An efficient distributed algorithm with application to COVID-19 data from heterogeneous clinical sites

Authors

Jiayi Tong, BS[1]
Chongliang Luo, PhD[1]
Md Nazmul Islam, PhD, MBA[2]
Natalie Sheils, PhD[2]
John Buresh, BS[2]
Mackenzie Edmondson, MS[1]
Peter A. Merkel, MD, MPH[1]
Ebbing Lautenbach, MD, MPH, MSCE[1]
Rui Duan, PhD[3]
Yong Chen, PhD[1]

Affiliation of the authors:

[1]Perelman School of Medicine, The University of Pennsylvania, Philadelphia, PA, USA
[2]UnitedHealth Group, Minnetonka, MN, USA
[3]School of Public Health, Harvard University, MA, USA

Correspondence author:

Yong Chen
University of Pennsylvania
Perelman School of Medicine
Blockley Hall 602, 423 Guardian Drive
Philadelphia, PA 19104
Office: 215-746-8155
E-mail: ychen123@upenn.edu

Word count: 3,844

ABSTRACT

Objectives: Integrating electronic health records (EHR) data from several clinical sites offers great opportunities to improve estimation with a more general population compared to analyses based on a single clinical site. However, sharing patient-level data across sites is practically challenging due to concerns about maintaining patient privacy. The objective of this study is to develop a novel distributed algorithm to integrate heterogeneous EHR data from multiple clinical sites without sharing patient-level data.

Materials and Methods: The proposed distributed algorithm for binary regression can effectively account for between-site heterogeneity and is communication-efficient. Our method is built on a pairwise likelihood function in the extended Mantel-Haenszel regression, which is known to be statistically highly efficient. We construct a surrogate pairwise likelihood function through approximating the target pairwise likelihood by its surrogate. We show that the proposed surrogate pairwise likelihood leads to a consistent and asymptotically normal estimator by effective communication without sharing individual patient-level data. We study the empirical performance of the proposed method through a systematic simulation study and an application with data of 14,215 COVID-19 patients from 230 clinical sites at UnitedHealth Group Clinical Research Database.

Results: The proposed method was shown to perform close to the gold standard approach under extensive simulation settings. When the event rate is <5%, the relative bias of the proposed estimator is 30% smaller than that of the meta-analysis estimator. The proposed method retained high accuracy across different sample sizes and event rates compared with meta-analysis. In the

data evaluation, the proposed estimate has a relative bias <9% when the event rate is <1%, whereas the meta-analysis estimate has a relative bias at least 10% higher than that of the proposed method.

Conclusions: Our simulation study and data application demonstrate that the proposed distributed algorithm provides an estimator that is robust to heterogeneity in event rates when effectively integrating data from multiple clinical sites. Our algorithm is therefore an effective alternative to both meta-analysis and existing distributed algorithms for modeling heterogeneous multi-site binary outcomes.

Keywords: distributed computing; heterogeneity; multi-site analysis; surrogate pairwise likelihood function.

INTRODUCTION

Electronic health records (EHR) data have become one of the most well-known data sources for medical and health research use. EHRs contain various elements of patient-level health information, including diagnoses, medications, procedures, imaging, and clinical notes [1-4]. Synthesis of this real-world evidence (RWE) from multiple clinical sites provides a larger sample size of the population compared to a single site study [5]. Analyses using larger populations benefit from better accuracy in estimation and prediction. Furthermore, the integration of research networks inside healthcare systems allows rapid translation and dissemination of research findings into evidence-based healthcare decision making to improve health outcomes, consistent with the idea of a learning health system [6-11].

In the past few years, several successful networks have been founded and become beneficial to multicenter research. One of them is the Observational Health Data Sciences and Informatics (OHDSI) consortium [12]. OHDSI was founded for the primary purpose of developing open-source tools that could be shared across multiple sites. OHDSI developed the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) for data standardization [13]. The OMOP allows each institution to transform the local EHR data to the CDM's standards. This procedure makes it feasible for the researchers to develop methods that can be simultaneously applied to the datasets from many institutions. The conversion and standardization of the data format decrease the probability of translation error and also increase the efficiency of data analysis. Another successful network is the National Pediatric Learning Health System (PEDSnet), a National Pediatric Learning Health System, within the PCORnet system [14,15]. This network contains eight large pediatric health systems in the US. Comprising clinical information from millions of children, PEDSnet offers the capacity to conduct multicenter pediatric research with

broad real-world evidence. The Sentinel System is another example of a multi-site network, a national electronic system for monitoring performance of FDA-regulated medical products [16].

In multi-center studies, maintaining privacy of patient data is a major challenge [17-19]. Due to data privacy policies, directly sharing patient-level data, especially demographic, comorbidity, and outcome data, is restricted and poorly feasible in practice. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) introduced a privacy rule to regulate use of protected health information (PHI) often found in EHRs, requiring de-identification of PHI before use in biomedical research [18]. De-identified PHI has been shown to be susceptible to re-identification, causing concern among patients [20,21].

In light of patient privacy concerns, many multicenter EHR-based studies currently conduct analyses by combining shareable summary statistics through meta-analysis [22-24]. While relatively simple to use, meta-analysis has been shown to result in biased or imprecise estimation in the context of rare outcomes, as well as with smaller sample sizes [25]. Other than meta-analysis, several distributed algorithms have been developed and considered in studies with multi-site data. In these distributed algorithms, a model estimation process is decomposed into smaller computational tasks that are distributed to each site. After parallel computation, intermediate results are transferred back to the coordinating center for final synthesis. Under this framework, there is no need to share patient-level data across sites. For example, GLORE (Grid Binary LOgistic Regression) was developed for conducting distributed logistic regressions [26], and WebDISCO (a Web service for distributed Cox model learning) was developed to fit the Cox proportional hazard model distributively and iteratively [27]. Both algorithms have been successfully deployed to the pSCANNER consortium [28]. Through iterative communication of aggregated information across the sites, these two algorithms provide accurate and lossless results,

which are equivalent to fitting a model on the pooled data from all sites. However, in practice these two methods can be time-consuming and communication-intensive due to the need for iteratively transferring data. To overcome this limitation, non-iterative privacy-preserving distributed algorithms for logistic regression (termed as ODAL) and Cox model (termed as ODAC) through the construction of a surrogate likelihood have been proposed [25,29,30].

However, all of the aforementioned distributed algorithms rely on the assumption that data across clinical sites are homogeneous. This assumption is often inaccurate in biomedical studies because it ignores the heterogeneity caused by the intrinsic differences across clinical sites or population characteristics. Ignoring heterogeneity across clinical sites can induce biases in estimating associations between the exposures of interest and outcomes [17,31]. Recently, a single Robust-ODAL algorithm was proposed to account for the heterogeneity across the clinical sites, but it only considers the limited situation when there exist a small number of outlying studies within the network [32].

One motivating example is the EHR data of 14,215 patients who were diagnosed with COVID-19 prior to June 29, 2020 from 230 sites in the UnitedHealth Group Clinical Research Database. There is a substantial difference in clinical practices across these sites due to such factors as geographical variability in disease patterns, variations in patients' characteristics, and regional differences in practice patterns. Therefore, developing methods to account for the heterogeneity in the data is especially needed when analyzing multi-site data within the networks.

To fill the above methodology gap, in this paper, we develop an effective privacy-preserving distributed data integration algorithm. We propose a distributed algorithm for binary regression to account for between-site heterogeneity by efficient communication (i.e., only requires one round of communication of aggregated information among the sites). Influenced by

the pairwise likelihood in the extended Mantel-Haenszel regression [31] and the idea proposed by Jordan et al. (2019) [33], we construct a surrogate pairwise likelihood through approximating the target pairwise likelihood by its surrogate. We show that the proposed surrogate pairwise likelihood leads to a consistent and asymptotically normal estimator, which is asymptotically equivalent to the maximum pairwise likelihood estimator based on the pooled data. This result is established based on U statistics, which is different from Jordan et al. (2019) [33]. We evaluate the empirical performance of the proposed method through simulation studies and apply the proposed method to investigate the associations between length of stay and the risk factors of interests. **Figure 1** shows the comparisons between the pooled analysis, meta-analysis method, iterative distributed algorithms, and the proposed method from various aspects. The proposed method can retain high estimation accuracy, protect patient privacy, handle heterogeneity, and save communication cost compared to the others.

[INSERT FIGURE 1 HERE]

MATERIALS AND METHODS

Surrogate Pairwise Likelihood

Suppose we have K different clinical sites. To keep the notation simple, we assume that each site has an equal number of n patients. Let $\{x_{ij}\}$ denote the collection of risk factors and $\{y_{ij}\}$ denote the independent response variable for the j-th patient in the i-th site where $j = 1, \ldots, n$ and $i = 1, \ldots, K$. The logistic regression model to characterize the association between the risk factors and the outcome is

$$logit\{Pr\,(y_{ij} = 1|x_{ij})\} = \alpha_i + \beta x_{ij} \quad (1)$$

where $logit(p) = log\,\{p/(1-p)\}$, α_i represents the site-specific prevalence of response variable, and β is the log odds ratio, meaning the association between risk factors x_{ij} and the outcome y_{ij}.

Following Liang (1987)'s extended Mantel-Haenszel regression, the pairwise likelihood can be constructed by conditioning $(y_{ij,}y_{il})$ on their order statistics. The pairwise likelihood for the i-th site can be written as

$$L_i(\beta) = \prod_{1\le j<l\le n}\left[1 + exp\,\{-\left(y_{ij} - y_{il}\right)\left(x_{ij} - x_{il}\right)^T\beta\}\right]^{-1} \quad (2)$$

We note that unlike generalized linear mixed effect model, where site-specific effects α_i's are assumed to follow a known distribution, the conditional pairwise likelihood eliminates the nuisance parameters $(\alpha_1, \ldots, \alpha_K)$ through the conditioning technique, hence avoids estimation of the nuisance parameters. Moreover, as studied in Liang (1987), the estimator, defined as the maximum of the pairwise likelihood, retains high statistical efficiency.

Now summing over all K sites, the overall likelihood function can be written as the product of L_i ,

$$L^*(\beta) = \prod_{i=1}^{K} L_i(\beta) = \prod_{i=1}^{K}\prod_{1\le j<l\le n}\left[1 + exp\,\{-\left(y_{ij} - y_{il}\right)\left(x_{ij} - x_{il}\right)^T\beta\}\right]^{-1} \quad (3)$$

which can be calculated if we have access to the patient-level data from all sites.

However, in practice, the individual patient-level data are only available at the local site and for the rest of the clinical sites in the network, we can only access aggregated information. Motivated by the surrogate likelihood in Jordan et al. (2019) which approximates the target pairwise likelihood by the likelihood from a single site, we propose a surrogate pairwise likelihood which can still handle the heterogeneity across the clinical sites.

For simplicity, we assume the first site as the local site, where we have access to the individual patient-level data. Let $l_1(\beta)$ denote the pairwise log-likelihood function for the local

site, $l_1(\beta) = \sum_{j<l}[1 + exp\{-(y_{1j} - y_{1l})(x_{1j} - x_{1l})^T\beta\}]^{-1}/\binom{n}{2}$. We construct the following surrogate log pairwise likelihood function $\tilde{l}_1(\beta)$ with the patient-level data from the local site, the initial value $\bar{\beta}$, and the aggregated information $\nabla l_i(\bar{\beta})$ and $\nabla^2 l_i(\bar{\beta})$. Specifically, we define

$$\tilde{l_1}(\beta) = l_1(\beta) + \{\nabla l(\bar{\beta}) - \nabla l_1(\bar{\beta})\}(\beta - \bar{\beta}) + \frac{(\beta-\bar{\beta})^T\{\nabla^2 l(\bar{\beta}) - \nabla^2 l_1(\bar{\beta})\}(\beta-\bar{\beta})}{2} \quad (4)$$

where $l_1(\beta)$ is the log pairwise likelihood function calculated from patient-level data in the local site; $\nabla^m l(\bar{\beta}) = K^{-1}\sum_{i=1}^{K}\nabla^m l_i(\bar{\beta}), m = 1, 2$ and $\{\nabla l_i(\bar{\beta})\}_{i=1,\dots K}$, $\{\nabla^2 l_i(\bar{\beta})\}_{i=1,\dots,K}$ are the first and second gradients of the surrogate pairwise likelihood function at $\bar{\beta}$ respectively. By maximizing the surrogate pairwise likelihood $\tilde{l}_1(\beta)$ we obtain the surrogate estimator $\tilde{\beta}$.

A natural choice of the initial value of $\bar{\beta}$ is the maximum likelihood estimator of the local site $l_1(\beta)$. Alternatively, since the performance of the surrogate estimator $\tilde{\beta}$ may depend on the choice of the initial values, we may use the inverse variance weighted average of the estimates from all sites, i.e.,

$$\bar{\beta} = (\sum_{i=1}^{K}\hat{V}_i^{-1})^{-1}\sum_{i=1}^{K}\hat{V}_i^{-1}\bar{\beta}_i \quad (5)$$

where $\bar{\beta}_i$ is the maximum of the pairwise likelihood and $\hat{V}_i = \hat{\Sigma}_{1,i}^{-1}(\bar{\beta}_i)\hat{\Sigma}_{2,i}(\bar{\beta}_i)\hat{\Sigma}_{1,i}^{-1}(\bar{\beta}_i)/\binom{n}{2}$ is the covariance matrix of $\bar{\beta}_i$ in the i-th site; $\hat{\Sigma}_{1,i}(\beta)$ and $\hat{\Sigma}_{2,i}(\beta)$ are functions of β for the i-th site. The definition of the covariance matrix, the asymptotic distribution of the surrogate estimator $\tilde{\beta}$, the derivation of the limiting distribution of $\tilde{\beta}$ are provided in Supplementary Appendix 1.

Algorithm

The proposed surrogate pairwise likelihood leads to the following algorithm.

Algorithm: Proposed method

Input: Patient-level data $\{x_{ij}\}$ and $\{y_{ij}\}$, where i denotes site index and j the observation index, where $i = 1, \dots, K$ and $j = 1, \dots, n$. Note that $\{x_{ij}\}$ and $\{y_{ij}\}$ are stored in the i-th site locally.

Output: Estimator $\tilde{\beta}$ of the association between $\{x_{ij}\}$ and $\{y_{ij}\}$.

Obtain $\bar{\beta}_i = arg\ max\ l_i(\beta)$ and $\hat{V}_i$ with patient-level data in the i-th site
Broadcast $\bar{\beta}_i$ and $\hat{V}_i$, and calculate initial value $\bar{\beta}$ with equation (5)
Suppose the 1st site is the local site that we have access to the individual patient-level data
Transfer $\bar{\beta}$ to the local site
for i in c(1:K) do
    Calculate $\nabla l_i(\bar{\beta})$, $\nabla^2 l_i(\bar{\beta})$
    Transfer the intermediate results to the local site
end for
Construct $\tilde{l}_1(\beta)$ as in equation (4) in the local site with $\bar{\beta}$, $\nabla l_i(\bar{\beta})$, $\nabla^2 l_i(\bar{\beta})$
Obtain $\tilde{\beta}$ by maximizing $\tilde{l}_1(\beta)$
Calculate variance of $\tilde{\beta}$ with equation (A.4) and (A.5) in Supplementary Appendix 1

In the following figure, we provide a graphical explanation to illustrate the implementation of the proposed algorithm.

[INSERT FIGURE 2 HERE]

REMARK 1: We implemented the proposed algorithm with R calling C programming language, which is a few dozen times faster than using R programming language only. Such implementation is necessary for the application of our algorithm to real-world settings where the number of patients in each site is relatively large.

REMARK 2: In the situation that each site is treated as the local site, each site can obtain its own surrogate pairwise likelihood estimate. These estimates can be further synthesized together with the inverse variance weighted average method to obtain an overall estimate.

Simulation Design

To evaluate the empirical performance of the proposed algorithm, we conduct a simulation study to cover a wide spectrum of practical settings. We set the total number of sites, K = 5 or 20 and sample size of each site is 1000 (i.e., the total numbers of patients are 5000 and 20,000 respectively.)

In our simulation study, we consider a setting where a binary outcome is associated with two risk factors, (x_1, x_2), where x_1 represents a continuous predictor (e.g., age) and x_2 is a binary predictor (e.g., sex, race). The binary outcome Y (e.g., presence/absence of hospitalization) is generated from a Bernoulli distribution, with the conditional probability specified by the following logistic regression model,

$$logit\{Pr\,(Y = 1|x)\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where β_1 and β_2 are the coefficients of x_1 and x_2 respectively, and β_0 is the intercept, characterizing the prevalence of the outcome Y. We set the true value of β_1 is 1 and of β_2 is -1. The distribution of x_1 for each study site is $Uni(-1, 1)$ to mimic the empirical distribution of variable "age", and x_2 is generated from a Bernoulli distribution with probability equal to 0.5 to mimic the empirical distribution of variable "sex".

We simulated three scenarios of the disease prevalence. The medians of the prevalence are 20%, 5%, and 0.5%. Specifically, the prevalence of the sites is randomly generated from a range of values as presented in **Figure 3**. We also simulated two scenarios of heterogeneity under each disease prevalence to mimic less heterogeneous cases (upper panel) and more heterogeneous (lower panel) cases, where the prevalence ranges are larger than those of the less heterogeneous cases.

[INSERT FIGURE 3 HERE]

Under each scenario, we compared the proposed method with the pairwise likelihood method (Liang, 1987), which can be treated as the gold standard and the commonly used meta-analysis. In the pairwise likelihood method, we assume that we have the access to all of the patient-level data. The simulation was conducted with 100 replications.

Data Evaluation

Our analytical dataset is composed of hospitalized patients who were diagnosed with COVID-19 prior to June 29, 2020 from a single large national health insurer, which covers a broad swath of the population. The data are from multiple EHR systems including EPIC [34], Cerner [35], and others. The data are recorded from $K = 230$ sites with $N = \sum_{i=1}^{K} n_i = 14{,}215$ insured (Commercial and Medicare) patients; see **Figure 4 (a)** for the details of inclusion-exclusion criteria and the distribution of COVID patients across the United States. Our objective is to develop an association model between clinical-and-demographic covariates (i.e., age, sex, line of business, and Charlson comorbidity index) and therapeutic patient outcomes. More details about the data quality are provided in Supplementary Appendix 2.

Outcomes are defined by combining both hospitalizations (days) and the status of patients being expired (i.e., a binary value taking value 1 if a patient is deceased or 0 otherwise). Consider three composite binary outcomes which take values 1 if the event occurs, and 0, otherwise. Here the events are defined as (a) LOS > 1 week and patient died, (b) LOS > 3 weeks and patient died, and (c) LOS > 4 weeks and patient died, respectively. **Figure 4 (b)** illustrates the prevalence rates of composite outcomes by 230 hospitals and **Figure 4 (c)** shows the number of COVID-19 hospitalizations included in the study across 47 states in the U.S. These two figures exhibit substantial variation in prevalence rates across sites. Moreover, patients admitted within the same

hospital are subject to somewhat similar care, administrative facilities, and treatments provided by the same physicians. This phenomenon leads us to treat sites as internally homogeneous and externally heterogeneous blocks. For details of the covariates, we refer to **Table 1**.

[INSERT FIGURE 4 HERE]

[INSERT TABLE 1 HERE]

RESULTS

Simulation studies results

For simplicity, we only present the results for the estimation of coefficient β_1 and the results for the other coefficients are similar. **Figure 3** shows the violin plot of the relative bias compared with the pairwise likelihood method under different numbers of sites and event rates. The first row in each panel is for the results when the total number sites, K = 5, and the second row is for K = 20. The black dashed line represents zero relative bias compared with the gold standard method. From the figure we observe that for all scenarios, the proposed method obtains smaller relative bias compared with meta-analysis. Importantly, as the event rate decreases under both less and more heterogeneous cases, the meta-analysis estimator is observed to have larger bias. When the event rate is <5%, the relative bias of the proposed estimator is 30% smaller than that of the meta-analysis estimator. In summary, the proposed method can provide better performance than the meta-analysis estimators to handle the heterogeneity across the clinical sites when the event is rare.

Data evaluation results

[INSERT FIGURE 5 HERE]

We primarily focus on estimating and comparing parameter estimates by the proposed method and the meta-analysis method. We stress that the parameter estimates need to be interpreted with caution since the effects' magnitudes or directions might be misleading without adjusting for potential confounders in the model. **Figure 5** illustrates the results obtained by the pairwise likelihood method (i.e., gold standard), the proposed method, and meta-analysis. As the prevalence rate decreases (i.e., in rare events), the proposed method outperforms meta-analysis in terms of estimating parameters. Specifically, the odds ratio (OR) of the proposed method remains closer to that of the gold standard approach, compared with the OR of meta-analysis. The proposed estimates have a relative bias <9% when the event rate is <1%, whereas the meta-analysis estimates have a relative bias at least 10% higher than that of the proposed method. This observation matches with that of the simulation study.

Besides, meta-analysis underestimates variance (or standard error of estimates) leading to far narrower confidence intervals relative to those of the gold standard method, especially for rare events. Ignoring between-and-within sites correlation in meta-analysis is likely to induce bias and underestimate uncertainty in parameter estimates leading to conflicting inference about the testing of significance of the effect size. For example, 95% confidence intervals of ORs for Charlson score based on meta-analysis does not contain OR value of one implying its significance, which is inconsistent with the inference based on the gold standard method. In contrast, the proposed method produces comparable inferences to the gold standard method.

DISCUSSION

In this paper, we proposed an effective privacy-preserving distributed algorithm for modeling binary outcomes while accounting for heterogeneity across clinical sites. Motivated by real-world

multicenter data, the proposed method requires transferring initial values and intermediate information instead of patient-level data. Our algorithm provides an estimator that is robust to heterogeneity in event rates. In simulations, the proposed method is shown to have higher accuracy than meta-analysis when the outcome is relatively rare, suggesting its utility in a rare-event context.

There are several advantages of our proposed algorithm compared to existing methods for privacy-preserving data analysis. Relative to meta-analysis, our method accesses patient data at a higher granularity while requiring minimal additional effort to institute. For multi-site studies operating under a common data model, such as OHDSI, analyses using our method can be carried out at individual sites concurrently without the need for any site-specific modifications. In addition, there are many benefits of using our method compared to existing distributed algorithms. First, compared to the iterative algorithms such as GLORE and WebDISCO [26,27], the proposed algorithm does not require iterative communication across the sites, leading to the reduction in communication costs and administrative efforts. Secondly, to implement the proposed method, the patient-level data are only required in one single site. For the other sites within the network, the aggregated information will be used instead of patient-level data transfer across the sites to construct the surrogate pairwise likelihood function. Given the understandable privacy- and proprietary-related sensitivities health systems have to provide "outside" collaborators with access to patient-level data, limiting the need to use such data to only one site would be extremely beneficial to a multi-site project in terms of feasibility, costs, and time. Thirdly, by canceling out the baseline probability function, the proposed method can handle the heterogeneity in the event rate between the sites. In addition, the proof of the proposed method is established based on U statistics and is different from that of Jordan et al. (2019) [33]. In terms of computational

complexity, the implementation of the proposed method is slow when the sample size is large compared with the traditional regression model. We thus implemented the algorithm with R calling C, which is a few dozen times faster than using the R programming language only. The R and C code will be made available at https://github.com/Penncil and through our R package: 'pda'.

To investigate the proposed non-iterative distributed algorithm, we can extend it in several aspects. First, the proposed pairwise likelihood function can only handle the heterogeneity of the intercepts in the regression model. To handle the other types of heterogeneity (e.g., heterogeneous effects of the predictors), more robust algorithms should be developed. Secondly, we are going to develop methods for other types of outcomes, such as continuous and time-to-event data. In addition, the development of distributed algorithms to handle the missingness in the longitudinal EHR is needed in the future. Lastly, we have been working on the development of the open-source software R package to implement the proposed distributed algorithm within a multicenter network. We believe that the proposed algorithm would be a robust method to account for the heterogeneity across multiple clinical sites, leading to a better data integration framework inside health systems.

CONCLUSION

The proposed distributed algorithm provides an estimator that is robust to heterogeneity in event rates when effectively integrating data from multiple clinical sites. Through a simulation study and a real-world use case using data from the UnitedHealth Group Clinical Research Database, the proposed method is shown to be an effective alternative to both meta-analysis and existing distributed algorithms for modeling heterogeneous multi-site binary outcomes.

FUNDING

This work was supported in part by the National Institutes of Health grants 1R01LM012607 and 1R01AI130460.

AUTHOR CONTRIBUTIONS

JT and YC designed methods and experiments; MI, NS, and JN provided the dataset from UnitedHealth Group Clinical Research Database for data analysis; CL, RD and YC guided the theoretical development and dataset generation for the simulation study; JT generated the simulation datasets, conducted simulation experiments; MI, NS and JB conducted data analysis of the EHR data from the UnitedHealth Group; all authors interpreted the results and provided instructive comments; JT drafted the main manuscript. All authors have approved the manuscript.

CONFLICT OF INTEREST STATEMENT

The authors have no competing interests to declare.

REFERENCES

1 Sherman RE, Anderson SA, Dal Pan GJ, *et al.* Real-world evidence—what is it and what can it tell us. *N Engl J Med* 2016;**375**:2293–7.

2 Food US, Administration D, Others. Use of real-world evidence to support regulatory decision-making for medical devices: guidance for industry and Food and Drug Administration staff. *Silver Spring, MD: US Food and Drug Administration* 2017.

3 Food US, Administration D, Others. Use of electronic health record data in clinical investigations: guidance for industry. *Silver Spring, MD: US Department of Health and Human Services* 2018.

4 Center for Drug Evaluation and Research. Submitting Documents Using Real-World Data and Real-World Evidence. *FDA Med Bull*www.fda.gov/regulatory-information/search-fda-guidance-documents/submitting-documents-using-real-world-data-and-real-world-evidence-fda-drugs-and-biologics-guidance

5 Bowens FM, Frye PA, Jones WA. Health information technology: integration of clinical workflow into meaningful use of electronic health records. *Perspect Health Inf Manag* 2010;**7**:1d.

6 Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. *Sci Transl Med* 2010;**2**:57cm29.

7 Weng C, Appelbaum P, Hripcsak G, *et al.* Using EHRs to integrate research with patient care: promises and challenges. *J Am Med Inform Assoc* 2012;**19**:684–7.

8 Greene SM, Reid RJ, Larson EB. Implementing the learning health system: from concept to action. *Ann Intern Med* 2012;**157**:207–10.

9 Smoyer WE, Embi PJ, Moffatt-Bruce S. Creating Local Learning Health Systems: Think Globally, Act Locally. *JAMA* 2016;**316**:2481–2.

10 Maro JC, Platt R, Holmes JH, *et al.* Design of a national distributed health data network. *Ann Intern Med* 2009;**151**:341–4.

11 Brown JS, Holmes JH, Shah K, *et al.* Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Med Care* 2010;**48**:S45–51.

12 Hripcsak G, Duke JD, Shah NH, *et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform* 2015;**216**:574–8.

13 Overhage JM, Ryan PB, Reich CG, *et al.* Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc* 2012;**19**:54–60.

14 Forrest CB, Margolis PA, Bailey LC, *et al.* PEDSnet: a National Pediatric Learning Health System. *J Am Med Inform Assoc* 2014;**21**:602–6.

15 Fleurence RL, Curtis LH, Califf RM, *et al.* Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc* 2014;**21**:578–82.

16 Platt R, Carnahan RM, Brown JS, *et al.* The US Food and Drug Administration's Mini-Sentinel program: status and direction. *Pharmacoepidemiol Drug Saf* 2012;**21**:1–8.

17 Wu H-DI. Effect of Ignoring Heterogeneity in Hazards Regression. In: Balakrishnan N, Nikulin MS, Mesbah M, *et al.*, eds. *Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life*. Boston, MA: : Birkhäuser Boston 2004. 239–50.

18 Arellano AM, Dai W, Wang S, *et al.* Privacy Policy and Technology in Biomedical Data Science. *Annu Rev Biomed Data Sci* 2018;**1**:115–29.

19 Loukides G, Denny JC, Malin B. The disclosure of diagnosis codes can breach research participants' privacy. *J Am Med Inform Assoc* 2010;**17**:322–7.

20 Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. *J Am Med Inform Assoc* 2010;**17**:169–77.

21 McGraw D. Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data. *J Am Med Inform Assoc* 2013;**20**:29–34.

22 Vashisht R, Jung K, Schuler A, *et al.* Association of Hemoglobin A1c Levels With Use of Sulfonylureas, Dipeptidyl Peptidase 4 Inhibitors, and Thiazolidinediones in Patients With Type 2 Diabetes Treated With Metformin. JAMA Network Open. 2018;**1**:e181755. doi:10.1001/jamanetworkopen.2018.1755

23 Hripcsak G, Ryan PB, Duke JD, *et al.* Characterizing treatment pathways at scale using the OHDSI network. *Proc Natl Acad Sci U S A* 2016;**113**:7329–36.

24 Boland MR, Parhi P, Li L, *et al.* Uncovering exposures responsible for birth season--disease effects: a global study. *J Am Med Inform Assoc* 2018;**25**:275–88.

25 Duan, R., Luo, C., Schuemie, M. H., Tong, J., Liang, J. C., Chang, H. H., Boland, M. R., Bian, J., Xu, H., Holmes, J. H.. Learning from local to global-an efficient distributed algorithm for modeling time-to-event data. *Journal of the American Medical Informatics Association*. 2020 July; 27(7):1028–1036.

26 Wu Y, Jiang X, Kim J, *et al.* G rid Binary LO gistic RE gression (GLORE): building shared models without sharing data. *J Am Med Inform Assoc* 2012;**19**:758–64.

27 Lu C-L, Wang S, Ji Z, *et al.* WebDISCO: a web service for distributed cox model learning without patient-level data sharing. *J Am Med Inform Assoc* 2015;**22**:1212–9.

28 Ohno-Machado L, Agha Z, Bell DS, *et al.* pSCANNER: patient-centered Scalable National Network for Effectiveness Research. *J Am Med Inform Assoc* 2014;**21**:621–6.

29 Duan R, Boland MR, Moore JH, Chen Y. ODAL: a one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. In: Altman RB, Dunker AK, Hunter L, Ritchie MD, Murray T, Klein TE, eds. *Pacific Symposium on Biocomputing* 2019. Singapore: World Scientific; 30–41.

30 Duan R, Boland MR, Liu Z, Liu Y, Chang HH, Xu H, Chu H, Schmid CH, Forrest CB, Holmes JH, Schuemie MJ, Chen Y. Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *Journal of the American Medical Informatics Association*. 2020 Mar;27(3):376-85.

31 Liang KY. Extended Mantel-Haenszel estimating procedure for multivariate logistic regression models. *Biometrics* 1987;**43**:289–99.

32 Tong J, Duan R, Li R, Schmid CH, Moore JH, Chen Y. Robust-ODAL: Learning from heterogeneous health systems without sharing patient-level data. *Pacific Symposium on Biocomputing* 2020;**25**:695–706.

33 Jordan MI, Lee JD, Yang Y. Communication-Efficient Distributed Statistical Inference. Journal of the American Statistical Association. 2019;**114**:668–81. doi:10.1080/01621459.2018.1429274

34 Epic. https://www.epic.com (accessed 15 Oct 2020).

35 Home. https://www.cerner.com (accessed 15 Oct 2020).

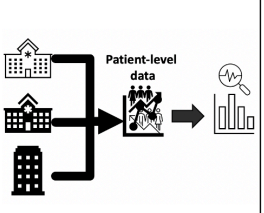
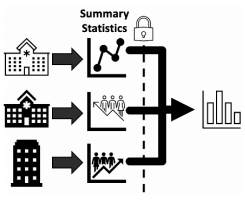
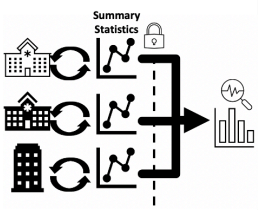
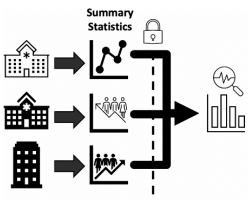
	Pooled Analysis	Meta Analysis	Iterative Distributed Algorithms	The Proposed Method
Workflow				
Accuracy	High	Varying (Not accurate for rare diseases)	High (lossless)	High
Protect privacy	NO	YES	YES	YES
Heterogeneity aware	YES	YES	NO	YES
Communication efficient	NO	YES	NO	YES

Figure 1: Comparisons between pooled analysis, meta-analysis, iterative distributed algorithms, and the proposed method. The proposed method can retain high accuracy when estimating association between exposures and outcome of interest. In addition, the proposed method can handle heterogeneity across the sites and protect patient privacy with efficient communication.

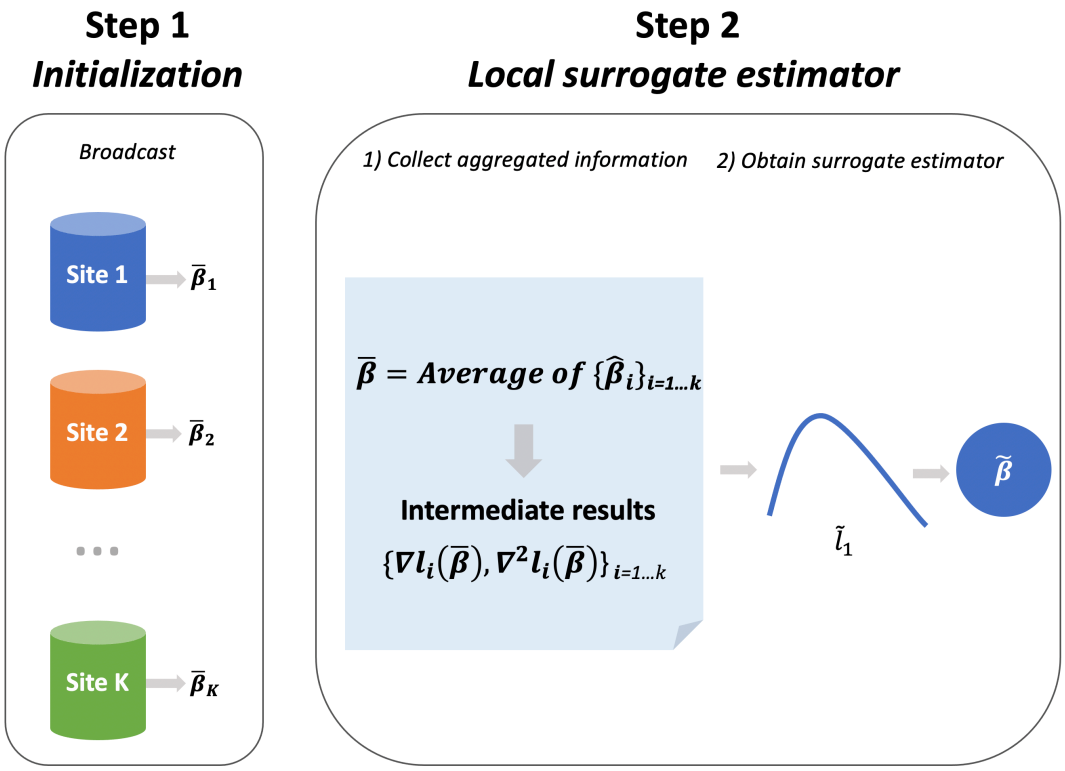


Figure 2: Illustration of the proposed method. Step I: Using data from each local site to estimate $\bar{\beta}_i$, where $i = 1, \ldots, K$ and broadcast the values to calculate the weighted initial value $\bar{\beta}$. Step II: With $\bar{\beta}$, calculating the intermediate terms $\nabla l_i(\bar{\beta})$, $\nabla^2 l_i(\bar{\beta})$ at each site and then transfer the results back to the local site. With the intermediate results and $\bar{\beta}$ to construct the surrogate pairwise log-likelihood function $\tilde{l}_1(\beta)$ in the local site. Maximizing $\tilde{l}_1(\beta)$ to obtain the estimator $\tilde{\beta}$.

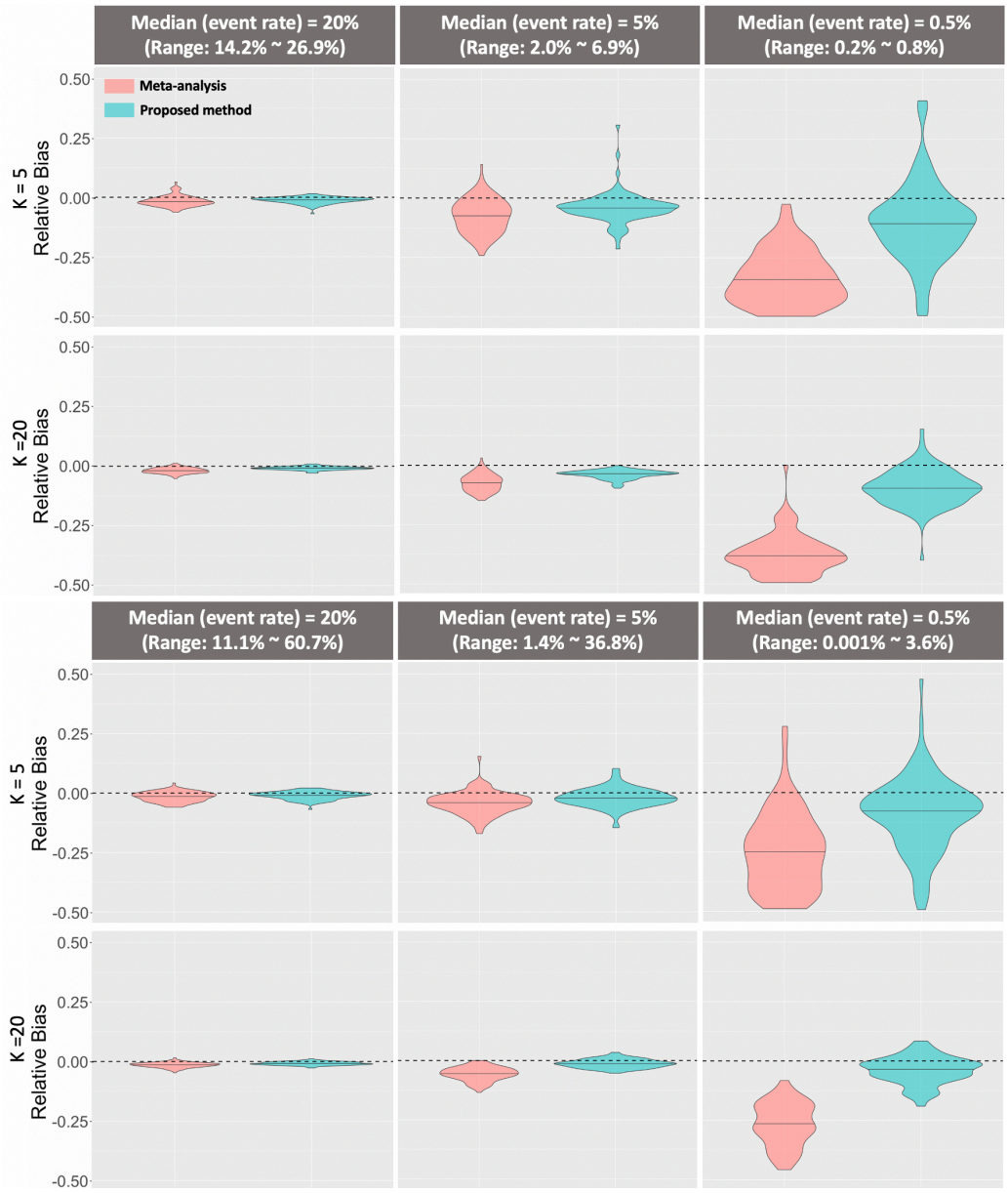


Figure 3: Upper panel: relative bias of β_1 estimation compared with the pairwise likelihood method under three scenarios with median prevalence 20%, 5%, and 0.5% when the total number of sites is 5 or 20 (i.e., K = 5 or 20). **Lower panel**: relative bias of β_1 estimation compared with the pairwise likelihood method under three scenarios with median prevalence 20%, 5%, and 0.5% with larger heterogeneity (i.e., larger prevalence range than upper panel) when the total number of sites is 5 or 20 (i.e., K = 5 or 20).

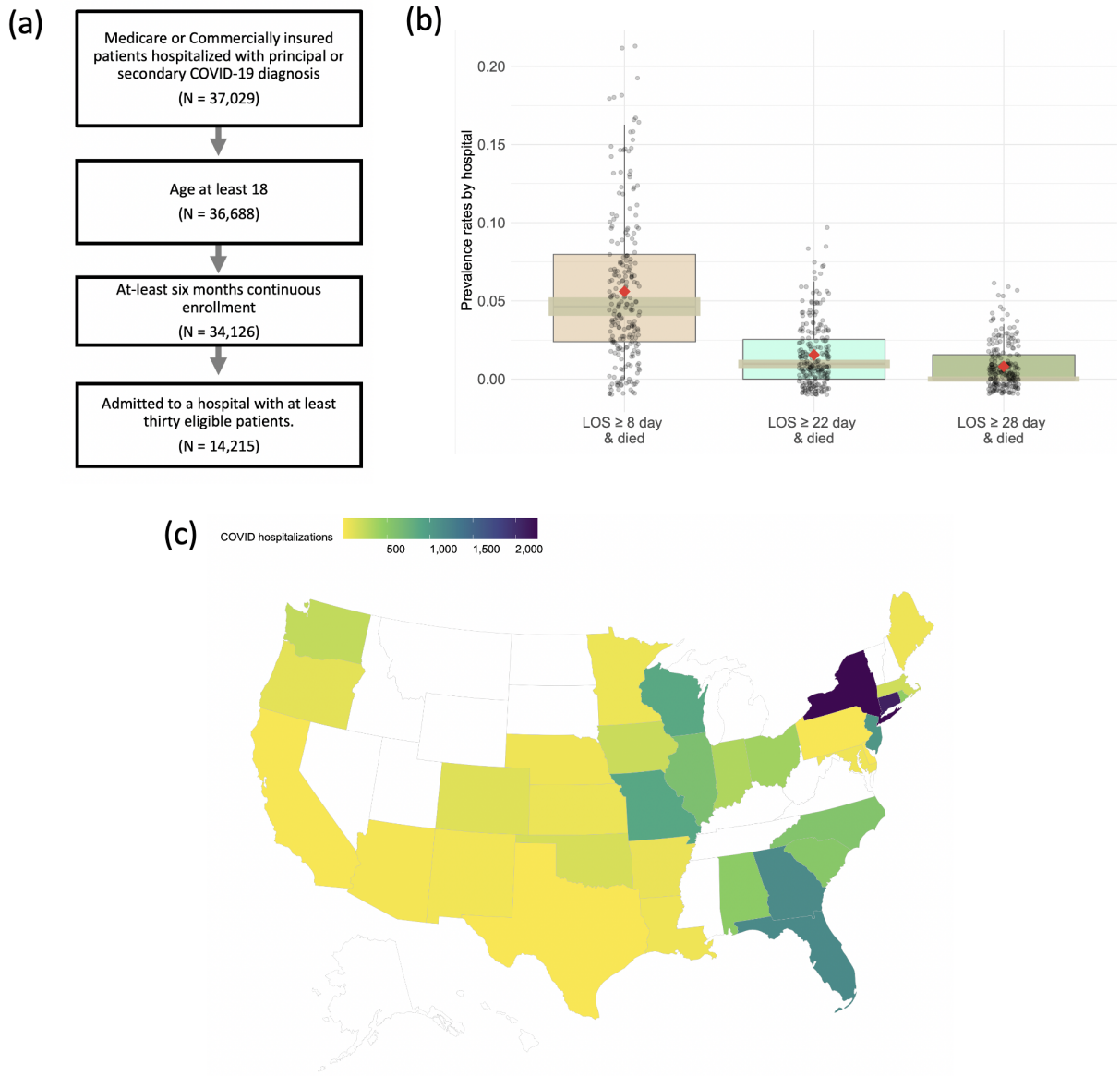


Figure 4: (a) Diagram of the patient inclusion-exclusion criteria. (b) Box plots of the prevalence rates of composite outcomes of 230 hospitals. (c) COVID-19 cases distribution: number of COVID-19 hospitalizations included in the study are represented across 47 states.

Number of patients		14,215
Number of hospitals		230
Patient Level Characteristics		
	Mean Age in years (Median, SD)	71.1 (73, 14.3)
	Sex	
	Male (%)	6,925 (48.7%)
	Female (%)	7,290 (51.3%)
	Mean Charlson Score (Median, SD)	3.4 (3.0, 3.0)
	Insurance Type	
	Medicare Advantage (%)	11,460 (80.6%)
	Commercial (%)	2,755 (19.4%)
Patient Outcomes		
	Mean Length of Stay in days (Median, SD)	10.2 (6, 12.6)
	Length of Stay ≥ 1 day and Died (%)	1,716 (12.7%)
	Length of Stay ≥ 8 day and Died (%)	843 (5.9%)
	Length of Stay ≥ 15 day and Died (%)	436 (3.1%)
	Length of Stay ≥ 22 day and Died (%)	234 (1.6%)
	Length of Stay ≥ 29 day and Died (%)	124 (0.9%)

Table 1. Summary characteristics of the 14,215 patients from 230 hospitals in our population.

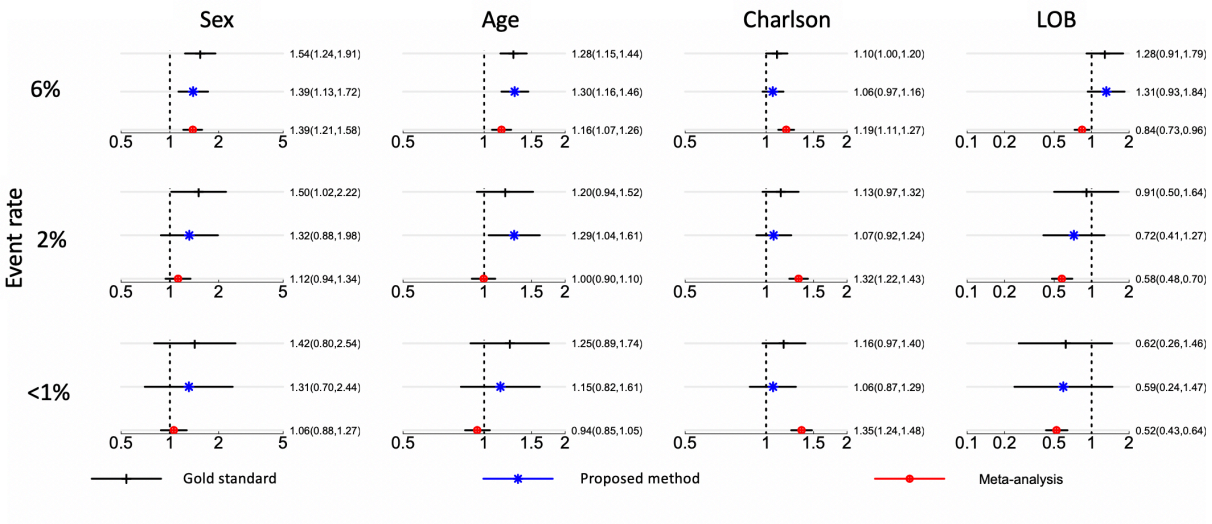


Figure 5: Point estimates and 95% confidence intervals (CI) for the association (in odds ratio scale) between the LOS (i.e., length of stay) and covariates (i.e., sex, age, Charlson score, line of business, from left to right). Each row represents an event rate of the outcome: 6%, 2%, and <1% from top to bottom. Each column represents the estimation of the covariate.