\in K} \right] \right] = 0$$

Notice that if we condition on $\{X_{ki}, Y_{ki}, T_{ki}\}_{i\in J_K^c} \bigcap \{X_{ki}\}_K$, $\pi(X_{ki}^T \tilde{\theta}_{K^c})$ can be considered as fixed.

Then, we can consider the truncated case, where $\mathcal{E}_5 = \left\{ \frac{3}{Kn} \sum_{(k,i)\in K} \left(\frac{T_{ki}}{\pi(X_{ki}^T \tilde{\theta}_{K^c})} - \frac{T_{ki}}{\pi(X_{ki}^T \theta^*)} \right) \varepsilon_{ki}^* \geq t \right\}$ for some t to be chosen later. In the meanwhile, we can denote:

$$A = \left\{ \frac{3}{Kn} \sum_{(k,i)\in K} \left\{ X_{ki}^T \left(\tilde{\theta}_{K^c} - \theta^* \right) \right\}^2 \leq C_L \left(\frac{s_1 \log(p \vee Kn)}{Kn} + \frac{s_1^3 \log^2(p \vee n)}{n^2} \right) \right\}.$$

Then, we can see that:

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}_5\right) &= \mathbb{E}\left[\mathbb{P}\left(\mathcal{E}_5\middle|\{T_{ki},X_{ki},Y_{ki}\}_{K^c}\bigcap\{X_{ki}\}_K\right)\right]\\ &= \mathbb{E}\left[\mathbb{P}\left(\mathcal{E}_5\middle|\{T_{ki},X_{ki},Y_{ki}\}_{K^c}\bigcap\{X_{ki}\}_K\right)\mathbb{1}\left\{A\right\}\right] + \mathbb{E}\left[\mathbb{P}\left(\mathcal{E}_5\middle|\{T_{ki},X_{ki},Y_{ki}\}_{K^c}\bigcap\{X_{ki}\}_K\right)\mathbb{1}\left\{A^c\right\}\right].\end{aligned}$$

Denote $\Delta_{Kn}^2 = \frac{3}{Kn}\sum_{(k,i)\in K}\left\{\left(\exp(-X_{ki}^T\theta^*) - \exp(-X_{ki}^T\tilde{\theta}_{K^c})\right)\right\}^2$. Since $T_{ki}\varepsilon_{ki}^*$ are sub-Gaussian, by assumption 10 and the fact that $T_{ki}\in\{0,1\}$, by Hoeffding inequality, we have

$$\begin{aligned}&\mathbb{E}\left(\exp\left(-\frac{CKnt^2}{\Delta_{Kn}^2}\right)\right)\\ &\leq \mathbb{E}\left(\exp\left(-\frac{CKnt^2}{\Delta_{Kn}^2}\right)\mathbb{1}\left\{\frac{3}{Kn}\sum_{(k,i)\in K}\left\{X_{ki}^T\left(\tilde{\theta}_{K^c}-\theta^*\right)\right\}^2 \leq C_L\left(\frac{s_1\log(p\vee Kn)}{Kn}+\frac{s_1^3\log^4(p\vee n)}{n^2}\right)\right\}\right)\\ &+\mathbb{P}\left(A^c\right)\\ &\leq \mathbb{E}\left(\exp\left(-\frac{CKnt^2}{C_L\left(\frac{s_1\log(p\vee Kn)}{Kn}+\frac{s_1^3\log^4(p\vee n)}{n^2}\right)}\right)\right)+\mathbb{P}\left(A^c\right).\end{aligned}$$

Taking $t^2 = C_L'\frac{\log(p\vee Kn)}{Kn}\left(\frac{s_1\log(p\vee Kn)}{Kn}+\frac{s_1^3\log^4(p\vee n)}{n^2}\right)$:

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}_5\right) &= \mathbb{P}\left(|\Delta_3| \geq C_L\sqrt{\frac{\log(p\vee Kn)}{Kn}}\left(\sqrt{\frac{s_1\log(p\vee Kn)}{Kn}}+\frac{s_1^{3/2}\log^2(p\vee n)}{n}\right)\right)\\ &\leq \frac{M}{(p\vee Kn)^8}.\end{aligned}$$

Thus, we can see that event

$$|\Delta_3| \leq \sqrt{\frac{\log(p\vee Kn)}{Kn}}\left(\sqrt{\frac{s_1\log(p\vee Kn)}{Kn}}+\frac{s_1^{3/2}\log^2(p\vee n)}{n}\right)$$

will hold with probability at least $1-\frac{M}{(p\vee Kn)^8}$. Combining the event and probability we obtained at Δ_1, Δ_2, and Δ_3, we have

$$\begin{aligned}&\mathbb{P}\left(|\tilde{\tau}_{1,J}-\hat{\tau}_{1,J}^*| \leq C_L\left(\frac{\sqrt{s_2(s_1\vee s_2)}\log(p\vee Kn)}{Kn}+\frac{s_1\sqrt{s_1s_2\log(p\vee Kn)\log^4(p\vee n)}}{n\sqrt{Kn}}\right)\right)\\ &\geq 1-\frac{M}{(p\vee n)^8}-\frac{M}{n^8}.\end{aligned}$$

Likewise, we can apply the same technique to $\tilde{\tau}_{1,K_2}$ or $\tilde{\tau}_{1,K_3}$. Combining the bound above, the desired bound can be obtained. ■

Theorem 13 *Under assumptions 1-6, we have*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\sqrt{Kn}(\tilde{\tau}_1-\tau_1^*)}{\sqrt{\widehat{V}}}\le x\right)-\Phi(x)\right|$$
$$\le \frac{M}{(p\vee n)^8}+\frac{M}{n^8}+C_L\left(\frac{\sqrt{s_2(s_1\vee s_2)}\log(p\vee Kn)}{\sqrt{Kn}}+\frac{s_1\sqrt{s_1s_2\log(p\vee Kn)\log^4(p\vee n)}}{n}\right),$$

where C_L is a sufficiently large constant, and M depends on C_L.

Proof We prove this theorem in two steps. Firstly, we control

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\sqrt{Kn}(\tilde{\tau}_1-\tau_1^*)}{\sqrt{V^*}}\le x\right)-\Phi(x)\right|.$$

Then, with the Berry-Esseen type bound established, we can replace the true value of the variance with the variance estimator to bound

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\sqrt{Kn}(\tilde{\tau}_1-\tau_1^*)}{\sqrt{\widehat{V}}}\le x\right)-\Phi(x)\right|.$$

For the first term, consider $\hat{\tau}_1^*$ defined as:

$$\hat{\tau}_1^* := \frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{X_{ki}^T\beta^*+\frac{T_{ki}}{\pi(X_{ki}^T\theta^*)}\left(Y_{ki}-X_{ki}^T\beta^*\right)\right\},$$

then we can see that the classical Berry-Esseen bound holds under assumption 12:

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\sqrt{Kn}(\hat{\tau}_1^*-\tau_1^*)}{\sqrt{V^*}}\le x\right)-\Phi(x)\right|\le\frac{C}{\sqrt{Kn}}.$$

Then, we may consider the event $\mathcal{E}_1=\{|\tilde{\tau}_1-\hat{\tau}_1^*|\le r\}$ where

$$r=C_L\left(\frac{\sqrt{s_2(s_1\vee s_2)}\log(p\vee Kn)}{Kn}+\frac{s_1\sqrt{s_1s_2\log(p\vee Kn)\log^4(p\vee n)}}{n\sqrt{Kn}}\right).$$

By theorem 12, we can realize that $\mathbb{P}(\mathcal{E}_1^c)\le\frac{M}{(p\vee n)^8}+\frac{M}{n^8}$.

For the first term, we may realize that:

$$\mathbb{P}\left(\frac{\sqrt{Kn}(\tilde{\tau}_1-\tau_1^*)}{\sqrt{V^*}}\le x\right)-\Phi(x)=\mathbb{P}\left(\frac{\sqrt{Kn}(\tilde{\tau}_1-\tau_1^*)}{\sqrt{V^*}}\le x,\mathcal{E}_1\right)-\Phi(x)+\mathbb{P}\left(\frac{\sqrt{Kn}(\tilde{\tau}_1-\tau_1^*)}{\sqrt{V^*}}\le x,\mathcal{E}_1^c\right).$$

It can be easily showed that:

$$\sup_{x\in\mathbb{R}}\mathbb{P}\left(\frac{\sqrt{Kn}(\tilde{\tau}_1-\tau_1^*)}{\sqrt{V^*}}\le x,\mathcal{E}_1\right)-\Phi\left(x\right)\le\frac{C}{\sqrt{Kn}}+\frac{M}{(p\vee n)^8}+\frac{M}{n^8}+C\sqrt{Kn}r,$$

where C is some positive constant. With similar argument, we can see that:

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\sqrt{Kn}(\tilde{\tau}_1-\tau_1^*)}{\sqrt{V^*}}\leq x\right)-\Phi\left(x\right)\right|\leq\frac{C}{\sqrt{Kn}}+\frac{M}{(p\vee n)^8}+\frac{M}{n^8}+C'\sqrt{Kn}r,$$

where C' is some positive constant.

By proposition 14,

$$\mathbb{P}\left(\left|\sqrt{\frac{\widehat{V}}{V^*}}-1\right|\leq C_L\left(\sqrt{\frac{(s_1\vee s_2)\log(p\vee Kn)}{Kn}}+\frac{s_1^{3/2}\log(p\vee n)}{n}\right)\right)\geq 1-\frac{M}{(p\vee n)^8}-\frac{M}{n^8}.$$

Thus, for the second result, we can apply the same technique as above. The same result shall be obtained. ■

A.2 Consistency Lemmas

Proposition 14 (Consistency of variance estimator) *The variance estimator satisfies*

$$|\widehat{V}-V^*|\leq C_L\left(\sqrt{\frac{(s_1\vee s_2)\log(p\vee Kn)}{Kn}}+\frac{s_1^{3/2}\log^2(p\vee n)}{n}\right),$$
$$\left|\sqrt{\frac{\widehat{V}}{V^*}}-1\right|\leq C_L\left(\sqrt{\frac{(s_1\vee s_2)\log(p\vee Kn)}{Kn}}+\frac{s_1^{3/2}\log^2(p\vee n)}{n}\right),$$

with probability at least $1-\frac{M}{(p\vee n)^8}-\frac{M}{n^8}$, where C_L is a sufficiently large constant and M depends on C_L.

Proof

Consider events defined as

$$\mathcal{E}_0 = \left\{ \|\tilde{\theta} - \theta^*\|_2 \leq C_L \left(\sqrt{\frac{s_1 \log(p \vee Kn)}{Kn}} + \frac{s_1^{3/2} \log^2(p \vee n)}{n} \right) \right\},$$

$$\mathcal{E}_1 = \left\{ \|\tilde{\beta} - \beta^*\|_2 \leq C_L \sqrt{\frac{s_2 \log(p \vee Kn)}{Kn}} \right\}$$

$$\mathcal{E}_2 = \left\{ \frac{1}{Kn} \sum_{j=1}^{m} \sum_{i=1}^{n} (X_{ki}^T (\tilde{\theta} - \theta^*))^2 \leq C_L \left(\frac{s_1 \log(p \vee Kn)}{Kn} + \frac{s_1^3 \log^4(p \vee n)}{n^2} \right), \right.$$

$$\left. \frac{1}{Kn} \sum_{j=1}^{m} \sum_{i=1}^{n} (X_{ki}^T (\tilde{\beta} - \beta^*))^2 \leq C_L \frac{s_2 \log(p \vee Kn)}{Kn} \right\},$$

$$\mathcal{E}_3 = \left\{ \left\| \frac{1}{Kn} \sum_{j=1}^{m} \sum_{i=1}^{n} \left\{ \left(\frac{T_{ki}}{\pi(X_{ki}^T \theta^*)} - 1 \right) X_{ki}^T \right\} \right\|_\infty \leq C_L \sqrt{\frac{\log(p \vee Kn)}{Kn}} \right\},$$

$$\mathcal{E}_4 = \left\{ |\tau_1^* - \tilde{\tau}_1| \leq \frac{C_L'}{\sqrt{Kn}} \right\},$$

$$\mathcal{E}_5 = \left\{ \left| \frac{1}{Kn} \sum_{j=1}^{m} \sum_{i=1}^{n} \left\{ \frac{T_{ki}}{\pi(X_{ki}^T \theta^*)^2} \left(Y_{ki} - X_{ki}^T \beta^* \right)^2 \right\} - \mathbb{E} \left[\frac{1}{\pi(X_{ki}^T \theta^*)} (Y_{ki} - X_{ki}^T \beta^*)^2 \right] \right| \leq C_L \sqrt{\frac{\log(p \vee Kn)}{Kn}} \right\},$$

$$\mathcal{E}_6 = \left\{ \left| \frac{1}{Kn} \sum_{j=1}^{m} \sum_{i=1}^{n} (\varepsilon_{ki}^*)^4 - \mathbb{E}\left((\varepsilon_{ki}^*)^4 \right) \right| \leq C_L, \left\| \frac{1}{Kn} \sum_{j=1}^{m} \sum_{i=1}^{n} \varepsilon_{ki} X_{ki} \right\|_\infty \leq C_L \sqrt{\frac{\log(p \vee Kn)}{Kn}} \right\},$$

$$\mathcal{E}_7 = \left\{ \left| \frac{1}{Kn} \sum_{j=1}^{m} \sum_{i=1}^{n} \left\{ \left(X_{ki}^T \beta^* - \tau_1^* \right)^2 \right\} - \mathbb{E}\left(\left(X_{ki}^T \beta^* - \tau_1^* \right)^2 \right) \right| \leq C_L \sqrt{\frac{\log(p \vee Kn)}{Kn}} \right\}.$$

By proposition 10 and 11 we may realize that $\mathcal{E}_0, \mathcal{E}_1$ will hold with probability at least $1 - \frac{M}{(p \vee n)^8} - \frac{M}{n^8}$. Also, $\mathcal{E}_2$ will hold with probability at least $1 - \frac{M}{(p \vee Kn)^8}$, which is the result of lemma 25. For event $\mathcal{E}_3$, it is a combination of theorem 12, assumption 11, union bound, and lemma 21. We can see that $\frac{T_{ki}}{\pi(X_{ki}^T \theta^*)} - 1$ is bounded by the strong ignorability assumption, with zero expected value. By lemma 21 and union bound, we can see events $\mathcal{E}_4, \mathcal{E}_6, \mathcal{E}_7$ will hold with probability at least $1 - \frac{M}{(p \vee Kn)^8}$. We can show the second part of $\mathcal{E}_6$ by the same technique. For the first part of $\mathcal{E}_6$, it can be derived by lemma 24. Thus, $\bigcap_{i=0}^7 \mathcal{E}_i$ will hold with probability at least $1 - \frac{M}{(p \vee n)^8} - \frac{M}{n^8}$.

We can show that, under $\bigcap_{i=0}^7 \mathcal{E}_i$:

$$|\widehat{V} - \widehat{V}^*| \leq C_L \left(\sqrt{\frac{(s_1 \vee s_2) \log(p \vee Kn)}{Kn}} + \frac{s_1^{3/2} \log^2(p \vee n)}{n} \right),$$

where $\widehat{V}^* := \frac{1}{Kn} \sum_{j=1}^{m} \sum_{i=1}^{n} \left\{ \frac{T_{ki}}{\pi(X_{ki}^T \theta^*)^2} \left(Y_{ki} - X_{ki}^T \beta^* \right)^2 + \left(X_{ki}^T \beta^* - \tau_1^* \right)^2 \right\}$.

Then, we show that:

$$|\widehat{V}^* - V^*| \le C_L' \sqrt{\frac{\log(p \vee Kn)}{Kn}}.$$

We may begin by rearranging the terms. Consider $\widehat{V} - \widehat{V}^*$, we have:

$$\begin{aligned}\widehat{V} - \widehat{V}^* &= \frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{\frac{T_{ki}}{\pi(X_{ki}^T\tilde{\theta})^2}\left(Y_{ki} - X_{ki}^T\tilde{\beta}\right)^2 + \left(X_{ki}^T\tilde{\beta} - \tilde{\tau}_1\right)^2\right\} \\ &\quad - \frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{\frac{T_{ki}}{\pi(X_{ki}\theta^*)^2}\left(Y_{ki} - X_{ki}^T\beta^*\right)^2 + \left(X_{ki}^T\beta^* - \tau_1^*\right)^2\right\}.\end{aligned}$$

This can be decomposed as:

$$\widehat{V} - \widehat{V}^* = \Delta_1 + \Delta_2,$$

where

$$\begin{aligned}\Delta_1 &= \frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{\frac{T_{ki}}{\pi(X_{ki}^T\tilde{\theta})^2}\left(Y_{ki} - X_{ki}^T\tilde{\beta}\right)^2\right\} - \frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{\frac{T_{ki}}{\pi(X_{ki}\theta^*)^2}\left(Y_{ki} - X_{ki}^T\beta^*\right)^2\right\}, \\ \Delta_2 &= \frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{\left(X_{ki}^T\tilde{\beta} - \tilde{\tau}_1\right)^2\right\} - \frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{\left(X_{ki}^T\beta^* - \tau_1^*\right)^2\right\}.\end{aligned}$$

Then, we can decompose Δ_1:

$$\Delta_1 = \Delta_{11} + \Delta_{12},$$

where

$$\begin{aligned}\Delta_{11} &= \frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{\frac{T_{ki}}{\pi(X_{ki}\theta^*)^2}\left(\left(Y_{ki} - X_{ki}^T\tilde{\beta}\right)^2 - \left(Y_{ki} - X_{ki}^T\beta^*\right)^2\right)\right\}, \\ \Delta_{12} &= \frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{\left(\frac{T_{ki}}{\pi(X_{ki}^T\tilde{\theta})^2} - \frac{T_{ki}}{\pi(X_{ki}^T\theta^*)^2}\right)\left(Y_{ki} - X_{ki}^T\tilde{\beta}\right)^2\right\}.\end{aligned}$$

Then, for Δ_{11}, we can see that

$$\begin{aligned}\Delta_{11} &= \frac{2}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{\frac{T_{ki}}{\pi(X_{ki}\theta^*)^2}\varepsilon_{ki}^*\left(X_{ki}^T(\tilde{\beta} - \beta^*)\right)\right\} + \frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{\frac{T_{ki}}{\pi(X_{ki}\theta^*)^2}\left(X_{ki}^T(\tilde{\beta} - \beta^*)\right)^2\right\} \\ &\le \left\|\frac{2}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{T_{ki}}{\pi(X_{ki}^T\theta^*)^2}\varepsilon_{ki}^* X_{ki}^T\right\|_\infty \left\|\tilde{\beta} - \beta^*\right\|_1 + \frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{X_{ki}^T(\tilde{\beta} - \beta^*)\right\}^2. \\ &\le \left\|\frac{2}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{T_{ki}}{\pi(X_{ki}^T\theta^*)^2}\varepsilon_{ki}^* X_{ki}^T\right\|_\infty \sqrt{s_2}\left\|\tilde{\beta} - \beta^*\right\|_2 + \frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{X_{ki}^T(\tilde{\beta} - \beta^*)\right\}^2.\end{aligned}$$

Under $\mathcal{E}_1 \bigcap \mathcal{E}_2 \bigcap \mathcal{E}_6$:

$$\Delta_{11} \le C_L \frac{s_2 \log(p \vee Kn)}{Kn}.$$

For Δ_{12}, we have:

$$\begin{aligned}
\Delta_{12} = & \frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{\left(\frac{T_{ki}}{\pi(X_{ki}^T\tilde{\theta})^2} - \frac{T_{ki}}{\pi(X_{ki}^T\theta^*)^2}\right) 2\varepsilon_{ki}^* \left(X_{ki}^T\tilde{\beta} - X_{ki}^T\beta^*\right)\right\} \\
& + \frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{\left(\frac{T_{ki}}{\pi(X_{ki}^T\tilde{\theta})^2} - \frac{T_{ki}}{\pi(X_{ki}^T\theta^*)^2}\right) \left(X_{ki}^T\tilde{\beta} - X_{ki}^T\beta^*\right)^2\right\} \\
& + \frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{\left(\frac{T_{ki}}{\pi(X_{ki}^T\tilde{\theta})^2} - \frac{T_{ki}}{\pi(X_{ki}^T\theta^*)^2}\right) (\varepsilon_{ki}^*)^2\right\}.
\end{aligned}$$

We can apply similar technique as above for the first two terms. While for the last one, by Cauchy inequality and mean value theorem:

$$\begin{aligned}
& \frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{\left(\frac{T_{ki}}{\pi(X_{ki}^T\tilde{\theta})^2} - \frac{T_{ki}}{\pi(X_{ki}^T\theta^*)^2}\right) (\varepsilon_{ki}^*)^2\right\} \\
\le & C\left(\frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{\left(\frac{T_{ki}}{\pi(X_{ki}^T\tilde{\theta})} + \frac{T_{ki}}{\pi(X_{ki}^T\theta^*)}\right) (\varepsilon_{ki}^*)^2\right\}^2\right)^{1/2} \left(\frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{X_{ki}^T(\tilde{\theta} - \theta^*)\right\}^2\right)^{1/2}.
\end{aligned}$$

Under $\mathcal{E}_0 \bigcap \mathcal{E}_2 \bigcap \mathcal{E}_5$

$$|\Delta_{12}| \le C_L \left(\sqrt{\frac{(s_1 \vee s_2)\log(p \vee Kn)}{Kn}} + \frac{s_1^{3/2}\log^2(p \vee n)}{n}\right).$$

For Δ_2, we have:

$$\begin{aligned}
|\Delta_2| \le 2 & \left(\frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{X_{ki}^T(\tilde{\beta} - \beta^*)\right\}^2 + (\tilde{\tau}_1 - \tau_1^*)^2\right. \\
& + \left(\frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{X_{ki}^T(\tilde{\beta} - \beta^*)\right\}^2\right)^{1/2} \left(\frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{X_{ki}^T\beta^* - \tau_1^*\right\}^2\right)^{1/2} \\
& \left. + \left(\frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{X_{ki}^T\beta^* - \tau_1^*\right\}^2\right)^{1/2} |\tilde{\tau}_1 - \tau_1^*|\right).
\end{aligned}$$

Thus, under $\bigcap_{i=0}^{7}\mathcal{E}_i$

$$|\widehat{V} - \widehat{V}^*| \le C_L \left(\sqrt{\frac{(s_1 \vee s_2)\log(p \vee Kn)}{Kn}} + \frac{s_1^{3/2}\log^2(p \vee n)}{n}\right).$$

Then,

$$\left|\widehat{V}^* - V^*\right| \leq \left|\frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{\frac{T_{ki}}{\pi(X_{ki}^T\theta^*)^2}\left(Y_{ki} - X_{ki}^T\beta^*\right)^2\right\} - \mathbb{E}\left(\frac{1}{\pi(X_{ki}^T\theta^*)}(Y_{ki} - X_{ki}^T\beta^*)^2\right)\right|$$

$$+\left|\frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left\{\left(X_{ki}^T\beta^* - \tau_1^*\right)^2\right\} - \mathbb{E}\left(\left(X_{ki}^T\beta - \tau_1^*\right)^2\right)\right|.$$

Under $\mathcal{E}_6$:

$$\left|\widehat{V}^* - V^*\right| \leq C_L\sqrt{\frac{\log(p \vee Kn)}{Kn}}.$$

Under the events above, since V^* is bounded and positive:

$$\left|\sqrt{\frac{\widehat{V}}{V^*}} - 1\right| \leq \left|\frac{\widehat{V} - V^*}{V^* + \sqrt{\widehat{V}V^*}}\right| \leq C_L\left(\sqrt{\frac{(s_1 \vee s_2)\log(p \vee Kn)}{Kn}} + \frac{s_1^{3/2}\log^2(p \vee n)}{n}\right).$$

■

A.3 Misspecified models

Proposition 15 *Under Assumptions 1-6, with θ^* replaced by θ^o, the proposed estimator satisfies*

$$|\tilde{\tau}_1 - \hat{\tau}_{1,ps}^o| \leq C_L\left(\frac{\sqrt{s_2(s_1 \vee s_2)}\log(p \vee Kn)}{Kn} + \frac{s_1\sqrt{s_1 s_2\log(p \vee Kn)\log^4(p \vee n)}}{n\sqrt{Kn}}\right)$$

with probability at least $1 - \frac{M}{n^8}$, where C_L is a sufficiently large constant and M is another constant depending on C_L.

Proof The proof is an analog of theorem 13, where auxiliary lemmas can be established in a likewise manner. ■

Lemma 16 *Under Assumptions 1-6, with β^* replaced by β^o, the proposed estimator satisfies*

$$|\tilde{\tau}_1 - \hat{\tau}_{1,om}^o| \leq C_L\left(\frac{\sqrt{s_2(s_1 \vee s_2)}\log(p \vee Kn)}{Kn} + \frac{s_1\sqrt{s_1 s_2\log(p \vee Kn)\log^4(p \vee n)}}{n\sqrt{Kn}}\right)$$

with probability at least $1 - \frac{M}{n^8}$, where C_L is a sufficiently large constant and M is another constant depending on C_L.

Proof The proof is an analog of theorem 13, where auxiliary lemmas can be established in a likewise manner. ■

A.4 Initial estimator

Lemma 17 *For $\bar{\theta} = \bar{\theta}_K$, we have:*

$$\left\|\bar{\theta} - \theta^*\right\|_2 \leq C_1 \sqrt{\frac{s_1 \log(p \vee n)}{n}}$$

holds with probability at least $1 - \frac{M}{(p \vee n)^8}$, where $J = K_1$, K_2, or K_3.

Proof By lemma 19, we can see that event defined as:

$$\mathcal{E}_0 = \left\{ Q_1(\theta^* + \delta) - Q_1(\theta^*) - \nabla Q_1(\theta^*)^T \delta \geq \mu \left\|\delta\right\|_2^2 - \mu' \sqrt{\frac{\log p}{n}} \left\|\delta\right\|_2 \left\|\delta\right\|_1 \right\}.$$

will hold with probability at least $1 - \frac{M}{(p \vee n)^8}$.

Then, under $\mathcal{E}_0$, we can apply corollary 1 of Negahban et al. (2012). Then, it shall be obtained that:

$$\left\|\bar{\theta}_K - \theta^*\right\|_2 \leq 3 \frac{\sqrt{s_1} \lambda_{\text{ps, ini}}}{C},$$

where

$$\lambda_{\text{ps, ini}} \geq \left\|\nabla Q_1(\theta^*)\right\|_\infty.$$

For $\nabla Q_1(\theta^*)$, we have:

$$\nabla Q_1(\theta^*) = \frac{1}{n} \sum_{i=1}^n (1 - T_{i1}) X_{i1S}^T - T_{i1} \exp\left(-X_{i1}^T \theta^*\right) X_{i1S}^T.$$

Since $\mathbb{E}(\nabla Q_1(\theta^*)) = 0$ and $T_{i1}, (1 - T_{i1})$ are bounded, with X_{i1S} being sub-Gaussian, by union bound and lemma 21,

$$\left\|\nabla Q_1(\theta^*)\right\|_\infty \leq \sqrt{\frac{\log(p \vee n)}{n}}.$$

Thus, we have:

$$\left\|\bar{\theta} - \theta^*\right\|_2 \leq C_1 \sqrt{\frac{s_1 \log(p \vee n)}{n}}.$$

■

A.5 Restricted strong convexity(RSC) conditions

Lemma 18 *Under assumptions, we can see that the event defined as:*

$$\sum_{i=1}^n T_{ki} \geq c_1 n + C_L \sqrt{n \log n}$$

will hold with probability at least $1 - \frac{M}{(p \vee n)^8}$ for some constant c_1, C_L.

Proof Let Z_{ki} be binomial random variables with probability c_0, where c_0 is the constant defined in 8. Then, we clearly have

$$\mathbb{P}\left(\left\{\sum_{i=1}^{n} T_{ki} \geq c_1 n + C_L\sqrt{n\log n}\right\}\right) \geq \mathbb{P}\left(\left\{\sum_{i=1}^{n} Z_{ki} \geq c_1 n + C_L\sqrt{n\log(p\vee n)}\right\}\right).$$

Clearly, Z_{ki} are sub-Gaussian random variables. Then, by lemma 22, the claimed bound can be obtained. ■

Lemma 19 *The loss function of propensity score model follows restricted strong convexity with probability at least* $1-\frac{M}{(p\vee n)^8}$*, where* M *is some positive constant. That is: for all* δ *s.t.* $\|\delta\|_2 \leq 1$

$$Q_1(\theta^*+\delta) - Q_1(\theta^*) - \nabla Q_1(\theta^*)^T\delta \geq \mu\|\delta\|_2^2 - \mu'\frac{\log p}{n}\|\delta\|_1^2,$$

where μ, μ' *is some positive constant.*

Proof Under lemma 18, the claim is a result of Proposition 2 in Negahban et al. (2009). As demonstrated in Negahban et al. (2009), Assumption 3, outlined below, is integral to the proof of Lemma 10. Serving as a foundational assumption for high-dimensional data analysis, this condition ensures that the loss function maintains sufficient convexity even in high-dimensional spaces where traditional convexity might not universally apply.

Assumption 3 (Design) *The minimal and maximal eigenvalues of* $\mathbb{E}[X_{ki}X_{ki}^T]$ *are contained in a bounded interval that does not contain zero.*

■

A.6 Concentration results

Lemma 20 *Consider* $\sum_{i=1}^n X_i$*, where* X_i *are zero-mean, independent sub-exponential random variables with parameter* $\alpha = a_i, \nu = \nu_i$*. Then,* $Y = \sum_{i=1}^n X_i$ *is a sub-exponential random variable with parameter* $\alpha = \max_i a_i, \nu = \sqrt{\sum_{i=1}^n \nu_i^2}$*.*

Lemma 21 (Bernstein Inequality for sub-exponential sums) *Consider* $\sum_{i=1}^n X_i$*, where* X_i *are zero-mean, independent sub-exponential random variables with parameter* $\alpha = a_i, \nu = \nu_i$*. Let* $\alpha = \max_i a_i, \nu = \sqrt{\sum_{i=1}^n \nu_i^2}$*, we then notice that*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq t\right) \leq \begin{cases} \exp(-\frac{nt^2}{2\nu^2/n}) & \textit{for } \ 0 \leq t \leq \frac{\nu^2}{n\alpha}, \\ \exp(-\frac{nt}{2\alpha}) & \textit{for } \ t > \frac{\nu^2}{n\alpha}. \end{cases}$$

Lemma 22 *Consider $\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ki}$, where X_{ki} are independent sub-exponential random variables with parameter $\alpha = a_{ki}, \nu = \nu_{ki}$ and common expectation $E(X_{ki})$, then the event*

$$\left|\frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ki} - \mathbb{E}(X_{ki})\right| \geq C_L\sqrt{\frac{\log Kn}{Kn}}$$

will hold with probability at most $\frac{M}{(Kn)^8}$, where C_L is some sufficiently large constant and M depends on C_L.

The proof is a direct application of lemma 21 and union bound.

Lemma 23 (Rosenthal (1970), Theorem 3) *Suppose that $\{X_i\}_{i=1}^n$ are zero-mean and independent random variables. For any $p \geq 1$, there exists a constant R_p that for any $p \in \mathbb{N}$:*

$$\mathbb{E}\left(\left(\sum_{i=1}^{n} X_i\right)^{2p}\right) \leq R_p\left(\sum_{i=1}^{n}\mathbb{E}(X_i^{2p}) + \left(\sum_{i=1}^{n}\mathbb{E}(X_i^2)\right)^p\right).$$

Lemma 24 (Tail bounds under moment conditions) *Suppose that $\{X_i\}_{i=1}^n$ are zero-mean and independent random variables such that, for some fixed integer $p \geq 1$, they satisfy the moment bound $\|X_i\|_{P,2p} \leq C_p$. Then*

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq \delta\right) \leq B_p\left(\frac{1}{\sqrt{n}\delta}\right)^{2p} \qquad \text{for all}\delta > 0.$$

Proof This lemma is a simple application of lemma 23. ■

Lemma 25 *Under the assumptions 3 and 10,*

1.

$$\frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}(X_{ki}^T(\hat{\theta} - \theta^*))^2 \leq C\|\hat{\theta} - \theta^*\|_2^2,$$

2.

$$\left\|\hat{\beta} - \beta^*\right\|_1 \leq (s_1 \vee s_2)\sqrt{\frac{\log(p \vee n)}{n}},$$
$$\frac{1}{Kn}\sum_{j=1}^{m}\sum_{i=1}^{n}(X_{ki}^T(\hat{\beta} - \beta^*))^2 \leq \frac{(s_1 \vee s_2)\log(p \vee n)}{n}$$

with probability at least $1 - \frac{M}{n^8}$, where C is some sufficiently large constant, and $\hat{\theta}, \hat{\beta}$ are regularized PS, OR estimator, respectively.

Proof For the first claim, it is a result of lemma 6 and lemma 9 of Bradic et al. (2019), while for the second claim, it is a result of lemma S4 of Ning et al. (2020). ■

A.7 Bias without transferring Hessians

We are providing more details on the scenario where covariate shift exists but only first-order gradients are transferred. In particular, an additional bias term is introduced when only first-order gradients are transferred, rather than Hessians. Please find the detailed explanation below:

From Theorem 5 of Jordan et al. (2018), the error bound is determined by

$$\left\|\widetilde{\theta}-\theta^*\right\|_2 \leq C\left(\sqrt{\frac{s_1\log(p\vee Kn)}{Kn}}+\sqrt{s_1}\left\|\nabla^2 L_1(\boldsymbol{\theta}^*)-\nabla^2 L_N(\boldsymbol{\theta}^*)\right\|_\infty\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_1\right),$$

where $\bar{\boldsymbol{\theta}}$ is the initial estimator. If Lasso is applied to the local data to obtain initial $\bar{\boldsymbol{\theta}}$, then $\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_1 \leq Cs_1\sqrt{\frac{\log p\vee n}{n}}$ holds.

- ***If there does not exist covariate shift***, we are assuming $\mathbb{E}(\nabla^2 L_1(\boldsymbol{\theta}^*))=\mathbb{E}(\nabla^2 L_k(\boldsymbol{\theta}^*))$ for all $k\in\{1,\ldots,K\}$, then we have

$$\left\|\nabla^2 L_1(\boldsymbol{\theta}^*)-\nabla^2 L_N(\boldsymbol{\theta}^*)\right\|_\infty \leq \sqrt{\frac{\log(p\vee n)}{n}}.$$

  Following Jordan et al's idea, the error bound is

$$\left\|\widetilde{\theta}-\theta^*\right\|_2 \leq C\left(\sqrt{\frac{s_1\log(p\vee Kn)}{Kn}}+s_1^{3/2}\frac{\log(p\vee n)}{n}\right),$$

  This error bound is better than the classical Lasso l_2-error bound $\sqrt{\frac{s_1\log(p\vee n)}{n}}$, which only uses the local data from a single local site.

- However, ***if there exists covariate shift***, i.e. $\mathbb{E}(\nabla^2 L_1(\boldsymbol{\theta}^*))\neq\mathbb{E}(\nabla^2 L_k(\boldsymbol{\theta}^*))$ for some $k\in\{1,\ldots,K\}$, we have

$$\left\|\nabla^2 L_1(\boldsymbol{\theta}^*)-\nabla^2 L_N(\boldsymbol{\theta}^*)\right\|_\infty = O(1).$$

  If we still follow Jordan et al's idea by only transferring first-order gradients, the error bound becomes

$$\left\|\widetilde{\theta}-\theta^*\right\|_2 \leq C\left(\sqrt{\frac{s_1\log(p\vee Kn)}{Kn}}+s_1^{3/2}\sqrt{\frac{\log(p\vee n)}{n}}\right),$$

  This error bound is worse than the classical Lasso l_2-error bound where only local data is used.

Therefore, when there exits covariate shift between different sites, with the aim of improving the convergence rate, we require not only transferring the first-order gradients but also the Hessians to obtain the improved rate

$$\left\|\tilde{\theta}-\theta^*\right\|_2 \leq C_L\left(\sqrt{\frac{s_1\log(p\vee Kn)}{Kn}}+\frac{s_1^{3/2}\log^2(p\vee n)}{n}\right).$$

Appendix B. Additional simulations

In this section, to provide a better understanding of the methodology, we conduct additional numerical simulations. We compare the performance of the proposed method from two perspectives: splitting data versus not splitting it, and the comparison between splitting patient-level data and splitting sites. Without loss of generality, for $k = 1, \dots, K$ and $i = 1, \dots, n$, the treatment T_{ki} are generated from a logistic regression with $\pi_{ki} = \text{expit}(-0.5 + 0.5X_{ki1} + 0.3X_{ki2} - 0.3X_{ki3} + 0.3X_{ki4} - 0.3X_{ki5})$, the potential outcomes satisfy $Y_{ki}(1) = 2 + 0.3X_{ki1} + 0.2X_{ki2} - 0.2X_{ki3} + 0.2X_{ki4} - 0.2X_{ki5} + \epsilon_{ki1}$ and $Y_{ki}(0) = 1 + 0.3X_{ki1} + 0.2X_{ki2} - 0.2X_{ki3} + 0.2X_{ki4} - 0.2X_{ki5} + \epsilon_{ki0}$, where ϵ_{ki1} and ϵ_{ki0} are i.i.d from $N(0, 1)$, while the p-dimensional covariates are generated from $\mathbf{X}_{ki} \sim N(0, \mathbf{\Sigma}_k)$. For simplicity, we only consider the heterogeneous case (II) in the paper, specifically

(II) **Heterogeneous covariates (i.e., covariate shift) with $p < n$:** We consider the dimension with $p = 100$ and the sample size in each site is fixed at $n = 200$, while the covariance matrix $\mathbf{\Sigma}_k$ is set to be $\Sigma_{k;st} = \rho_k^{|s-t|}$, where $\rho_k \sim \text{Uniform}(0.2, 0.8)$ for $k = 1, \dots, K$. In this case, the simple size is larger than the dimension, and there is a shift in the distribution of covariates across sites.

The comparison results of splitting data versus not splitting it are depicted in Figure 6, where DisC^{2}o-HD-1 and DisC^{2}o-HD-2 represent our proposed approaches involving the splitting of K, while DisC^{2}o-HD-1-WS and DisC^{2}o-HD-2-WS denote the corresponding methods without splitting any data. We can see that the performance of DisC^{2}o-HD-2 are close to DisC^{2}o-HD-2-WS when K is large, while DisC^{2}o-HD-1 can outperform DisC^{2}o-HD-1-WS when K is large. In summary, while we employ data splitting for proof convenience, in numerical analysis, it's also feasible to apply the proposed method without splitting any data.

In the comparison between splitting patient-level data and splitting sites scenario, we fixed $K = 15$ and compare the ATE estimation error by repeating the process 100 times. The DisC^{2}o-HD-1 and DisC^{2}o-HD-2 are our proposed approaches involving splitting K, while the DisC^{2}o-HD-1-SN and DisC^{2}o-HD-2-SN approaches involve splitting patient-level data n. As depicted in Figure 7, it is evident that the DisC^{2}o-HD-1 and DisC^{2}o-HD-2 methods exhibit greater robustness and yield smaller ATE estimation errors compared to the DisC^{2}o-HD-1-SN and DisC^{2}o-HD-2-SN methods, thereby supporting our decision to split K.

References

Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):597–623, 2018.

Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed testing and estimation under sparse high dimensional models. *Annals of statistics*, 46(3):1352, 2018.

Fred M Behlen and Stephen B Johnson. Multicenter patient records research: security policies and tools. *Journal of the American Medical Informatics Association*, 6(6):435–443, 1999.

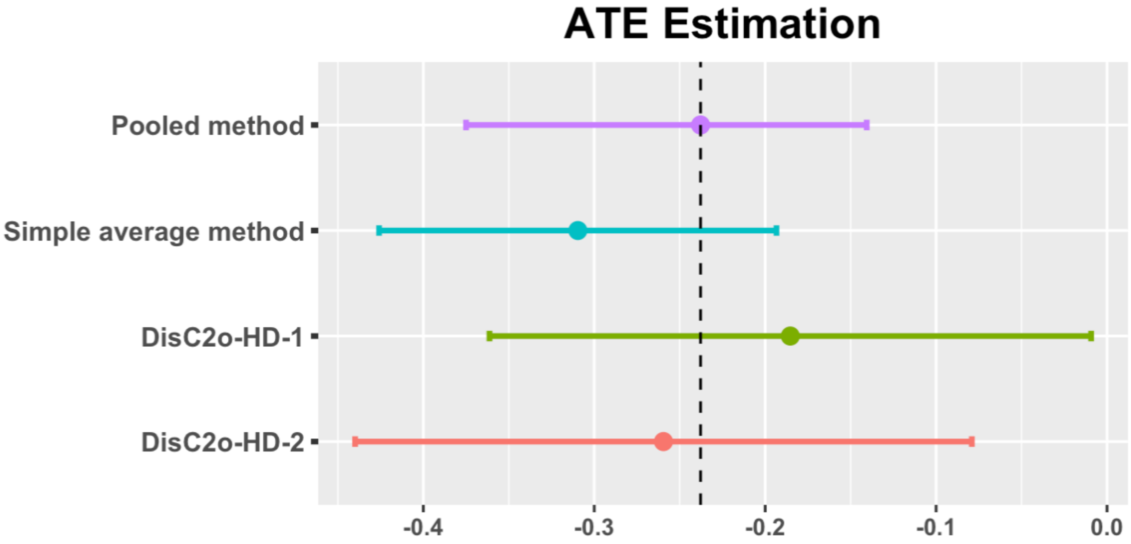

Figure 5: Data analysis results on the investigation on the impact of the vaccine on Post-Acute Sequelae of SARS-CoV-2 infection (PASC) in children during the Omicron period.

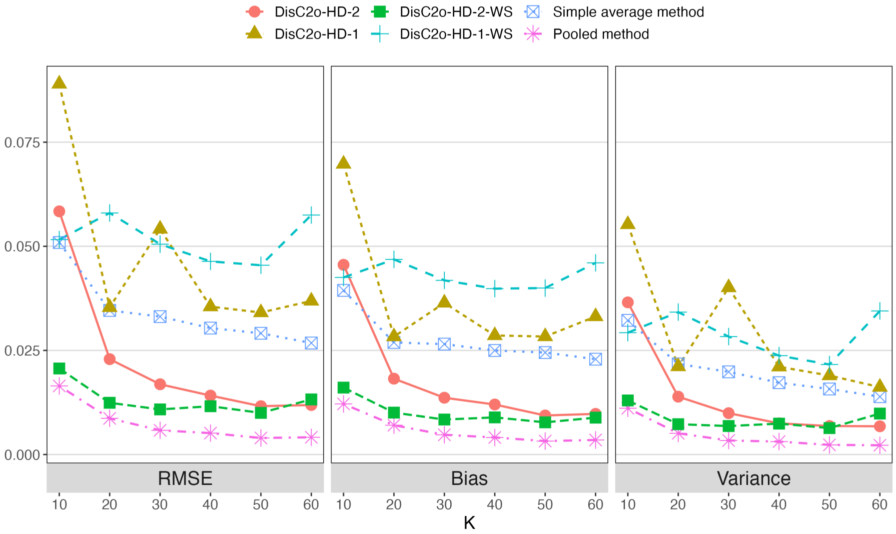

Figure 6: Comparison results of different methods under scenario (II) – heterogeneous covariates with $p < n$

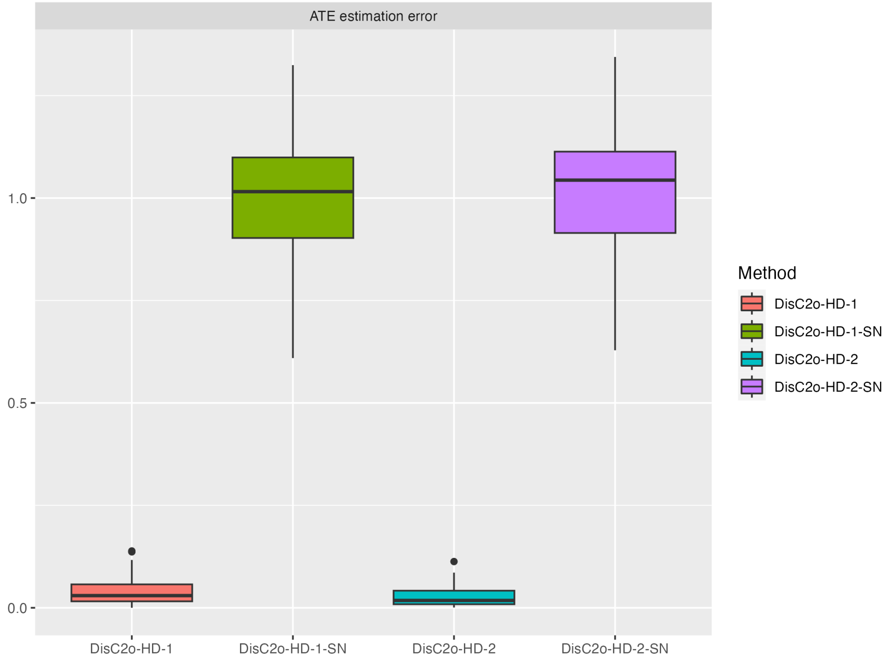

Figure 7: Comparison results of different methods under scenario (II) – heterogeneous covariates with $p < n$

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81 (2):608–650, 2014.

Alexandre Belloni, Victor Chernozhukov, Ivan Fernandez-Val, and Christian Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1): 233–298, 2017.

Jelena Bradic, Stefan Wager, and Yinchu Zhu. Sparsity double robust inference of average treatment effects. *arXiv preprint arXiv:1905.00744*, 2019.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Francis S Collins, Kathy L Hudson, Josephine P Briggs, and Michael S Lauer. Pcornet: turning a dream into reality, 2014.

Rui Duan, Mary Regina Boland, Jason H Moore, and Yong Chen. Odal: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. In *BIOCOMPUTING 2019: Proceedings of the Pacific Symposium*, pages 30–41. World Scientific, 2018.

Rui Duan, Mary Regina Boland, Zixuan Liu, Yue Liu, Howard H Chang, Hua Xu, Haitao Chu, Christopher H Schmid, Christopher B Forrest, John H Holmes, et al. Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *Journal of the American Medical Informatics Association*, 27(3):376–385, 2020a.

Rui Duan, Chongliang Luo, Martijn J Schuemie, Jiayi Tong, C Jason Liang, Howard H Chang, Mary Regina Boland, Jiang Bian, Hua Xu, John H Holmes, et al. Learning from local to global: An efficient distributed algorithm for modeling time-to-event data. *Journal of the American Medical Informatics Association*, 27(7):1028–1036, 2020b.

Rui Duan, Yang Ning, and Yong Chen. Heterogeneity-aware and communication-efficient distributed statistical inference. *Biometrika*, 109(1):67–83, 2022.

Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

Rachael L Fleurence, Lesley H Curtis, Robert M Califf, Richard Platt, Joe V Selby, and Jeffrey S Brown. Launching pcornet, a national patient-centered clinical research network. *Journal of the American Medical Informatics Association*, 21(4):578–582, 2014.

Charles P Friedman, Adam K Wong, and David Blumenthal. Achieving a nationwide learning health system. *Science translational medicine*, 2(57):57cm29–57cm29, 2010.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2998560.

Larry Han, Jue Hou, Kelly Cho, Rui Duan, and Tianxi Cai. Federated adaptive causal estimation (face) of target treatment effects. *arXiv preprint arXiv:2112.09313*, 2021.

Miguel A Hernán and James M Robins. Causal inference, 2010.

David A Hirshberg and Stefan Wager. Augmented minimax linear estimation. *The Annals of Statistics*, 49(6):3206–3227, 2021.

George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, et al. Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. *Studies in health technology and informatics*, 216:574, 2015.

Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 2018.

Jason D Lee, Qiang Liu, Yuekai Sun, and Jonathan E Taylor. Communication-efficient sparse regression. *The Journal of Machine Learning Research*, 18(1):115–144, 2017.

Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Advances in neural information processing systems*, 22, 2009.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical science*, 27(4):538–557, 2012.

Whitney K Newey and James R Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195, 2017.

Yang Ning, Peng Sida, and Kosuke Imai. Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika*, 107(3):533–554, 2020.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Haskell P Rosenthal. On the subspaces of $\mathcal{L}^p (p > 2)$ spanned by sequences of independent random variables. *Israel Journal of Mathematics*, 8(3):273–303, 1970.

Tamara D Simon, Wren Haaland, Katherine Hawley, Karen Lambka, and Rita Mangione-Smith. Development and validation of the pediatric medical complexity algorithm (pmca) version 3.0. *Academic pediatrics*, 18(5):577–580, 2018.

Zhiqiang Tan. Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *The Annals of Statistics*, 48(2):811–837, 2020a.

Zhiqiang Tan. Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika*, 107(1):137–158, 2020b.

Tanayott Thaweethai, Sarah E Jolley, Elizabeth W Karlson, Emily B Levitan, Bruce Levy, Grace A McComsey, Lisa McCorkell, Girish N Nadkarni, Sairam Parthasarathy, Upinder Singh, et al. Development of a definition of postacute sequelae of sars-cov-2 infection. *JAMA*, 329(22):1934–1946, 2023.

Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Mark J van der Laan, Sherri Rose, Wenjing Zheng, and Mark J van der Laan. Cross-validated targeted minimum-loss-based estimation. *Targeted learning: causal inference for observational and experimental data*, pages 459–474, 2011.

Jialei Wang, Mladen Kolar, Nathan Srebro, and Tong Zhang. Efficient distributed learning with sparsity. In *International conference on machine learning*, pages 3636–3645. PMLR, 2017.

Robert WM Wedderburn. Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika*, 61(3):439–447, 1974.

Qiong Wu, Jiayi Tong, Bingyu Zhang, Dazheng Zhang, Jiajie Chen, Yuqing Lei, Yiwen Lu, Yudong Wang, Lu Li, Yishan Shen, et al. Real-world effectiveness of bnt162b2 against infection and severe diseases in children and adolescents. *Annals of Internal Medicine*, 177(2):165–176, 2024.

Yonghui Wu, Jeremy L Warner, Liwei Wang, Min Jiang, Jun Xu, Qingxia Chen, Hui Nian, Qi Dai, Xianglin Du, Ping Yang, et al. Discovery of noncancer drug effects on survival in electronic health records of patients with cancer: a new paradigm for drug repurposing. *JCO clinical cancer informatics*, 3:1–9, 2019.

Elke Wynberg, Alvin X Han, Anders Boyd, Hugo DG van Willigen, Anouk Verveen, Romy Lebbink, Karlijn van der Straten, Neeltje Kootstra, Marit J van Gils, Colin Russell, et al. The effect of sars-cov-2 vaccination on post-acute sequelae of covid-19 (pasc): A prospective cohort study. *Vaccine*, 40(32):4424–4431, 2022.

Ruoxuan Xiong, Allison Koenecke, Michael Powell, Zhu Shen, Joshua T Vogelstein, and Susan Athey. Federated causal inference in heterogeneous observational data. *arXiv preprint arXiv:2107.11732*, 2021.

Hua Xu, Melinda C Aldrich, Qingxia Chen, Hongfang Liu, Neeraja B Peterson, Qi Dai, Mia Levy, Anushi Shah, Xue Han, Xiaoyang Ruan, et al. Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. *Journal of the American Medical Informatics Association*, 22(1):179–191, 2015.

Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14(68): 3321–3363, 2013. URL http://jmlr.org/papers/v14/zhang13b.html.