Heterogeneity-aware and communication-efficient distributed statistical inference

BY RUI DUAN

Department of Biostatistics, Harvard University, Boston, Massachusetts 02115, U.S.A.
rduan@hsph.harvard.edu

YANG NING

Department of Statistics and Data Science, Cornell University, Ithaca, New York, 14853, U.S.A.
yn265@cornell.edu

YONG CHEN

Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.
ychen123@upenn.edu

SUMMARY

In multicenter research, individual-level data are often protected against sharing across sites. To overcome the barrier of data sharing, many distributed algorithms, which only require sharing aggregated information, have been developed. The existing distributed algorithms usually assume the data are homogeneously distributed across sites. This assumption ignores the important fact that the data collected at different sites may come from various sub-populations and environments, which can lead to heterogeneity in the distribution of the data. Ignoring the heterogeneity may lead to erroneous statistical inference. In this paper, we propose distributed algorithms which account for the heterogeneous distributions by allowing site-specific nuisance parameters. The proposed methods extend the surrogate likelihood approach (Wang et al., 2017; Jordan et al., 2018) to the heterogeneous setting by applying a novel density ratio tilting method to the efficient score function. The proposed algorithms maintain the same communication cost as the existing communication-efficient algorithms. We establish a non-asymptotic risk bound for the proposed distributed estimator and its limiting distribution in the two-index asymptotic setting which allows both sample size per site and the number of sites to go to infinity. In addition, we show that the asymptotic variance of the estimator attains the Cramér-Rao lower bound when the number of sites is in rate smaller than the sample size at each site. Finally, we use simulation studies and a real data application to demonstrate the validity and feasibility of the proposed methods.

Some key words: Data integration; distributed inference; efficient score; surrogate likelihood; two-index asymptotics

1. INTRODUCTION

The growth of availability and variety of clinical data has induced the trend of multicenter research (Sidransky et al., 2009). Multicenter research confers many distinct advantages over single-center studies, including the ability to study rare exposures/outcomes that require larger

2

sample sizes, accelerating the discovery of more generalizable findings, and bringing together investigators who share and leverage resources, expertise, and ideas (Cheng et al., 2017). Since individual-level information is often protected by privacy regularities and rules, directly pooling data across multiple clinical sites is less feasible or requires large amount of operational efforts (Barrows Jr & Clayton, 1996). As a consequence, healthcare systems need more effective tools for evidence synthesis across clinical sites.

Distributed algorithms, also known as "divide-and-conquer" procedures, have been applied to multicenter studies. In the classical divide-and-conquer framework, the entire data set is split into multiple subsets and the final estimator is obtained by averaging the local estimators computed using the data from each subset (Li et al., 2013; Chen & Xie, 2014; Lee et al., 2017; Tian & Gu, 2016; Zhao et al., 2016; Lian & Fan, 2017; Battey et al., 2018; Wang et al., 2019). The class of methods adopts the same principle as meta-analysis in the area of evidence synthesis and systematic review, where the local estimates are combined through a fixed effect or random effects model (DerSimonian & Laird, 1986). When the number of research sites is relatively small, these averaging type of methods are able to perform equally well as the combined analysis using data from all the sites (Hedges, 1983; Olkin & Sampson, 1998; Battey et al., 2018). When the number of research sites is large, as we will demonstrate in the simulation studies, these averaging methods may not be as good as the combined analysis. More importantly, when studying rare conditions, some clinical sites do not have enough number of cases to achieve the asymptotic properties. In such cases, the averaging methods can be suboptimal.

Recently, Wang et al. (2017) and Jordan et al. (2018) proposed a novel surrogate likelihood approach, which approximates the higher order derivatives of the global likelihood by using the likelihood function in a local site. This method has low communication cost and improves the performance of the average method especially when the number of sites is large, see Duan et al. (2019) for a real data application to pharamcoepidemiology. From the practical perspective, the surrogate likelihood approach endowed a highly feasible framework for sharing sensitive data in a collaborative environment, especially in biomedical sciences, where the lead investigators often have access to the individual-level data in their home institute, and the collaborative investigators from other sites are willing to share summary statistics but not individual-level information.

Most of the aforementioned distributed algorithms assumed that the data at different sites are independently and identically distributed. However, a prominent concern in multi-center analysis is that there may exist a non-negligible degree of heterogeneity across sites because the samples collected in different sites may come from different sub-populations and environments. One concrete example is the Observational Health Data Sciences and Informatics consortium, which contains over 82 clinical databases from over 20 countries around the world (Hripcsak et al., 2015). The amount of heterogeneity cannot be ignored when implementing distributed algorithms in such healthcare networks.

To the best of our knowledge, Zhao et al. (2016) is the only work in this area that considers a similar heterogeneous setting. They generalized the divide-and-conquer approach by averaging all the local estimators and studied theoretical properties under the partially linear model. Different from this work, we propose to account for the heterogeneous distributions via a general parametric likelihood framework by allowing site-specific nuisance parameters. In particular, we extend the surrogate likelihood function approach to a surrogate estimating equation approach, and propose a density-ratio tilted surrogate efficient score function which only requires the individual-level data from a local site and summary statistics from the other sites. To reduce the influence of estimation of the site-specific nuisance parameters, we propose to use the efficient score function for distributed inference rather than the score function as in Jordan et al. (2018). We further adjust for the degree of heterogeneity by applying a novel density ratio tilt-

ing method to the efficient score function. We refer the resulting score function to the surrogate efficient score function. The estimator is defined as the root of this function. We show that the communication cost of the proposed algorithm is of the same order as Jordan et al. (2018) assuming no heterogeneity and therefore is communication-efficient. Theoretically, we show that our estimator approximates the global maximum likelihood estimator with a faster rate than the average approach in the two-index asymptotic setting; see Remarks 3 and 4. From the inference perspective, our estimator attains the Cramér-Rao lower bound whereas the average approach has larger asymptotic variance and is not efficient when the number of sites is less than the sample size at each site; see Remark 6. We show that the proposed estimator outperforms the average approach in numerical studies.

2. WANG ET AL. (2017) AND JORDAN ET AL. (2018)'S SURROGATE LIKELIHOOD APPROACH FOR HOMOGENEOUS DISTRIBUTIONS

In this section, we briefly review the surrogate likelihood approach for distributed inference by Wang et al. (2017) and Jordan et al. (2018). Consider a general parametric likelihood framework, where the random variable Y follows the density function $f(y; \theta)$ indexed by a finite dimensional unknown parameter θ. In the distributed inference problem, we suppose there are K different sites. Denote $\{Y_{ij}\}$ to be the i-th observation in the j-th site. For notation simplicity, we assume that each site has equal sample size n. The existing works on distributed inference such as Wang et al. (2017) and Jordan et al. (2018) further assume that all the observations are independently and identically distributed across sites, $Y_{ij} \sim f(y; \theta)$. Under this assumption, the combined log likelihood function can be written as

$$L(\theta) = \frac{1}{Kn} \sum_{j=1}^{K} \sum_{i=1}^{n} \log f(y_{ij}; \theta) := \frac{1}{K} \sum_{j=1}^{K} L_j(\theta),$$

where $L_j(\theta) = \sum_{i=1}^{n} \log f(y_{ij}; \theta)/n$ is the log-likelihood function obtained at each site. Due to the communication constraint and privacy concerns, one cannot directly combine data across multiple sites to compute the maximum likelihood estimator. Motivated by the following Taylor expansion of the combined likelihood function around some initial value $\bar{\theta}$,

$$L(\theta) = L(\bar{\theta}) + \nabla L(\bar{\theta})^{\mathrm{T}}(\theta - \bar{\theta}) + \sum_{k=2}^{\infty} \frac{1}{k!} \nabla^k L(\bar{\theta})(\theta - \bar{\theta})^{\otimes k}, \tag{1}$$

Wang et al. (2017) and Jordan et al. (2018) proposed to construct a surrogate likelihood function by approximating all the higher-order derivatives in equation (1) using the individual-level data in one of the K sites (such as the first site). When the data are identically and independently distributed across sites, it holds that $\nabla^k L_1(\bar{\theta}) - \nabla^k L(\bar{\theta}) = o_P(1)$ for any $k \geq 0$, where $L_1(\theta)$ is the log-likelihood at the first site. Thus, $\nabla^k L_1(\bar{\theta})$ is an asymptotically unbiased surrogate of $\nabla^k L(\bar{\theta})$. However, in a distributed framework, communicating $\nabla L_j(\theta)$ from site j to site 1 requires to transfer only $O(d)$ numbers where d is the dimension of θ, whereas communicating higher order derivatives can be very costly. Replacing $\nabla^k L(\bar{\theta})$ with $\nabla^k L_1(\bar{\theta})$, the communication of the higher-order derivatives across sites can be avoided. Hence, by replacing $\sum_{k=2}^{\infty} \nabla^k L(\bar{\theta})(\theta - \bar{\theta})^{\otimes k}/k!$ with $\sum_{k=2}^{\infty} \nabla^k L_1(\bar{\theta})(\theta - \bar{\theta})^{\otimes k}/k!$, which also equals to $L_1(\theta) - \nabla L_1(\bar{\theta})^{\mathrm{T}}(\theta - \bar{\theta})$ and dropping the terms independent of θ, the surrogate likelihood is defined as

$$\tilde{L}(\theta) := L_1(\theta) + \{\nabla L(\bar{\theta}) - \nabla L_1(\bar{\theta})\}^{\mathrm{T}} \theta. \tag{2}$$

4

From the perspective of estimating equations, the surrogate likelihood approach is equivalent to a surrogate score approach which approximates the combined score function $\nabla L(\theta)$ by

$$\tilde{S}(\theta) := \nabla L(\bar{\theta}) + \nabla L_1(\theta) - \nabla L_1(\bar{\theta}).$$

The theoretical properties of the estimator obtained by maximizing the surrogate likelihood function (or solving the surrogate score function) have been thoroughly studied; see Wang et al. (2017) and Jordan et al. (2018) for details. The theoretical properties of the estimator obtained by maximizing the surrogate likelihood function (or solving the surrogate score function) have been thoroughly studied; see Wang et al. (2017) and Jordan et al. (2018) for details.

3. SURROGATE EFFICIENT SCORE METHOD FOR HETEROGENEOUS DISTRIBUTIONS

We consider a heterogeneous setting by assuming the i-th observation in the j-th site satisfies

$$Y_{ij} \sim f(y; \theta_j), \quad \text{for } i \in \{1, \ldots, n\} \text{ and } j \in \{1, \ldots, K\},$$

where the unknown parameter θ_j can be decomposed into $\theta_j = (\beta, \gamma_j) \in R^d$. In this partition, β is a p-dimensional parameter of interest assumed to be common in every site, which is the main motivation of evidence synthesis, and the $(d-p)$-dimensional nuisance parameter γ_j is allowed to be different across sites. The true value of θ_j is denoted by θ_j^*.

If all patient-level data could be pooled together, the combined log-likelihood function is

$$L_N(\beta, \Gamma) = \frac{1}{Kn} \sum_{j=1}^{K} \sum_{i=1}^{n} \log f(y_{ij}; \beta, \gamma_j) := \frac{1}{K} \sum_{j=1}^{K} L_j(\theta_j),$$

where $L_j(\theta) = \sum_{i=1}^{n} \log f(y_{ij}; \theta_j)/n$ and $\Gamma = \{\gamma_j\}_{j \in \{1, \ldots, K\}} \in R^{(d-p)K}$. In a distributed setting, the method reviewed in Section 2 is not directly applicable due to the following two reasons: the higher order derivatives of the log likelihood function in any site is a biased surrogate of the corresponding higher order derivatives of $L_N(\beta, \Gamma)$, and the total number of nuisance parameters $\dim(\Gamma) = (d-p)K$ increases with sample size n if we allow K to increase with n.

With the site-specific nuisance parameters, we propose to approximate the efficient score function instead of the score function. Motivated from theories of semiparametric models, the efficient score function is a way of reducing the influence of the less accurate estimation of the site-specific γ_j which is essentially a projection of the score function of β on the space that is orthogonal to the space spanned by the score function of nuisance parameter γ_j (Van der Vaart, 2000). In our setting, it is defined as

$$s_j(y; \beta, \gamma_j) = \nabla_\beta \log f(y; \beta, \gamma_j) - I_{\beta\gamma}^{(j)} I_{\gamma\gamma}^{(j)-1} \nabla_\gamma \log f(y; \beta, \gamma_j)$$

where $I_{\gamma\gamma}^{(j)}$ and $I_{\gamma\beta}^{(j)}$ are the corresponding submatrices of the information matrix in the j-th site, i.e., $I^{(j)} = \mathbb{E}\{-\nabla^2 L_j(\theta_j^*)\}$. In parametric models, the estimator of β obtained from solving the efficient score function defined above has the asymptotic variance reaching the Cramer-Rao lower bound, which is considered as an efficient estimator. In addition, it satisfies that $E\{\nabla_\gamma s_j(y; \beta, \gamma_j)\} = 0$, which shows it is less sensitive to small perturbations of the nuisance parameter γ_j. We then define

$$S(\beta, \Gamma) = \frac{1}{Kn} \sum_{j=1}^{K} \sum_{i=1}^{n} s_j(y_{ij}; \beta, \gamma_j).$$

Treating the above combined efficient score function as a target function, we aim to construct a surrogate efficient score equation to approximate the target function using individual-level data from the first site and summary-level data from the other sites. To explain the origin of our estimator, we first consider an ideal situation where we know the true parameter value γ_j^*. Using the key idea of the surrogate likelihood approach, we aim to construct a function $g^*(y; \beta)$ in the first site such that

$$E_{\theta_1^*}\{\nabla_\beta^k g^*(Y_{i1}; \beta)\} = E\{\nabla_\beta^k S(\beta, \Gamma^*)\}, \tag{3}$$

holds for any $k \geq 1$, where we use $E_{\theta_j^*}(\cdot)$ to denote the expectation with respect to the distribution $f(y, \beta^*, \gamma_j^*)$, $E(\cdot)$ to denote the expectation with respect to the joint distribution of the full data, and Γ^* to denote the true value of Γ. The right hand side of equation (3) can be written as

$$E\{\nabla_\beta^k S(\beta, \Gamma^*)\} = \frac{1}{K} \sum_{j=1}^{K} E_{\theta_j^*}\{\nabla_\beta^k s_j(Y_{ij}; \beta, \gamma_j^*)\}.$$

However, the function $g^*(Y_{i1}; \beta)$ only involves samples in the first local site, which follows the distribution $f(y, \beta^*, \gamma_1^*)$ different from $f(y, \beta^*, \gamma_j^*)$ for $j \neq 1$. To achieve equation (3), we propose to construct $g^*(y; \beta)$ by using the density ratio tilting method

$$g^*(y; \beta) = \frac{1}{K} \sum_{j=1}^{K} \frac{f(y; \beta^*, \gamma_j^*)}{f(y; \beta^*, \gamma_1^*)} s_j(y_{ij}; \beta, \gamma_j^*),$$

where the density ratio $f(y; \beta^*, \gamma_j^*)/f(y; \beta^*, \gamma_1^*)$ is the adjustment that accounts for the heterogeneity of the distributions. It can be shown that $E_{\theta_1^*}\{\nabla_\beta^k g^*(Y_{i1}; \beta)\} = E\{\nabla_\beta^k S(\beta, \Gamma^*)\}$ holds for any $k \geq 0$ and observation Y_{i1} in the first local site (see Supplementary Material for details).

The map $g^*(y; \beta)$ cannot be computed in practice as it depends on the unknown parameters β^*, γ_j^*, and the information matrix $I^{(j)}$. Nevertheless, a natural surrogate can be used instead, by plugging in some initial estimators $\bar{\beta}$ and $\bar{\gamma}_j$, and replacing the matrix $I^{(j)}$ by its density ratio tilting estimator $\tilde{H}^{(1,j)}$, defined as

$$\tilde{H}^{(1,j)} = -\frac{1}{n} \sum_{i=1}^{n} \nabla^2 \log f(y_{i1}; \bar{\beta}, \bar{\gamma}_j) \frac{f(y_{i1}; \bar{\beta}, \bar{\gamma}_j)}{f(y_{i1}; \bar{\beta}, \bar{\gamma}_1)}.$$

We then have

$$g(y; \beta, \bar{\beta}, \bar{\Gamma}) = \frac{1}{K} \sum_{j=1}^{K} \left[\frac{f(y; \bar{\beta}, \bar{\gamma}_j)}{f(y; \bar{\beta}, \bar{\gamma}_1)} \left\{ \nabla_\beta \log f(y; \beta, \bar{\gamma}_j) - \tilde{H}_{\beta\gamma}^{(1,j)} \{\tilde{H}_{\gamma\gamma}^{(1,j)}\}^{-1} \nabla_\gamma \log f(y; \beta, \bar{\gamma}_j) \right\} \right].$$

We denote $U_1(\beta; \bar{\beta}, \bar{\Gamma}) = \sum_{i=1}^{n} g(y_{i1}; \beta, \bar{\beta}, \bar{\Gamma})/n$, and define the surrogate efficient score function as

$$\tilde{U}(\beta; \bar{\beta}, \bar{\Gamma}) = U_1(\beta; \bar{\beta}, \bar{\Gamma}) + \frac{1}{K} \sum_{j=1}^{K} \{\nabla_\beta L_j(\bar{\beta}, \bar{\gamma}_j) - \bar{H}_{\beta\gamma}^{(j)}(\bar{H}_{\gamma\gamma}^{(j)})^{-1} \nabla_\gamma L_j(\bar{\beta}, \bar{\gamma}_j)\} - U_1(\bar{\beta}; \bar{\beta}, \bar{\Gamma}),$$

where $\bar{H}_{\beta\gamma}^{(j)} = \nabla_{\beta\gamma} L_j(\bar{\beta}, \bar{\gamma}_j)$ and $\bar{H}_{\gamma\gamma}^{(j)} = \nabla_{\gamma\gamma} L_j(\bar{\beta}, \bar{\gamma}_j)$. Recall that $U_1(\beta; \bar{\beta}, \bar{\Gamma})$ is constructed based on the samples in Site 1. Thus the surrogate efficient score only requires to transfer a p-dimensional score vector $S_j(\bar{\beta}, \bar{\gamma}_j) = \nabla_\beta L_j(\bar{\beta}, \bar{\gamma}_j) - \bar{H}_{\beta\gamma}^{(j)}(\bar{H}_{\gamma\gamma}^{(j)})^{-1} \nabla_\gamma L_j(\bar{\beta}, \bar{\gamma}_j)$ from each site together with some initial estimators. The surrogate efficient score estimator $\tilde{\beta}$ is obtained by

6

solving the following equation for β within Site 1,

$$\tilde{U}(\beta; \bar{\beta}, \bar{\Gamma}) = 0. \tag{4}$$

In Section 4, we show that the estimation accuracy of the above estimator $\tilde{\beta}$ can be further improved by iterating the above surrogate efficient score procedures. The method is summarized in the following algorithm. The estimator $\tilde{\beta}$ defined in equation (4) is equivalent to the estimator with $T = 1$ in the following algorithm, which is also known as a oneshot procedure.

Algorithm 1. Algorithm for the proposed surrogate efficient score estimator

1.   In Site $j = 1$ to $j = K$
2.   Obtain and broadcast $(\bar{\beta}_j, \bar{\gamma}_j) = \arg\max_{\beta,\gamma_j} L_j(\beta, \gamma_j)$;
3.   Choose a proper weight w_j and obtain $\bar{\beta} = \sum_{j=1}^{K} w_j \bar{\beta}_j / \{\sum_{j=1}^{K} w_j\}$;
4.   Calculate and transfer $S_j(\bar{\beta}, \bar{\gamma}_j)$ to Site 1;
5.   In Site 1
6.   Construct $\tilde{U}(\beta; \bar{\beta}, \bar{\Gamma})$ using $\bar{\beta}$, $\{\bar{\gamma}_j\}$, and $\{S_j(\bar{\beta}, \bar{\gamma}_j)\}$;
7.   Obtain $\tilde{\beta}^{(1)}$ by solving $\tilde{U}(\beta; \bar{\beta}, \bar{\Gamma}) = 0$;
8.   If $T = 1$, output $\tilde{\beta}^{(1)}$;
9.   If $T \geq 2$, for $t = 2$ to $t = T$:
10.   Broadcast $\bar{\beta}^{(t)} = \tilde{\beta}^{(t-1)}$;
11.   In Site $j = 1$ to $j = K$
12.   Transfer $\bar{\gamma}_j^{(t)} = \arg\max_{\gamma_j} L_j(\bar{\beta}^{(t)}, \gamma_j)$ and $S_j(\bar{\beta}^{(t)}, \bar{\gamma}_j^{(t)})$ to Site 1;
13.   In Site 1
14.   Construct $\tilde{U}(\beta; \bar{\beta}^{(t)}, \bar{\Gamma}^{(t)})$ using $\bar{\beta}^{(t)}$, $\{\bar{\gamma}_j^{(t)}\}$ and $\{S_j(\bar{\beta}^{(t)}, \bar{\gamma}_j^{(t)})\}$;
15.   Obtain $\tilde{\beta}^{(t)}$ by solving $\tilde{U}(\beta; \bar{\beta}^{(t)}, \bar{\Gamma}^{(t)}) = 0$;
16. Output $\tilde{\beta}^{(T)}$

Remark 1. The broadcast step (line 2) in the above algorithm can be done by transferring $\bar{\theta}_j$ from each site to Site 1, and Site 1 returns the initial estimator $\bar{\beta}$ to all the sites. It can also be done by uploading all $\bar{\theta}_j$ to a shared repository and obtaining $\bar{\beta}$ at each site. The initial estimator $\bar{\beta}$ is chosen as a weighted average of the local estimators $\bar{\beta}_j$. When $w_j = 1$ for all j, $\bar{\beta} = \sum_{j=1}^{K} \bar{\beta}_j / K$ is the average estimator (Zhao et al., 2016). We can also choose w_j to be the sample size of each site in the unbalanced design. When w_j is chosen as the inverse of the estimated variance of $\bar{\beta}_j$, the resulting estimator $\bar{\beta}$ is referred to as the fixed effect meta-analysis estimator. In this paper, we simply choose $w_j = 1$. The total communication cost per iteration is to transfer $O(Kd)$ numbers across all sites, where d is the dimension of $\theta_j = (\beta, \gamma_j)$. Comparing to the homogeneous setting, the communication cost is of the same order as Jordan et al. (2018), and is communication-efficient.

Remark 2. To further reduce the computational complexity of solving the surrogate efficient score function, we can approximate the combined efficient score function $S(\beta, \Gamma)$ via one-step Taylor expansion,

$$S(\beta, \Gamma) \approx S(\bar{\beta}, \bar{\Gamma}) + \nabla_\beta S(\bar{\beta}, \bar{\Gamma})(\beta - \bar{\beta}) + \nabla_\Gamma S(\bar{\beta}, \bar{\Gamma})(\Gamma - \bar{\Gamma}).$$

First, the property of the efficient score implies $\nabla_\Gamma S(\bar{\beta}, \bar{\Gamma}) \approx 0$ so that the last term can be neglected. Next, we replace the Hessian matrix $\nabla_\beta S(\bar{\beta}, \bar{\Gamma})$ computed by pooling over all the samples with the local surrogate $\nabla_\beta \check{U}_1(\bar{\beta})$, where $\check{U}_1(\beta) = U_1(\beta; \bar{\beta}, \bar{\Gamma})$. The resulting linear

approximation $S(\bar{\beta}, \bar{\Gamma}) + \nabla_\beta \check{U}_1(\bar{\beta})(\beta - \bar{\beta})$ as an estimating function of β defines the following estimator

$$\tilde{\beta}^O = \bar{\beta} - \{\nabla_\beta \check{U}_1(\bar{\beta})\}^{-1} S(\bar{\beta}, \bar{\Gamma}).$$

If we treat $\bar{\beta}$ as an initial estimator, the above estimator $\tilde{\beta}^O$ can be also viewed as a one-step estimator with a local surrogate of the Hessian matrix. Hereafter, our discussion will be focused on the estimator from Algorithm 1, as we show in Lemma S12 in Supplementary Materials that this one-step estimator shares the same theoretical properties as the estimator from Algorithm 1. When calculating the inverse of $(\bar{H}_{\gamma\gamma}^{(j)})^{-1}$ becomes a bottleneck of computation, we proposed a modified algorithm in Appendix A of Supplementary Material.

4. MAIN RESULTS

In this section, we study the theoretical properties of the surrogate efficient score estimator $\tilde{\beta}^{(T)}$ obtained from Algorithm 1. For convenience, we use C, C_1 and C_2 to denote positive constants which can vary from place to place. For sequence $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ ($a_n \gtrsim b_n$) if there exist a constant C such that $a_n \leq Cb_n$ ($a_n \geq Cb_n$) for all n. We first introduce the following assumptions.

Assumption 1. The parameter space of β, denoted by $\mathcal{B}$, is a compact and convex subset of $\mathbb{R}^p$. The true value β^* is an interior point of $\mathcal{B}$.

Assumption 2 (*Local Strong Convexity*). Define the expected second-order derivative of the negative log likelihood function to be $I^{(j)}(\theta_j) = E_{\theta_j^*}\{-\nabla^2 \log f(Y_{ij}; \theta_j)\}$. There exist positive constants (μ_-, μ_+), such that for any $j \in \{1, \ldots, K\}$, the population Hessian matrix $I^{(j)}(\theta_j^*)$ satisfies

$$\mu_- I_d \preceq I^{(j)}(\theta_j^*) \preceq \mu_+ I_d,$$

where I_d is the d dimensional identity matrix. Here, we use the notation that $A \preceq B$ for two matrices A and B if $A - B$ is positive semi-definite.

Assumption 3. For all $j \in \{1, \ldots, K\}$ and $i \in \{1, \ldots, n\}$, all components in $\nabla \log f(Y_{ij}, \theta_j^*)$ and $\nabla^2 \log f(Y_{ij}, \theta_j^*)$ are sub-exponential random variables.

Assumption 4 (*Identifiability*). For any $j \in \{1, \ldots, K\}$, we denote $F_j(\beta, \gamma_j) = E_{\theta_j^*}\{\log f(Y_{ij}; \beta, \gamma_j)\}$. The parameter (β^*, γ_j^*) is the unique maximizer of $F_j(\beta, \gamma_j)$.

Assumption 5 (*Smoothness*). For each $j \in \{1, \ldots, K\}$, let $\bar{\eta}_j = (\bar{\beta}, \bar{\gamma}_1, \bar{\gamma}_j)$, and define

$$H(\theta_j; y) = \nabla^2 \log f(y; \beta, \gamma_j),$$

and

$$\tilde{H}(\beta, \bar{\eta}_j; y) = \nabla^2 \log f(y; \beta, \bar{\gamma}_j) \frac{f(y; \bar{\beta}, \bar{\gamma}_j)}{f(y; \bar{\beta}, \bar{\gamma}_1)}.$$

Define $U_\theta(\rho) = \{\theta; \|\theta - \theta\|_2 \leq \rho\}$ for some radius $\rho > 0$. There exist some function $m_1(y)$ and $m_2(y)$, where $m_1(Y_{ij})$ and $m_2(Y_{ij})$ are sub-exponentially distributed for all $j \in \{1, \ldots, K\}$ and $i \in \{1, \ldots, n\}$, such that for any θ_j and $\theta_j' \in U_{\theta_j}(\rho)$, we have

$$\|H(\theta_j; y) - H(\theta_j'; y)\|_2 \leq m_1(y)\|\theta - \theta_j'\|_2.$$

8

And for any $\beta, \beta' \in U_\beta(\rho)$, $\bar{\eta}_j, \bar{\eta}'_j \in U_{\eta_j}(\rho)$, we have

$$\|\tilde{H}(\beta, \bar{\eta}_j; y) - \tilde{H}(\beta', \bar{\eta}'_j; y)\|_2 \leq m_2(y)\{\|\beta - \beta'\|_2 + \|\bar{\eta}_j - \bar{\eta}'_j\|_2\}.$$

Assumptions 1, 2, 4, and 5 are standard assumptions in the distributed inference literature; see Jordan et al. (2018). Assumption 3 is a general distributional requirements of the data, which covers a wide range of parametric models.

When all individual-level data can be pooled together, the global maximum likelihood estimator $(\hat{\beta}, \hat{\Gamma}) = \arg\max_{\beta, \Gamma} L_N(\beta, \Gamma)$ is considered as the gold standard in practice. The asymptotic property of the global estimator has been studied in Li et al. (2003) under the asymptotic regime $K/n \to c \in (0, \infty)$. Our first result characterizes a non-asymptotic bound for the distance between the global maximum likelihood estimator $\hat{\beta}$ and the true parameter value β^*.

LEMMA 1. *Under Assumptions 1-5, the global maximum likelihood estimator $\hat{\beta}$ satisfies*

$$E\|\hat{\beta} - \beta^*\|_2 \leq \frac{C_1}{(Kn)^{1/2}} + \frac{C_2}{n}$$

for some positive constants C_1 and C_2 not related to n and K.

To the best of our knowledge, this is one of the first nonasymptotic results on the rate of convergence of the maximum likelihood estimator in the presence of site-specific nuisance parameters. Under the classical two-index asymptotics setting, we allow the number of sites K to grow with the sample size n. As a result, the dimension of nuisance parameters Γ also increases with n. Let $N = Kn$ denote the total sample size. This lemma implies that the convergence rate of $\hat{\beta}$ is of order $O_p(N^{-1/2})$ when $K/n = O(1)$, which attains the optimal rate of convergence with known nuisance parameters. However, when $K/n \to \infty$, the estimator has a slower rate $O_p(1/n)$. In particular if n is fixed, the global maximum likelihood estimator is no longer consistent, which is known as the Neyman-Scott problem (Neyman et al., 1948).

In the following, we characterize the difference between the proposed estimator and the maximum likelihood estimator $\hat{\beta}$. We first focus on the estimator $\tilde{\beta}$ defined in equation (4), which is identical to $\tilde{\beta}^{(1)}$ in algorithm 1 with the number of iterations $T = 1$.

THEOREM 1. *Suppose Assumptions 1-5 hold. In Algorithm 1, if the number of iterations $T = 1$, assuming $n \gtrsim \log K$, we have*

$$\mathbb{E}\|\tilde{\beta}^{(1)} - \hat{\beta}\|_2 \leq \frac{C}{n}.$$

where C is a positive constant not related to n and K.

The above theorem shows that the proposed estimator with only one iteration converges to the global estimator $\hat{\beta}$ with a rate only depending on n. Together with Lemma 1, we obtain $\mathbb{E}\|\tilde{\beta}^{(1)} - \beta^*\|_2 \lesssim 1/(Kn)^{1/2} + 1/n$. In other words, the estimator has the same rate of convergence as the global maximum likelihood estimator.

Remark 3. When $K/n \to 0$, we showed in Lemma S.10 of Supplementary Material that the average estimator $\bar{\beta}^{(1)} = \sum_{j=1}^{K} \bar{\beta}_j^{(1)}/K$ defined in algorithm 1 satisfies $\mathbb{E}\|\bar{\beta}^{(1)} - \hat{\beta}\|_2 \gtrsim 1/(Kn)^{1/2}$. By comparing with the bound in Theorem 1, we have

$$\mathbb{E}\|\tilde{\beta}^{(1)} - \hat{\beta}\|_2 \leq C\left(\frac{K}{n}\right)^{1/2} \mathbb{E}\|\bar{\beta}^{(1)} - \hat{\beta}\|_2. \tag{5}$$

Thus our estimator $\tilde{\beta}^{(1)}$ is closer to the global maximum likelihood estimator than the average estimator under the condition $K/n \to 0$.

Our next result shows that after at least one iteration the estimator $\tilde{\beta}^{(T)}$ in Algorithm 1 with $T \geq 2$ has a tighter bound than $\tilde{\beta}^{(1)}$ in Theorem 1.

THEOREM 2. *Suppose all the assumptions in Theorem 1 hold. In Algorithm 1, if the number of iterations $T \geq 2$, we have*

$$\mathbb{E}\|\tilde{\beta}^{(T)} - \hat{\beta}\|_2 \leq \frac{C_1}{(K)^{1/2}n} + \frac{C_2}{n^{3/2}},$$

where C_1 and C_2 are positive constants not related to n and K.

Remark 4. The above theorem implies that when $K/n \to 0$, for any $T \geq 2$

$$\mathbb{E}\|\tilde{\beta}^{(T)} - \hat{\beta}\|_2 \leq Cn^{-1/2}\mathbb{E}\|\bar{\beta}^{(1)} - \hat{\beta}\|_2, \tag{6}$$

which improves the result in equation (5). When K is relatively small, our estimator $\tilde{\beta}^{(T)}$ with $T \geq 2$ is closer to the global maximum likelihood estimator by a factor of $n^{-1/2}$ than the average estimator. We also see an interesting fact that the dimension of the nuisance parameters has no effect on the relative error $\mathbb{E}\|\tilde{\beta}^{(T)} - \hat{\beta}\|_2/\mathbb{E}\|\bar{\beta}^{(1)} - \hat{\beta}\|_2$. This dimension-free phenomenon provides an explanation of why the proposed estimator consistently outperforms the average method in our simulation studies in Section 6.

Our next theorem establishes the asymptotic normality of the proposed estimator.

THEOREM 3. *Suppose all the assumptions in Theorem 1 hold. Define $I_{\beta|\gamma} = \sum_{j=1}^{K} I_{\beta|\gamma}^{(j)}/K$, where $I_{\beta|\gamma}^{(j)}$ is the partial information matrix of β defined as $I_{\beta|\gamma}^{(j)} = I_{\beta\beta}^{(j)} - I_{\beta\gamma}^{(j)}(I_{\gamma\gamma}^{(j)})^{-1}I_{\gamma\beta}^{(j)}$. Assuming $K = Cn^r$ for some fixed $r \in [0,1)$, we have for any $T \geq 1$, as $n \to \infty$,*

$$Kn(\tilde{\beta}^{(T)} - \beta^*)^{\mathrm{T}}I_{\beta|\gamma}(\tilde{\beta}^{(T)} - \beta^*) \to \chi_p^2.$$

To obtain the $\sqrt{(Kn)}$-asymptotic normality of the proposed estimator $\tilde{\beta}^{(T)}$, we have to restrict to the setting $K = Cn^r$ for some $r \in [0,1)$. In particular, when $K/n \to C \in (0,\infty)$ or equivalently $r = 1$, Li et al. (2003) showed that the maximum likelihood estimator $\hat{\beta}$ is asymptotically biased, that is $(Kn)^{1/2}I_{\beta|\gamma}^{1/2}(\hat{\beta} - \beta^*) \to N(b, I_p)$, for some $b \neq 0$. Since the proposed estimator $\tilde{\beta}^{(T)}$ (with $T \geq 2$) satisfies $\tilde{\beta}^{(T)} - \hat{\beta} = O_p(K^{-1/2}n^{-1} + n^{-3/2})$ by Theorem 2, it implies that the same asymptotic distribution holds for $\tilde{\beta}^{(T)}$, $(Kn)^{1/2}I_{\beta|\gamma}^{1/2}(\tilde{\beta}^{(T)} - \beta^*) \to_d N(b, I_p)$ for the same $b \neq 0$ if $r = 1$. The same limiting distribution also holds for $T = 1$. This leads to a phase transition of the limiting distribution of $\tilde{\beta}^{(T)}$ at $r = 1$. As a result, the condition $r \in [0,1)$ is essential for the asymptotic unbiasedness of $\tilde{\beta}^{(T)}$ and cannot be further relaxed.

Remark 5. The choice of the initial value $\bar{\theta}_j^{(1)}$ in line 3 of Algorithm 1 is not necessarily restricted to the local maximum likelihood estimator. Due to the use of the efficient score, the impact of the initial estimators of the nuisance parameters is alleviated. We can show that the conclusions of Theorem 1-3 still hold if $\bar{\theta}_j^{(1)}$ is replaced with any $\sqrt{n}$-consistent estimator.

It is well known that the average estimator is fully efficient under the homogeneous setting; see e.g., Battey et al. (2018) and Jordan et al. (2018). However, the following proposition shows that this estimator is no longer efficient under the considered heterogeneous setting.

PROPOSITION 1. *Recall that the average estimator is $\bar{\beta} = \sum_{j=1}^{K} \bar{\beta}_j / K$, where $(\bar{\beta}_j, \bar{\gamma}_j) = \arg\max_{\beta, \gamma_j} L_j(\beta, \gamma_j)$. Suppose all the conditions in Theorem 3 hold. We have as $n \to \infty$,*

$$Kn(\bar{\beta} - \beta^*)^{\mathrm{T}} \left\{ \frac{1}{K} \sum_{j=1}^{K} I_{\beta|\gamma}^{(j)}{}^{-1} \right\}^{-1} (\bar{\beta} - \beta^*) \to \chi_p^2.$$

Remark 6. In this remark, we compare the asymptotic variance of $\tilde{\beta}^{(T)}$ in Theorem 3 and $\bar{\beta}$ in Proposition 1. Our proposed estimator $\tilde{\beta}^{(T)}$ is efficient in the sense that its asymptotic variance is equal to the Cramér-Rao lower bound, i.e., $\lim_{K \to \infty} I_{\beta|\gamma} = \lim_{K \to \infty} \{\sum_{j=1}^{K} I_{\beta|\gamma}^{(j)}/K\}^{-1}$, for any $T \geq 1$. On the other hand, the average estimator is not efficient as $\lim_{K \to \infty} \sum_{j=1}^{K} I_{\beta|\gamma}^{(j)}{}^{-1}/K \succeq \lim_{K \to \infty} I_{\beta|\gamma}^{-1}$.

Finally, to construct the confidence interval of β^*, we need to provide a consistent estimator for the averaged partial information matrix $I_{\beta|\gamma}$. In the following theorem, we apply the density ratio tilting approach to estimate the variance using data only from the first local site.

THEOREM 4. *Suppose all the assumptions in Theorem 3 hold. Define*

$$\tilde{I}^{(j)} = -\frac{1}{Kn} \sum_{j=1}^{K} \sum_{i=1}^{n} \frac{f(y_{i1}, \tilde{\beta}^{(T)}, \bar{\gamma}_j^{(T)})}{f(y_{i1}, \tilde{\beta}^{(T)}, \bar{\gamma}_1^{(T)})} \nabla^2 \log f(y_{i1}; \tilde{\beta}^{(T)}, \bar{\gamma}_j^{(T)}),$$

and $\tilde{I}_{\beta|\gamma}^{(j)} = \tilde{I}_{\beta\beta}^{(j)} - I_{\beta\gamma}^{(j)}(\tilde{I}_{\gamma\gamma}^{(j)})^{-1}\tilde{I}_{\gamma\beta}^{(j)}$. We have as $n \to \infty$,

$$Kn(\tilde{\beta}^{(T)} - \beta^*)^{\mathrm{T}} \tilde{I}_{\beta|\gamma}^{(j)}(\tilde{\beta}^{(T)} - \beta^*) \to \chi_p^2.$$

5. REDUCE THE INFLUENCE OF THE LOCAL SITE

In a practical collaborative research network, each site can act as the local site and obtain an estimate using Algorithm 1. To further reduce the impact of the choice of the local site and improve the stability of the algorithm, we can combine these estimates in Algorithm 1 by an average approach. At the j-th site, we define the site-specific surrogate score function to be

$$\tilde{U}_j(\beta; \bar{\beta}, \bar{\Gamma}) = U_j(\beta; \bar{\beta}, \bar{\Gamma}) + \frac{1}{K} \sum_{k=1}^{K} \{\nabla_\beta L_k(\bar{\beta}, \bar{\gamma}_k) - \bar{H}_{\beta\gamma}^{(k)}(\bar{H}_{\gamma\gamma}^{(k)})^{-1}\nabla_\gamma L_k(\bar{\beta}, \bar{\gamma}_j)\} - U_j(\bar{\beta}; \bar{\beta}, \bar{\Gamma}),$$

where

$$U_j(\beta; \bar{\beta}, \bar{\Gamma}) = \frac{1}{Kn} \sum_{k=1}^{K} \sum_{i=1}^{n} \frac{f(y_{ij}; \bar{\beta}, \bar{\gamma}_k)}{f(y_{ij}; \bar{\beta}, \gamma_j)} \left\{ \nabla_\beta \log f(y_{ij}; \beta, \bar{\gamma}_k) - \tilde{H}_{\beta\gamma}^{(j,k)}\{\tilde{H}_{\gamma\gamma}^{(j,k)}\}^{-1}\nabla_\gamma \log f(y_{ij}; \beta, \bar{\gamma}_k) \right\}$$

and

$$\tilde{H}^{(j,k)} = -\frac{1}{n} \sum_{i=1}^{n} \nabla^2 \log f(y_{ij}; \bar{\beta}, \bar{\gamma}_k) \frac{f(y_{ij}; \bar{\beta}, \bar{\gamma}_k)}{f(y_{ij}; \bar{\beta}, \bar{\gamma}_j)}.$$

The surrogate score function $\tilde{U}_j(\beta; \bar{\beta}, \bar{\Gamma})$ is obtained using the individual-level data in the j-th site and summary-level data from the other $K-1$ sites. In this case, each site can obtain a surrogate efficient score estimator $\tilde{\beta}_j$ by solving $\tilde{U}_j(\beta; \bar{\beta}, \bar{\Gamma}) = 0$, and we further combine these estimators by $\tilde{\beta}_{all} = \sum_{j=1}^{K} \tilde{\beta}_j/K$. The algorithm is summarized below.

Algorithm 2. Algorithm for computing $\tilde{\beta}_{all}$

1. In Site $j = 1$ to $j = K$
2.   Obtain an initial estimator $\bar{\theta}_j = (\bar{\beta}_j, \bar{\gamma}_j)$ for the parameter β and γ_j;
3.   Broadcast $\bar{\theta}_j$;
4.   Choose $w_j^{(1)}$ and obtain $\bar{\beta}^{(1)} = \sum_{j=1}^{K} w_j^{(1)} \bar{\beta}_j / \{\sum_{j=1}^{K} w_j^{(1)}\}$;
5.   Obtain and broadcast $S_j(\bar{\beta}, \bar{\gamma}_j)$;
6.   Construct $\tilde{U}_j(\beta; \bar{\beta}, \bar{\Gamma})$ using $\bar{\beta}$, $\{\bar{\gamma}_j\}$ and $\{S_k(\bar{\beta}, \bar{\gamma}_k)\}_{1 \le k \le K}$;
7.   Obtain and broadcast $\tilde{\beta}_j$;
8. Obtain and output $\tilde{\beta}_{all}$

6. SIMULATION STUDY

We consider a logistic regression between a binary outcome Y and a binary exposure X, controlling for a confounding variable Z. It is assumed that for data in the k-th site, we have

$$\text{logit}\{Pr(Y = 1 \mid X, Z)\} = \gamma_{0k} + \beta X + \gamma_{1k} Z.$$

We set the true value of $\beta = -1$ for all the K sites. The nuisance parameters γ_{0k} and γ_{1k} are generated from the uniform distribution $\mathrm{U}(a - 1, a + 1)$ and the uniform distribution $\mathrm{U}(-2, 2)$, respectively. The binary exposure X is generated from a Bernoulli distribution with probability b, and the confounder variable is generated by $Z \sim N(X - 0.3, 1)$. Under each setting, we set the sample size n to be 100 and the number of sites K to be 10 or 50. We compare the performance of five different methods: (1) The estimator from averaging all local maximum likelihood estimators (Average); (2) The surrogate likelihood method in Jordan et al. (2018) assuming the homogeneous model (Homo); (3) The proposed estimator in Algorithm 1 with $T = 1$ (i.e., the oneshot algorithm) (M1); (4) The proposed estimator in Algorithm 1 with $T = 2$ (M2), and (5) The proposed estimator in Algorithm 2 (M3).

We first investigate how the prevalence of binary events in a regression model influences the performance of the compared methods. We vary the value of a and the probability b which control the prevalence of the exposure and the outcome, and consider the following four scenarios: (1) both the outcome and the exposure are common; (2) the outcome is rare and the exposure is common; (3) the outcome is common and the exposure is rare; (4) both the outcome and the exposure are rare. The parameter values for a and b are presented in Table S1 of Supplementary Material. We observed from Figure 1 that the surrogate likelihood method which ignores the heterogeneity has substantial bias in all settings. When both the outcome and exposure are common, all the proposed estimators and the average estimator perform well. The average estimator starts to show large bias when either the outcome or the exposure is rare. Our oneshot estimator (M1) reduces the bias of the average estimator in all settings. However, when both outcome and exposure are rare, the oneshot algorithm illustrates non-negligible bias and relatively large variation. Through only one more iteration, our estimator (M2) has sizeably improved performance and becomes more stable in the rare outcome or exposure setting. The estimator in Algorithm 2 (M3) also outperforms the oneshot method in terms of reducing the bias and variance, especially when both the outcome and the exposure are rare. When we increase the number of sites, the bias of the average method remains the same while the variation is smaller. Our proposed algorithms, however, are able to take advantage of the increased total sample size and provide estimates with smaller bias.

12

We then investigate whether the level of heterogeneity influences the performance of the compared methods. To alter the level of heterogeneity, we generate the nuisance parameters γ_{0k} from the uniform distribution $\mathrm{U}(-v, v)$, and γ_{1k} from the uniform distribution $\mathrm{U}(-2v, 2v)$. We increase v from 0.1 to 4. When v is 0.1, the heterogeneity of the values of nuisance parameters is small across sites. We observe from Figure 2 that all methods work comparably well, including the surrogate likelihood method which ignores the heterogeneity. As v increases from 0.1 to 4, the bias of the surrogate likelihood method is increasing. The average approach, in theory, is a consistent estimator, but is shown to have larger bias and variation when v is increasing. The proposed three methods have similar performance under all settings. We can observe a slightly increasing variation and more outlying points of the proposed methods when v is large. Overall, the proposed methods are more robust to the change of v compared to the other two methods.

Finally, we investigate how the dimension of the nuisance parameters affects the performance of the compared methods. We generate Z from $N(0, I_q)$, and the corresponding coefficient vector γ_{1k} is generated from q independent uniform distributions $\mathrm{U}(-1, 1)$. Including the intercept, the total dimension of the nuisance parameters is denoted as $d_\gamma = q + 1$. We increase d_γ from 2 to 14. From Figrue 3, we see as d_γ increases, the estimation errors of all compared methods become larger. M2 has slightly better performance compared to M1 and M3, implying that iterations might help reduce the bias. The average approach, however, has the worst performance compared to the other approaches when d_γ is large.
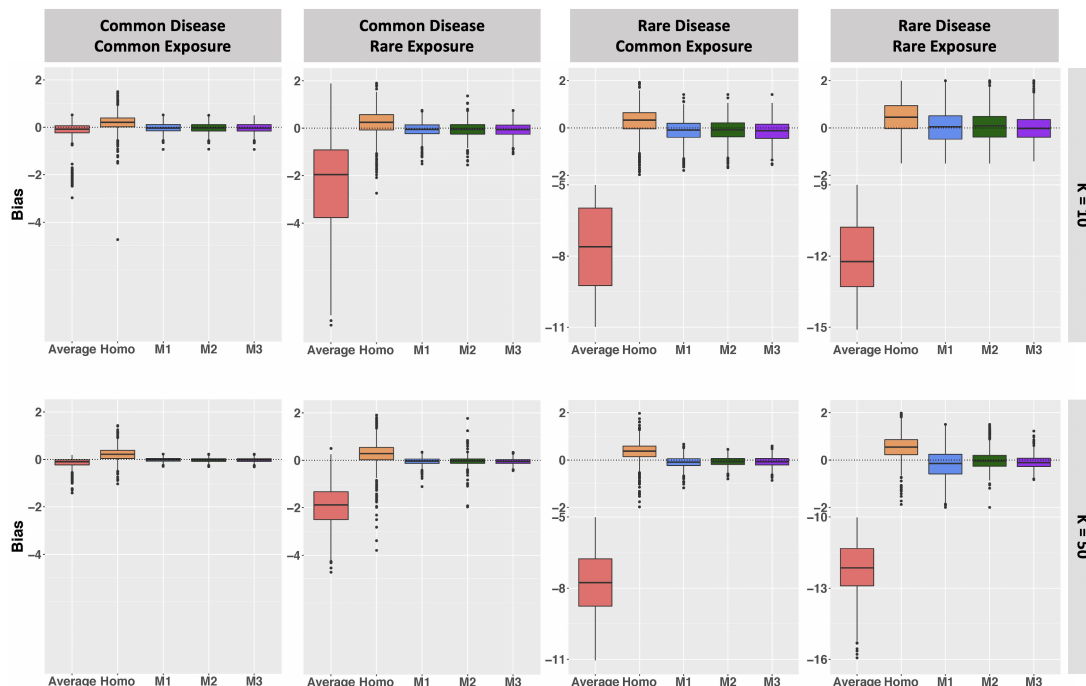
In sum, the increase of the rareness of the disease or exposure, the level of heterogeneity and the dimension of the model can increase the estimation errors of all compared methods. Ignoring heterogeneity in a distributed setting can lead to substantial amount of bias, and our proposed methods can greatly improve the estimation accuracy based on the commonly used average approach.

7. Real data application

We applied the proposed algorithms to a dataset from a distributed research network. We aim to use data from five sites to quantify the association between mental disorders, including major depression and anxiety, with the risk of opioid use disorder using a logistic regression model. Each participating site extracted electronic health records between 01/01/2012 and 03/01/2019 for patients who had opioid prescription (including Codeine, Fentanyl, Hydromorphone, Meperidine, Methadone, Morphine, Oxycodone, Tramadol, Hydrocodone, Buprenorphine), and no cancer or diagnosis of opioid use disorder before their first prescription. Among these patients who were exposed to opioid, a case of opioid use disorder is defined as having first diagnosis of opioid use disorder within 12 months after their first prescription and a control is defined as having no diagnosis of opioid use disorder in the entire time window. We obtained in total 1458 cases from the five clinical sites, and we randomly selected 2908 controls to maintain a case-control ratio of 1:2. In addition to the two risk factors of interest (i.e., major depression and anxiety), we requested a list of relevant covariate variables to be adjusted in the regression model, including age, gender, race (non-Hispanic White vs others), alcohol-related disorders, pain, cannabis-related disorder, cocaine-related disorder, nicotine-related disorder, smoking status (ever-smoker, non-smoker, and unknown), as well as the Charlson comorbidity index (Quan et al., 2005). Records with missing values were removed, resulting in sample sizes of 680, 1311, 920, 270, and 1106 from Site 1 to Site 5, respectively; see Appendix C of Supplementary Material for more information about data processing.

In the logistic regression model, we treated the coefficients of major depression and anxiety to be common parameters across sites, and all the other coefficients including the intercept were
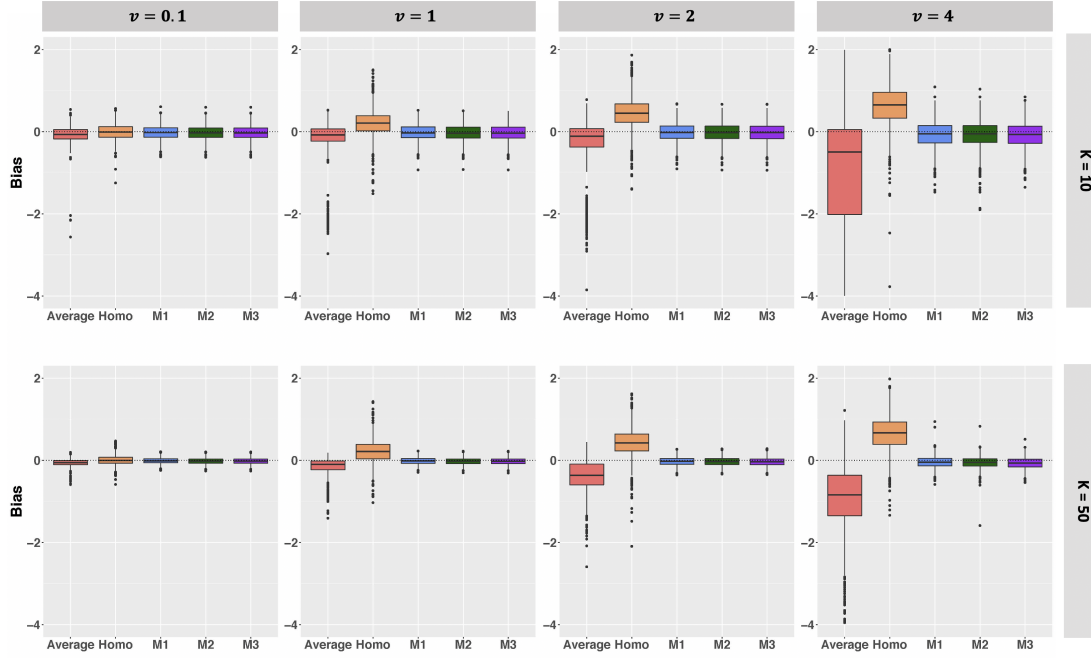
Fig. 1. Boxplots of the estimates (subtracted by the true parameter value) under the four parameter settings in Table 1. Each site has a sample size 100, and each setting is replicated 1000 times.

assumed to be site-specific. We chose Site 1 as the local site and applied our methods M1, M2, M3, and the average approach. The estimated log odds ratios with their 95% confidence intervals are shown in Figure 4. We observed consistent results from the three proposed methods for both anxiety and depression, suggesting one round of communication already led to stable estimation results. Anxiety was identified to be statistically significantly associated with opioid use disorder by all the methods. The relative difference based on the point estimates is about 18.4% comparing the average approach to M1. All methods failed to identify significant association between depression and opioid use disorder, possibly due to the relatively low prevalence (9%) of depression in the overall sample and the limited sample size. We observed opposite signs of the point estimates obtained from the proposed methods and the average approach, leading to a large relative difference of 164.3%. Since it has been shown in many studies that depression is associated with increased risk of developing opioid use disorder (Martins et al., 2012; Sullivan, 2018), the negative association estimated by the average approach can be less reliable. More details of model fitting results can be found in the Appendix C of Supplementary Material.

This real data application demonstrated the feasibility of implementing the proposed distributed algorithms in real-world distributed research networks. Although the average approach is easy to implement in practice, the proposed methods provide more reliable parameter estimates with only one extra step of sharing aggregate data.

Fig. 2. Boxplots of the estimates (subtracted by the true parameter value) when v takes values in $(0.1, 1, 2, 4)$, and K varies from 10 to 50. Each site has a sample size 100, and each setting is replicated 1000 times.

8. DISCUSSION

Motivated from a practical consideration that the data stored at different clinical sites are often heterogeneously distributed, we propose a surrogate efficient score approach for distributed inference. Our approach provides flexibility to allow site-specific nuisance parameters, and bridges the gap in the current research on healthcare distributed research networks. There are several future research directions. To account for the large number of clinical, environmental and genetic related variables in the modern healthcare datasets, it will be interesting to extend our method to the high-dimensional settings where either the dimension of β or the dimension of the nuisance parameters is larger than the sample size. Moreover, to extend the scope of the proposed framework, it would be of interest to relax the parametric assumption by using methods such as the generalized estimating equations (Liang & Zeger, 1986) and the generalized methods of moments (Hansen, 1982). However, as the density ratio tilting relies on the distributional assumption, it may require new methodological development to adjust for the heterogeneity under these new settings. Another practical challenge is that some sites only have a subset of all covariates. Recent work including Kundu et al. (2019) and Zhang et al. (2020) proposed novel methods to integrate summary statistics from external datasets with different covariate information. It is of interest to develop distributed inference that can handle heterogeneity and account for incomplete covariate information across sites. These topics are currently under investigation and will be reported in the future.

Fig. 3. Boxplots of the estimates (subtracted by the true parameter value) when d_γ takes values in $(2, 6, 10, 14)$, and K varies from 10 to 50. Each site has a sample size 100, and each setting is replicated 1000 times.
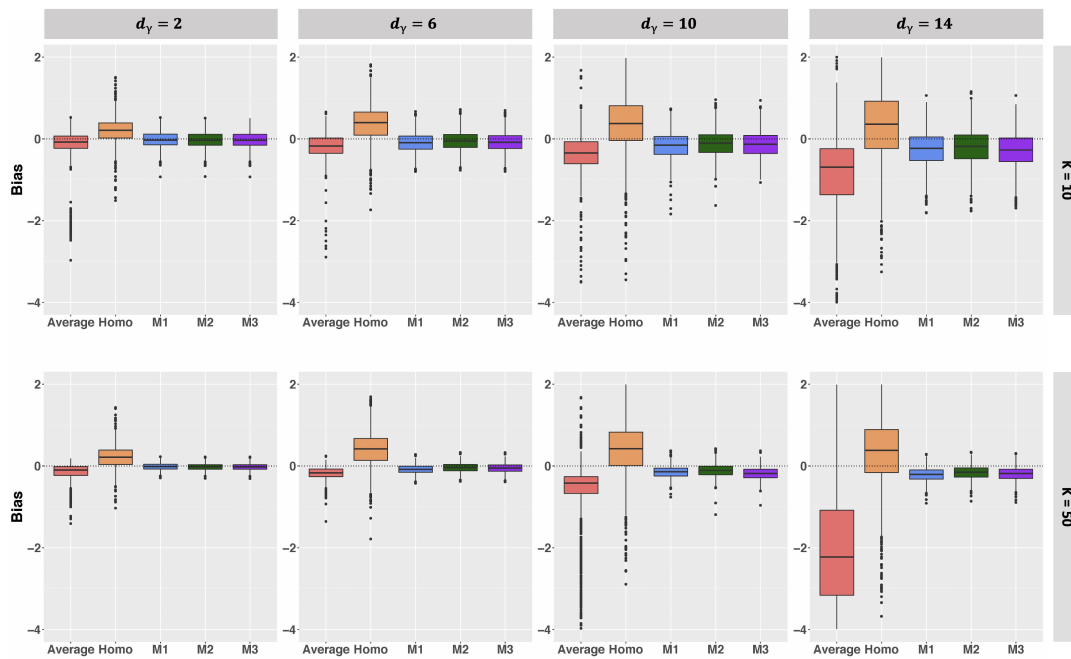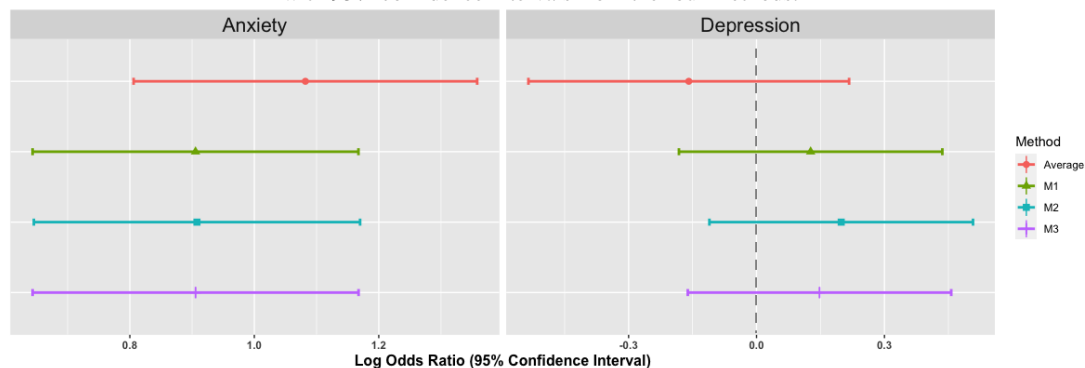

Fig. 4. Estimated log odds ratios of anxiety and depression, with 95% confidence intervals from the four methods.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes additional results of simulation and real data application, and proofs for Theorems 1-4, Lemma 1 and Proposition 1.

ACKNOWLEDGEMENT

16

REFERENCES

BARROWS JR, R. C. & CLAYTON, P. D. (1996). Privacy, confidentiality, and electronic medical records. *Journal of the American Medical Informatics Association* **3**, 139–148.

BATTEY, H., FAN, J., LIU, H., LU, J. & ZHU, Z. (2018). Distributed testing and estimation under sparse high dimensional models. *Annals of Statistics* **46**, 1352–1382.

CHEN, X. & XIE, M.-G. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica* , 1655–1684.

CHENG, A., KESSLER, D., MACKINNON, R., CHANG, T. P., NADKARNI, V. M., HUNT, E. A., DUVAL-ARNOULD, J., LIN, Y., PUSIC, M. & AUERBACH, M. (2017). Conducting multicenter research in healthcare simulation: Lessons learned from the inspire network. *Advances in Simulation* **2**, 6.

DERSIMONIAN, R. & LAIRD, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**, 177–188.

DUAN, R., BOLAND, M. R., MOORE, J. H. & CHEN, Y. (2019). ODAL: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. *Pacific Symposium on Biocomputing* , 30–41.

HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society* , 1029–1054.

HEDGES, L. V. (1983). Combining independent estimators in research synthesis. *British Journal of Mathematical and Statistical Psychology* **36**, 123–131.

HRIPCSAK, G., DUKE, J. D., SHAH, N. H., REICH, C. G., HUSER, V., SCHUEMIE, M. J., SUCHARD, M. A., PARK, R. W., WONG, I. C. K., RIJNBEEK, P. R. et al. (2015). Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. *Studies in Health Technology and Informatics* **216**, 574–578.

JORDAN, M. I., LEE, J. D. & YANG, Y. (2018). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association* , 1–14.

KUNDU, P., TANG, R. & CHATTERJEE, N. (2019). Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika* **106**, 567–585.

LEE, J. D., LIU, Q., SUN, Y. & TAYLOR, J. E. (2017). Communication-efficient sparse regression. *The Journal of Machine Learning Research* **18**, 115–144.

LI, H., LINDSAY, B. G. & WATERMAN, R. P. (2003). Efficiency of projected score methods in rectangular array asymptotics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**, 191–208.

LI, R., LIN, D. K. & LI, B. (2013). Statistical inference in massive data sets. *Applied Stochastic Models in Business and Industry* **29**, 399–409.

LIAN, H. & FAN, Z. (2017). Divide-and-conquer for debiased l 1-norm support vector machine in ultra-high dimensions. *The Journal of Machine Learning Research* **18**, 6691–6716.

LIANG, K.-Y. & ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

MARTINS, S. S., FENTON, M. C., KEYES, K. M., BLANCO, C., ZHU, H. & STORR, C. L. (2012). Mood and anxiety disorders and their association with non-medical prescription opioid use and prescription opioid-use disorder: longitudinal evidence from the national epidemiologic study on alcohol and related conditions. *Psychological medicine* **42**, 1261–1272.

NEYMAN, J., SCOTT, E. L. et al. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32.

OLKIN, I. & SAMPSON, A. (1998). Comparison of meta-analysis versus analysis of variance of individual patient data. *Biometrics* , 317–322.

QUAN, H., SUNDARARAJAN, V., HALFON, P., FONG, A., BURNAND, B., LUTHI, J.-C., SAUNDERS, L. D., BECK, C. A., FEASBY, T. E. & GHALI, W. A. (2005). Coding algorithms for defining comorbidities in icd-9-cm and icd-10 administrative data. *Medical care* , 1130–1139.

SHENKMAN, E., HURT, M., HOGAN, W., CARRASQUILLO, O., SMITH, S., BRICKMAN, A. & NELSON, D. (2018). Oneflorida clinical research consortium: linking a clinical and translational science institute with a community-based distributive medical education model. *Academic Medicine* **93**, 451.

SIDRANSKY, E., NALLS, M. A., AASLY, J. O. et al. (2009). Multicenter analysis of glucocerebrosidase mutations in parkinson's disease. *New England Journal of Medicine* **361**, 1651–1661.

SULLIVAN, M. D. (2018). Depression effects on long-term prescription opioid use, abuse, and addiction. *The Clinical journal of pain* **34**, 878–884.

TIAN, L. & GU, Q. (2016). Communication-efficient distributed sparse linear discriminant analysis. *arXiv preprint arXiv:1610.04798* .

VAN DER VAART, A. W. (2000). *Asymptotic statistics*, vol. 3. Cambridge university press.

WANG, J., KOLAR, M., SREBRO, N. & ZHANG, T. (2017). Efficient distributed learning with sparsity. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.

WANG, X., YANG, Z., CHEN, X. & LIU, W. (2019). Distributed inference for linear support vector machine. *Journal of Machine Learning Research* **20**, 1–41.

ZHANG, H., DENG, L., SCHIFFMAN, M., QIN, J. & YU, K. (2020). Generalized integration model for improved statistical inference by leveraging external summary data. *Biometrika* **107**, 689–703.

ZHAO, T., CHENG, G. & LIU, H. (2016). A partially linear framework for massive heterogeneous data. *Annals of Statistics* **44**, 1400–1437.

[*Received ???? 2018. Revised ???? 2018*]