




## Research and Applications

# Learning from local to global: An efficient distributed algorithm for modeling time-to-event data

Rui Duan,<sup>1,†</sup> Chongliang Luo,<sup>1,†</sup> Martijn J. Schuemie ,<sup>2,†</sup> Jiayi Tong,<sup>1</sup> C. Jason Liang,<sup>3</sup> Howard H. Chang,<sup>4</sup> Mary Regina Boland ,<sup>1</sup> Jiang Bian,<sup>5,6</sup> Hua Xu ,<sup>7</sup> John H. Holmes,<sup>1</sup> Christopher B. Forrest,<sup>8</sup> Sally C. Morton,<sup>9</sup> Jesse A. Berlin,<sup>10</sup> Jason H. Moore,<sup>1</sup> Kevin B. Mahoney,<sup>11</sup> and Yong Chen<sup>1</sup>

<sup>1</sup>Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, USA, <sup>2</sup>Janssen Research and Development LLC, Titusville, New Jersey, USA, <sup>3</sup>Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland, USA, <sup>4</sup>Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, USA, <sup>5</sup>Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, Florida, USA, <sup>6</sup>Cancer Informatics and eHealth Core, University of Florida Health Cancer Center, Gainesville, Florida, USA, <sup>7</sup>School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA, <sup>8</sup>Applied Clinical Research Center, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA, <sup>9</sup>Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA, <sup>10</sup>Johnson & Johnson, Titusville, NJ, USA, and <sup>11</sup>University of Pennsylvania Health System, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

<sup>†</sup>These authors contributed equally.

Corresponding Author: Yong Chen, PhD, Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania School of Medicine, 423 Guardian Drive, Philadelphia, PA 19104; ychen123@upenn.edu

Received 23 October 2019; Revised 27 February 2020; Editorial Decision 25 March 2020; Accepted 28 March 2020

## ABSTRACT

**Objective:** We developed and evaluated a privacy-preserving One-shot Distributed Algorithm to fit a multicenter Cox proportional hazards model (ODAC) without sharing patient-level information across sites.

**Materials and Methods:** Using patient-level data from a single site combined with only aggregated information from other sites, we constructed a surrogate likelihood function, approximating the Cox partial likelihood function obtained using patient-level data from all sites. By maximizing the surrogate likelihood function, each site obtained a local estimate of the model parameter, and the ODAC estimator was constructed as a weighted average of all the local estimates. We evaluated the performance of ODAC with (1) a simulation study and (2) a real-world use case study using 4 datasets from the Observational Health Data Sciences and Informatics network.

**Results:** On the one hand, our simulation study showed that ODAC provided estimates nearly the same as the estimator obtained by analyzing, in a single dataset, the combined patient-level data from all sites (ie, the pooled estimator). The relative bias was <0.1% across all scenarios. The accuracy of ODAC remained high across different sample sizes and event rates. On the other hand, the meta-analysis estimator, which was obtained by the inverse variance weighted average of the site-specific estimates, had substantial bias when the event rate is <5%, with the relative bias reaching 20% when the event rate is 1%. In the Observational Health Data Sciences and Informatics network application, the ODAC estimates have a relative bias <5% for 15 out of 16 log hazard ratios, whereas the meta-analysis estimates had substantially higher bias than ODAC.

**Conclusions:** ODAC is a privacy-preserving and noniterative method for implementing time-to-event analyses across multiple sites. It provides estimates on par with the pooled estimator and substantially outperforms the meta-analysis estimator when the event is uncommon, making it extremely suitable for studying rare events and diseases in a distributed manner.

**Key words:** Cox proportional hazards model, data integration, distributed algorithm, electronic health record, meta-analysis

## INTRODUCTION

Real-world data (RWD) such as electronic health records (EHRs) and health plan claims are playing an increasing role in generating real-world evidence to support healthcare decision making.<sup>1</sup> Patient data (eg, diagnoses, medications, labs, and clinical notes) are routinely collected and entered into the EHRs during clinical care delivery. When compared with cross-sectional observational data, the detailed longitudinal information contained in the EHR enables time-to-event modeling, also known as survival analysis. Incorporating time into statistical models that predict occurrence of outcomes enables a better understanding of the predictors of when an event occurs rather than merely whether it occurred.<sup>2–4</sup> The Cox proportional hazards model (hereafter referred to as the Cox model) is one of the most commonly used time-to-event models, and has been widely applied in EHR-based studies for treatment evaluation, risk factor identification, and individual risk prediction.<sup>5</sup>

The last few years have witnessed an increasing number of clinical research networks, curating and using immense collections of health system EHRs and health plan claims data. Two prominent examples are (1) the Observational Health Data Sciences and Informatics (OHDSI) network—an international network of observational health databases that cover more than half a billion patient records in a common data model (CDM) and (2) the national Patient-Centered Clinical Research Network (PCORnet)—a network covering more than 100 million patients in the United States.<sup>6,7</sup> These large data consortia strive to provide platforms and tools to integrate heterogeneous RWD from a diverse range of healthcare organizations. Multicenter analyses using RWD from these clinical research networks have been increasing rapidly in large part because they can improve the generalizability of the study results by increasing the sample size.

Despite the benefits of multicenter analyses, direct sharing of patient-level data across institutions may be challenging because of concerns related to patient and institutional privacy. Studies that include institutional data from multiple countries face legal and regulatory barriers to sharing patient-level data. The OHDSI network, which is an international consortium, uses a federated model in which the patient-level data are stored at local institutions and only aggregated information are shared. Thus, multicenter analyses have to be based on the summary statistics obtained from each site.<sup>8–11</sup> The results are often combined through meta-analysis, specifically, some form of weighted average where the weight for each estimate, eg, could be the inverse of the site-specific variance (for a fixed-effect model). However, when the outcomes or exposures are rare, or some of the sites may have small sample size, the accuracy (in terms of bias and precision) of the meta-analysis may be poor, as will be shown later in both the simulation studies and with real-world examples.

Distributed computing has been considered in many applications,<sup>12–17</sup> in which the model estimation is decomposed into smaller computational tasks that are computed locally at each site and then returned to the site instituting the study. For example, a distributed algorithm for conducting logistic regression, GLORE (Grid Binary Logistic Regression),<sup>13</sup> and the WebDISCO (a web service for distributed Cox model learning) for fitting the Cox model<sup>14</sup> were developed and successfully implemented in the pSCANNER (patient-centered Scalable National Network for Effectiveness Research) consortium.<sup>18</sup> Through multiple rounds of iterative communication of in-

formation across sites, these algorithms can be lossless (ie, the results are equivalent to the results from fitting the model on the dataset created by pooling the individual-level data across all sites). However, owing to the need to iteratively transfer data across sites, it is time-consuming and labor-intensive in practice. Thus, development of non-iterative distributed algorithms, which require neither data sharing nor iterative communication back and forth between sites, is an active research. For example, Chen et al<sup>12</sup> developed a lossless noniterative distributed algorithm for linear regression. Duan et al<sup>17</sup> proposed a noniterative distributed algorithm for logistic regression through the construction of a surrogate likelihood.

In this article, we propose the One-shot Distributed Algorithm for Cox model (ODAC). A unique challenge in developing distributed algorithms for the Cox model, compared with our earlier work on logistic regression,<sup>17</sup> is that the log partial likelihood function is not equivalent to the summation of local log partial likelihood functions from each site. Each component of the log partial likelihood function involves a nonlinear function of data from all patients who are at risk at a certain time point. By designing a novel initialization step, our algorithm extends the surrogate likelihood approach<sup>16</sup> to deal with the more complicated Cox partial likelihood function, while maintaining the property of only requiring communication of aggregated information between sites. We show that our proposed algorithm is noniterative while achieving high accuracy (small bias) in both a simulation study and a real-world use case that studies the risk factors of acute myocardial infarction (AMI) and stroke using claims data from 4 different databases in the OHDSI network.

## MATERIALS AND METHODS

### Cox model

We first introduce the notation and basic assumptions of the Cox model. Let  $X$  be a vector denoting  $p$  risk factors and let  $T$  be the time-to-event for the outcome of interest. The Cox model assumes the hazard at time  $t$  follows

$$\lambda(t|X) = \lambda_0(t)\exp(\beta^T X),$$

where  $\lambda_0(t)$  is the baseline hazard function, and  $\beta$  is the vector of regression coefficients which are interpreted as the log hazard ratios (HRs). The observed data are  $\{T_i, \delta_i, x_i\}$  for the  $i$ -th subject, with  $\delta_i = 0$  indicating censoring and  $\delta_i = 1$  indicating an event,  $i = 1, \dots, N$ . For a given time  $t$ , we denote  $R(t) = \{i; T_i \geq t\}$ , which is the risk set at time  $t$ . The log partial likelihood function is constructed as

$$L(\beta) = \frac{1}{n} \sum_{i=1}^n \delta_i \log \frac{\exp(\beta^T x_i)}{\sum_{j \in R(T_i)} \exp(\beta^T x_j)},$$

which does not require estimation of the baseline hazard  $\lambda_0(t)$ . The parameter  $\beta$  is estimated by the value that maximizes the log partial likelihood function.<sup>5</sup>

### Parameter estimation in a distributed network via ODAC

Now suppose that we have data stored in  $K$  different clinical sites and denote  $n_j$  to be the sample size of the data at the  $j$ -th site. Denote the total sample size to be  $N = \sum_{j=1}^K n_j$ . For the  $i$ -th patient at the  $j$ -th site, we observe  $\{T_{ij}, \delta_{ij}, x_{ij}\}$ . The risk set in site  $j$  at time  $t$  is

defined as  $R_j(t) = \{i; T_{ij} \geq t\}$ , which contains all the subjects in site  $j$  that have not experienced an event or been censored at time  $t$ . The combined risk set over all the  $K$  sites is  $R(t) = \{(i, j); T_{ij} \geq t\}$ . With a slight abuse of notation, we further denote  $R_{ij}$  to be the risk set at time  $T_{ij}$ , ie,  $R_{ij} = R(T_{ij})$ . Denote the set of event times at the  $j$ -th site to be  $T_j = \{T_{ij} : \delta_{ij} = 1\}$ , and the total set of unique event time to be  $\mathcal{T} = \cup_j T_j$ . Assume there are in total  $d$  unique event time points  $t_1, \dots, t_d \in \mathcal{T}$ . Ideally, if all the data could be pooled together, the combined partial likelihood function could be expressed as

$$L(\beta) = \frac{1}{N} \sum_{j=1}^K \sum_{i=1}^{n_j} \delta_{ij} \log \frac{\exp(\beta^T x_{ij})}{\sum_{(s,k) \in R_{ij}} \exp(\beta^T x_{sk})} \quad (1)$$

and a pooled estimator  $\hat{\beta}$  could be obtained by maximizing the combined partial likelihood function. However, in reality, transferring patient-level data across sites is almost always impossible due to patient privacy concerns; thus, each site can only access their own data.

Inspired by the surrogate likelihood approach developed in Jordan et al<sup>16</sup> and Duan et al,<sup>17</sup> we aim to construct a proxy of the combined partial likelihood function, which we call a surrogate likelihood function. The idea is to construct a function that is close to the combined partial likelihood function defined in equation 1 near a neighborhood of the true parameter value. One naive choice of the surrogate likelihood function (eg, at the  $j$ -th site) is the local partial likelihood function  $L_j(\beta)$  constructed as

$$L_j(\beta) = \frac{1}{n_j} \sum_{i=1}^{n_j} \delta_{ij} \log \frac{\exp(\beta^T x_{ij})}{\sum_{s \in R_j(T_{ij})} \exp(\beta^T x_{sj})} \quad (2)$$

which only utilizes the patient-level data at the  $j$ -th site. However, this approximation does not utilize any information from other sites. Our goal is to further improve the accuracy of the approximation by borrowing aggregated information from other sites.

Define  $\tilde{\beta}$  to be an initial value that is in a neighborhood of the true value of the parameter  $\beta$ . We propose the surrogate likelihood function obtained at site  $j$  to be

$$\tilde{L}_j(\beta) = L_j(\beta) + \langle \nabla L(\tilde{\beta}) - \nabla L_j(\tilde{\beta}), \beta \rangle + \frac{1}{2} (\beta - \tilde{\beta})^T \{ \nabla^2 L(\tilde{\beta}) - \nabla^2 L_j(\tilde{\beta}) \} (\beta - \tilde{\beta}) \quad (3)$$

for  $j = 1, \dots, K$ , where  $L_j(\beta)$  is the local likelihood function defined in equation (2),  $\nabla$  and  $\nabla^2$  denote the first and second order gradients of a function (explicit forms of  $\nabla L_j(\tilde{\beta})$ ,  $\nabla L(\tilde{\beta})$ ,  $\nabla^2 L_j(\tilde{\beta})$  and  $\nabla^2 L(\tilde{\beta})$  can be found in the [Supplementary Appendix](#)). Intuitively, the surrogate likelihood function in equation (3) utilizes the local likelihood function  $L_j(\beta)$  as a baseline function, and it also adds a first-order term and a second-order term,  $\langle \nabla L(\tilde{\beta}) - \nabla L_j(\tilde{\beta}), \beta \rangle$  and  $\frac{1}{2} (\beta - \tilde{\beta})^T \{ \nabla^2 L(\tilde{\beta}) - \nabla^2 L_j(\tilde{\beta}) \} (\beta - \tilde{\beta})$ , to alter the shape of the local likelihood function around  $\tilde{\beta}$  toward the combined partial likelihood function.

In the construction of the surrogate likelihood function  $\tilde{L}_j(\beta)$ , once  $\tilde{\beta}$  is obtained, the terms  $L_j(\beta)$ ,  $\nabla L_j(\tilde{\beta})$  and  $\nabla^2 L_j(\tilde{\beta})$  can be calculated at the  $j$ -th site without needing to transfer information. And the constructing components  $\nabla L(\tilde{\beta})$  and  $\nabla^2 L(\tilde{\beta})$  can be calculated distributively from all sites with the help of some preliminary summary-level statistics  $U_j(\mathcal{T}) = (U_j(t_1), \dots, U_j(t_d))$ ,  $W_j(\mathcal{T}) = (W_j(t_1), \dots, W_j(t_d))$  and  $Z_j(\mathcal{T}) = (Z_j(t_1), \dots, Z_j(t_d))$ , where for each time point  $t$  in  $\mathcal{T}$ ,  $U_k(t) = \sum_{i \in R_k(t)} \exp(\tilde{\beta}^T x_{ik})$ ,  $W_k(t) = \sum_{i \in R_k(t)} \exp(\tilde{\beta}^T x_{ik}) x_{ik}$ , and  $Z_k(t) = \sum_{i \in R_k(t)} \exp(\tilde{\beta}^T x_{ik}) x_{ik} x_{ik}^T$ . The detailed

steps of distributively calculating  $\nabla L(\tilde{\beta})$  and  $\nabla^2 L(\tilde{\beta})$  can be found in the [Supplementary Appendix](#). Importantly, these gradients are all aggregated such that patient privacy is protected.

Regarding the initial value  $\tilde{\beta}$ , we recommend using the inverse variance weighted average of the estimates obtained by fitting a Cox model at each site, that is,

$$\tilde{\beta} = \left( \sum_{j=1}^K \hat{V}_j^{-1} \right)^{-1} \sum_{j=1}^K \hat{V}_j^{-1} \hat{\beta}_j \quad (4)$$

where  $\hat{\beta}_j = \text{argmax}_{\beta} L_j(\beta)$  is the estimator from of Cox model fitted on data at the  $j$ -th site, and  $\hat{V}_j$  is the estimated variance of  $\hat{\beta}_j$ . After constructing the surrogate likelihood function, we can obtain a surrogate likelihood estimator at each site by

$$\tilde{\beta}_j = \text{argmax}_{\beta} \tilde{L}_j(\beta). \quad (5)$$

The variance of  $\tilde{\beta}_j$  is defined to be  $\tilde{V}_j$ , and it can be calculated using equation 5.5 in the [Supplementary Appendix](#). Finally, we combine all the local surrogate estimator  $\tilde{\beta}_j$  using inverse variance weighted averaging in the same format as in equation 5. We summarize the ODAC algorithm in the following and also provide a schematic illustration in [Figure 1](#).

#### ODAC algorithm

##### (1) Initialization

In site  $k = 1$  to  $K$ ,

**do**

Fit a Cox model and obtain the local estimate  $\hat{\beta}_k$  and the variance estimate  $\hat{V}_k$ ; **broadcast**  $\hat{\beta}_k$ ,  $\hat{V}_k$ , and the set of unique event time points in site  $k$ .

**end**

##### (2) Local surrogate estimator

In Site  $k = 1$  to  $K$ ,

**do**

obtain  $\tilde{\beta}$  using (4), and all the unique event time points across all sites  $t_1 \dots t_d$ ; calculate and broadcast the intermediate summary-level statistics  $U_j(\mathcal{T})$ ,  $W_j(\mathcal{T})$  and  $Z_j(\mathcal{T})$ ; construct the surrogate likelihood  $\tilde{L}_k(\beta)$  by (3) treating the  $k$ -th site as the local site; obtain and broadcast  $\tilde{\beta}_k$  and the variance  $\tilde{V}_k$ ;

**end**

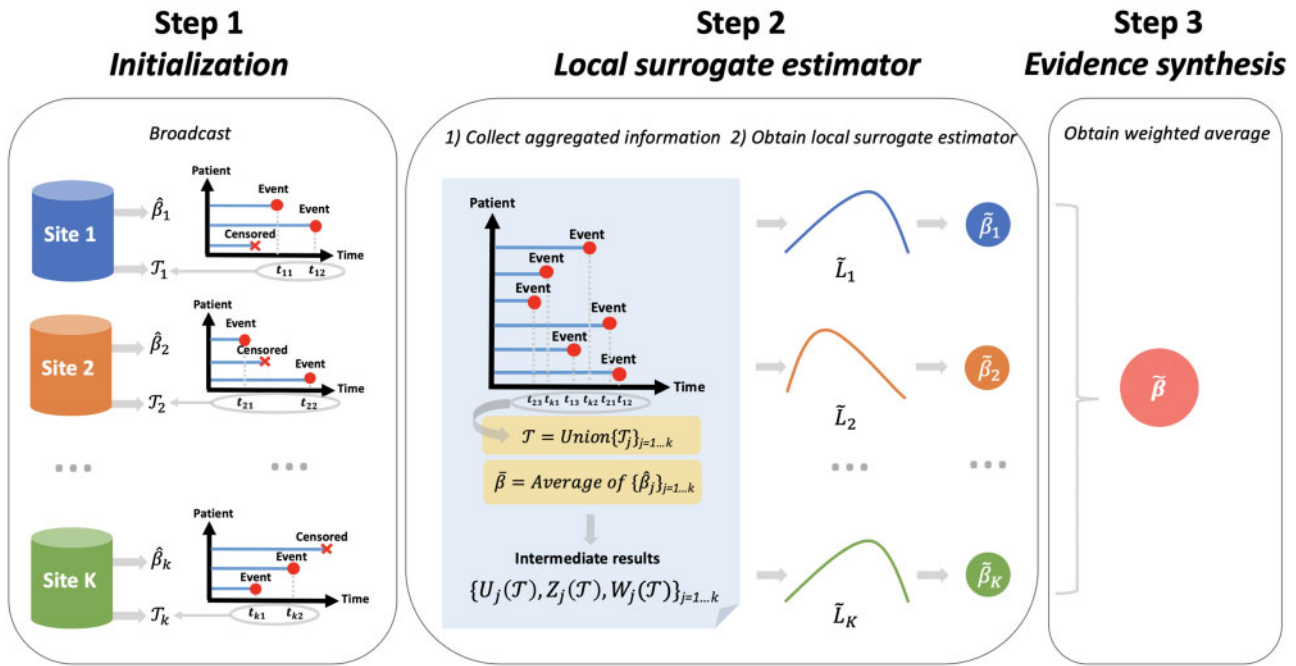
##### (3) Evidence synthesis

Obtain  $\tilde{\beta}$  using (5) by plugging in  $\tilde{\beta}_k$  and  $\tilde{V}_k$ .

**Return**  $\tilde{\beta}$ .

#### A step-by-step illustration of ODAC using a simulated network with 2 sites

To better explain each step in the ODAC algorithm, we simulate 2 datasets that mimic a distributed network with 2 clinical sites. Site A has 100 patients and site B has 50 patients, and the goal is to fit a Cox model to study the association between the survival time and treatment (new drug vs placebo) adjusting for age. We generated age from a uniform distribution with a range from 20 to 60, and the treatment status was generated from a Bernoulli distribution with probability 0.5 of being in each arm. Before fitting the model, we standardized age by subtracting the mean age and dividing by the standard error. The survival time of each patient was then generated



**Figure 1.** Schematic illustration of ODAC to fit a multicenter Cox proportional hazards model algorithm. The first step is initialization, in which each site reports the local estimate of the log hazard ratio ( $\hat{\beta}_j$ ), and the set of event times ( $T_j$ ). In the second step, each site calculates the average of all local estimates, and obtain the union of all event times. Using this information, each site shares the intermediate results ( $U_j(T)$ ,  $Z_j(T)$ ,  $W_j(T)$ ). Next, each site combines the intermediate results and the local patient-level information to construct a surrogate likelihood function defined in equation 3, and obtains the local surrogate estimates by optimizing the surrogate likelihood function. The last step is the evidence synthesis, in which all the local surrogate estimates are combined through a weighted average.

from a Weibull proportional hazards model, in which the baseline hazard follows a Weibull-distribution with scale 200 and shape 20 and the regression parameters for the treatment option and age are set to -1 and 1, respectively. We generated noninformative censoring from a Weibull distribution to keep the event rate at approximately 5%. Figure 2 provides a detailed step-by-step illustration of the implementation of the ODAC algorithm to this simulated dataset.

In the first step, the 2 sites fit the Cox model locally and share the initial estimates of the log HRs and their variances [ $\hat{\beta}_A = (-1.69, 1.58)$ ,  $\hat{\beta}_B = (-0.60, 2.02)$ ,  $\hat{V}_A = (1.20, 0.90)$ ,  $\hat{V}_B = (1.51, 1.91)$ ]. In addition, each site reports their set of unique event times. In site A, there are 6 time points at which an event occurs, that is,  $T_A = (21, 23, 24, 31, 32, 36)$ , and in site B, we have  $T_B = (25, 27, 39)$ . Therefore, in the first step, site A shares  $\hat{\beta}_A$ ,  $\hat{V}_A$ ,  $T_A$  and site B shares  $\hat{\beta}_B$ ,  $\hat{V}_B$ ,  $T_B$ .

In the second step, each site calculates the inverse variance weighted initial estimator  $\tilde{\beta} = (\hat{V}_A^{-1} \hat{\beta}_A + \hat{V}_B^{-1} \hat{\beta}_B) / (\hat{V}_A^{-1} + \hat{V}_B^{-1}) = (-0.68, 1.72)$ , and obtains the combined set of event times  $T$  which contains 9 unique time points. Then for each time  $t \in T$ , site A calculates a scalar  $U_A(t)$ , a  $p$ -dimensional vector  $W_A(t)$ , and a matrix containing  $p^2$  numbers  $Z_A(t)$  (for simple illustration, we vectorize  $Z_A(t)$  in Figure 2). Therefore, in this step, in total each site shares 42 numbers that do not contains patient-level information. Using these numbers, sites A and B are able to construct the surrogate likelihood function within each site, and then obtain and share the surrogate likelihood estimates  $\tilde{\beta}_A = (-0.75, 1.54)$ ,  $\tilde{\beta}_B = (-0.76, 1.53)$ , as well as their corresponding covariance matrices  $\tilde{V}_A$  and  $\tilde{V}_B$ .

In the evidence synthesis step, the final estimator  $\tilde{\beta}_{all}$  is obtained as the weighted average of  $\tilde{\beta}_A$  and  $\tilde{\beta}_B$  with weights equal to the inverse of  $\tilde{V}_A$  and  $\tilde{V}_B$ . In this example, we obtain  $\tilde{\beta}_{all} = (-0.76, 1.54)$ , and the variance of  $\tilde{\beta}_{all}$  is estimated by

$\tilde{V}_{all} = (\tilde{V}_A^{-1} + \tilde{V}_B^{-1})^{-1}$ , which is closer to the true parameter value than the local estimates.

### A real-world use case using OHDSI data

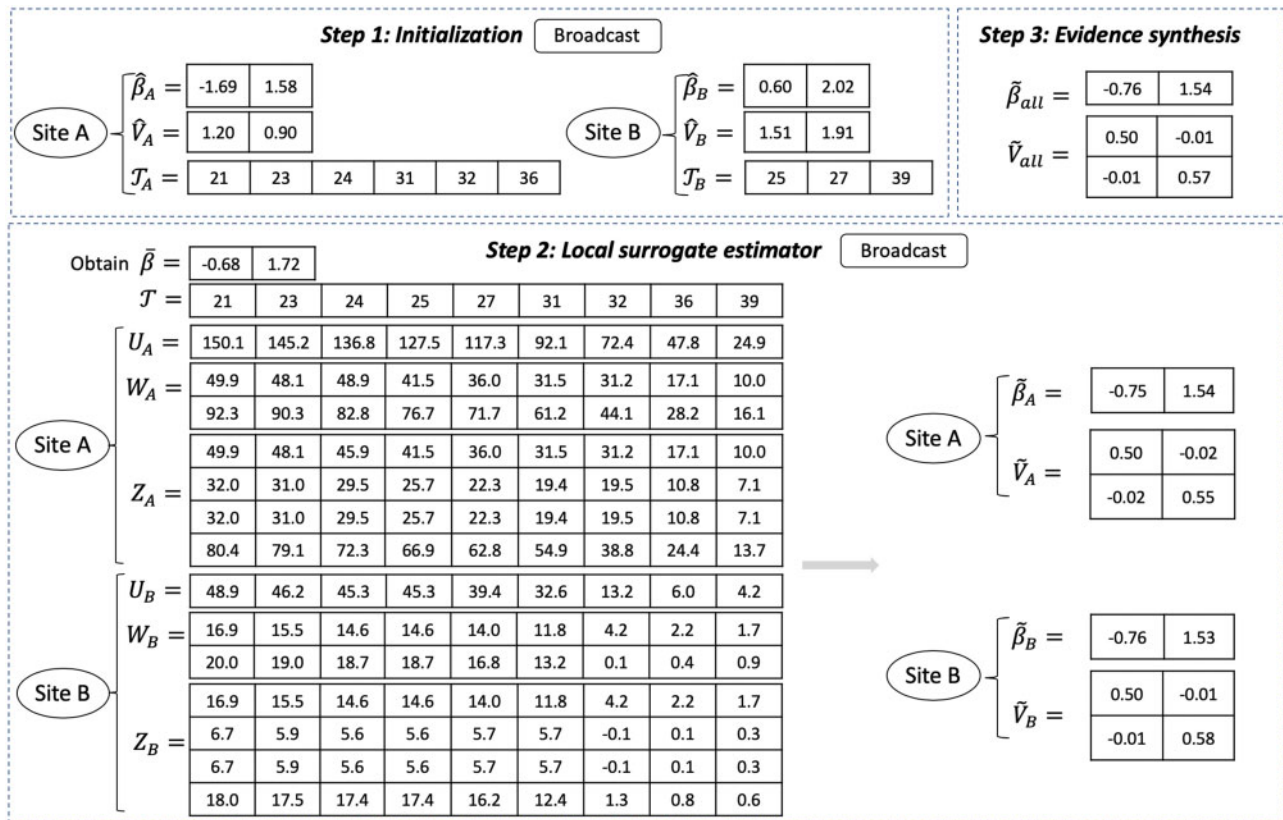
We used ODAC to study the risk factors of AMI and stroke in a population with pharmacologically treated major depressive disorder using data from 4 different U.S. insurance claims databases that have been converted to the Observational Medical Outcomes Partnership CDM<sup>19</sup> in the OHDSI network. Both AMI and stroke were defined as the occurrence of the respective diagnosis codes in an inpatient or emergency room setting. We only counted the first occurrence of each condition per patient in order to preserve independence.

We fit Cox regression models for the 2 outcomes and corresponding risk factors. For AMI, we included known observed risk factors: age, gender, alcohol dependence, hyperlipidemia, hypertensive disorder, depression, obesity, and type 2 diabetes.<sup>20,21</sup> Similarly, for stroke the following known risk factors were included: congestive heart failure, coronary arteriosclerosis, hyperlipidemia, ischemic heart disease, renal failure, hypertensive disorder, transient cerebral ischemia, and type 2 diabetes.<sup>22</sup> We compared our ODAC estimator with the pooled estimator and the estimator from meta-analysis.

### A numerical experiment to demonstrate the benefit of ODAC in studying rare events

To demonstrate the benefit of ODAC comparing to meta-analysis in study rare events, we designed the following numerical experiment. A pooled dataset of  $N = 10\,000$  subjects were generated based on Weibull proportional hazards model, where the baseline hazard follows a Weibull-distribution with scale 200 and shape 20. We generated 2 covariates from independently and identically distributed uniform dis-





**Figure 2.** Step-by-step illustration of ODAC to fit a multicenter Cox proportional hazards model algorithm using a simulation dataset. In the first step, the 2 sites share the estimated log hazard ratios and their variances ( $\hat{\beta}_A = (-1.69, 1.58)$ ,  $\hat{\beta}_B = (-0.60, 2.02)$ ,  $\hat{V}_A = (1.20, 0.90)$ ,  $\hat{V}_B = (1.51, 1.91)$ ), as well as their set of unique event times,  $\mathcal{T}_A = (21, 23, 24, 31, 32, 36)$  and  $\mathcal{T}_B = (25, 27, 39)$ . In the second step, each site calculates the inverse variance-weighted initial estimator  $\tilde{\beta} = (-0.68, 1.72)$ , and obtains the combined set of event times  $\mathcal{T}$ , which contains 9 unique time points. Then, each site shares 63 numbers in this step. Using these numbers, sites A and B obtain and share the surrogate likelihood estimates  $\tilde{\beta}_A = (-0.75, 1.54)$ ,  $\tilde{\beta}_B = (-0.76, 1.53)$ , as well as the covariance matrix  $\tilde{V}_A$  and  $\tilde{V}_B$ . In the evidence synthesis step, the final estimator  $\tilde{\beta}_{all}$  is obtained by the weighted average of  $\tilde{\beta}_A$  and  $\tilde{\beta}_B$  where  $\tilde{\beta}_{all} = (-0.76, 1.54)$ .

tributions and the true log HRs were set to be  $\beta = (-1, 2)$ . We set the event rate (number of cases over number of subjects) as 20%, 5%, 2%, and 1% by appropriately modifying the distribution of censoring times. The pooled data were distributed over  $K = 10$  clinical sites, with 5 large and 5 small sites. We set the relative sizes of the large to small sites to be 1000/1000, 1250/750, 1500/500. Under each scenario, we compared the ODAC estimator with the meta-analysis estimator over 1000 replications. Because the pooled estimator can be considered a gold standard, the bias to the pooled estimator is used as the metric to evaluate the performance of each method. For simplicity of illustration, we only present the results for estimation of coefficient  $\beta_2$ , as the simulation results for  $\beta_1$  are similar.

## RESULTS

### Application to OHDSI data

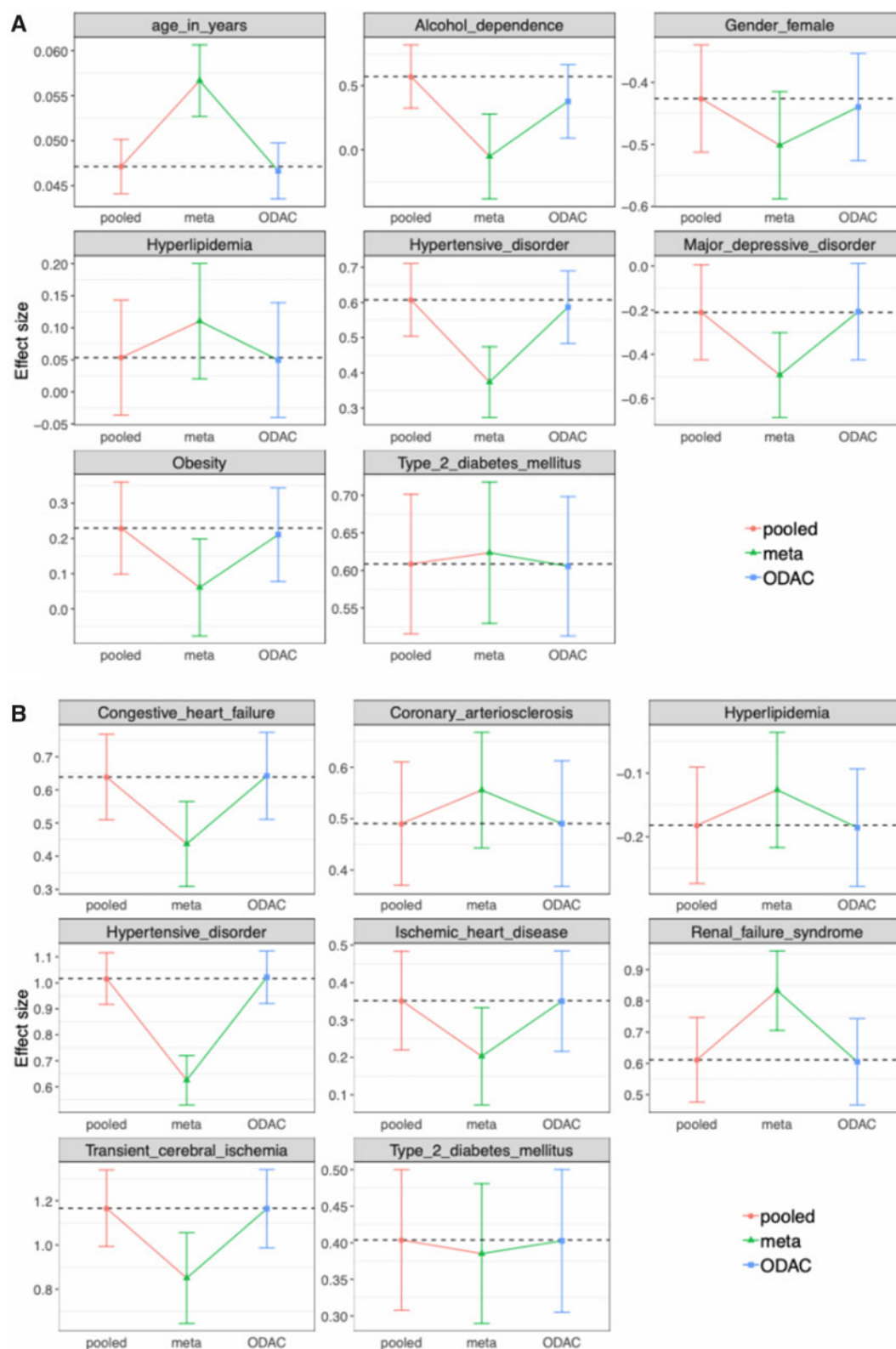
The summaries of patients' characteristics of the 4 datasets are listed in Table 1. The overall event rates are below 1% for both outcomes. The Medicare data (ie, IBM MarketScan Medicare Supplemental Database) were observed to have more elderly patients, and therefore a higher prevalence for both the outcome and risk factors of interest.

Figure 3 shows the estimated log HRs from the 3 methods as well as their 95% confidence intervals (CIs). The ODAC provided HR estimates are nearly identical to the pooled estimates for most of

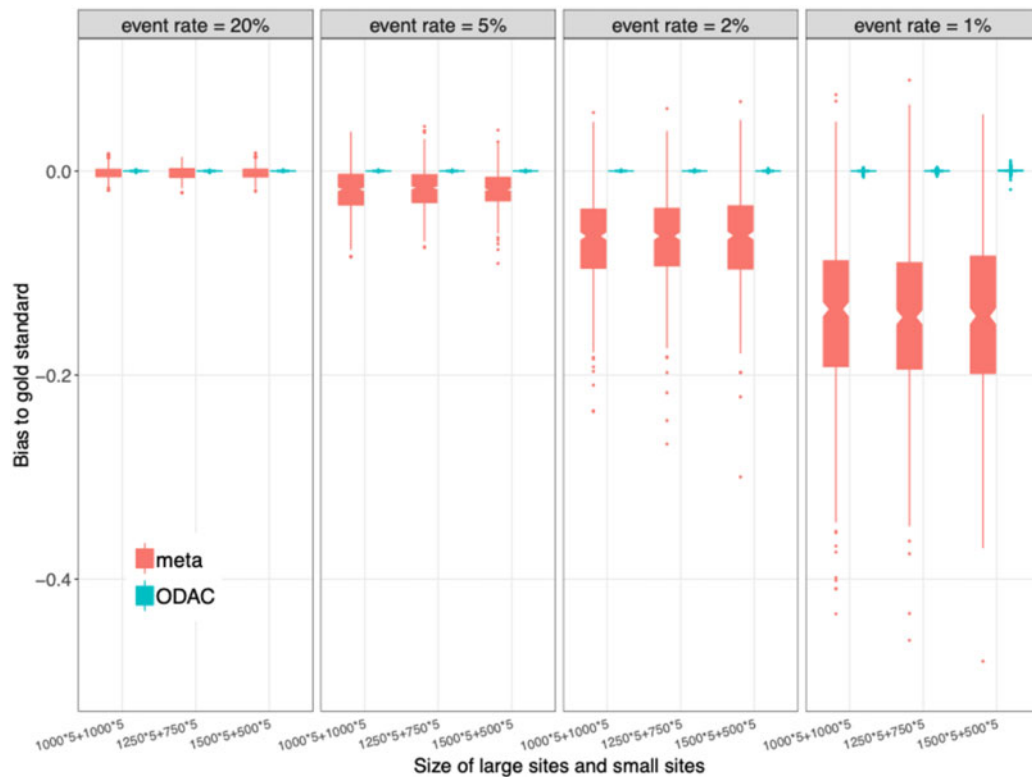
**Table 1.** Characteristics of the 4 claims datasets at the Observational Health Data Sciences and Informatics

Dataset	CCAIE	MDCD	MDCR	Optum
Subjects	64 222	59 861	69 164	62 348
Median age, y	43	35	71	47
Female, %	69.21	73.82	68.08	69.68
Congestive heart failure, %	0.70	3.06	7.58	2.61
Hypertensive disorder, %	20.81	31.80	57.70	32.96
Ischemic heart disease, %	1.70	3.82	10.27	4.10
Type 2 diabetes mellitus, %	7.49	14.63	21.83	12.71
Coronary arteriosclerosis, %	2.39	4.92	18.43	5.75
Renal failure syndrome, %	0.69	2.67	2.31	2.49
Transient cerebral ischemia, %	0.41	0.64	2.32	0.71
Hyperlipidemia, %	20.96	22.00	43.21	33.85
Obesity, %	7.15	16.54	6.71	9.62
Alcohol dependence, %	1.79	2.94	1.01	2.29
Major depressive disorder, %	4.17	3.55	3.16	3.34
Acute myocardial infarction, %	0.26	0.75	2.03	0.51
Stroke, %	0.24	0.73	1.75	0.58

<sup>a</sup>The 4 claims datasets are the IBM MarketScan Commercial Claims and Encounters Data (CCAIE), IBM MarketScan Medicaid Multi-State Medicaid Database, IBM MarketScan Medicare Supplemental Database (MDCR), and OptumDe-Identified Clinformatics.



**Figure 3.** (A) Estimated log hazard ratios with 95% confidence intervals for risk factors of acute myocardial infarction using pooled analysis (red), meta-analysis (green), and One-shot Distributed Algorithm to fit a multicenter Cox proportional hazards model (ODAC) (blue). (B) Estimated log hazard ratios with 95% confidence intervals for risk factors of stroke using pooled analysis on the combined dataset across all sites (red), meta-analysis (green), and ODAC (blue).



**Figure 4.** Boxplot of relative bias to the gold standard (pooled estimator obtained by fitting Cox model on the combined dataset across all sites) in different simulation settings. The 2 methods illustrated in the plot are meta-analysis (red) and One-shot Distributed Algorithm to fit a multicenter Cox proportional hazards model (ODAC) (blue). The event rate varies from 20% to 1%, and there are 3 different sample size distributions: (1) all 10 sites have 1000 samples, (2) 5 sites with 1250 samples and 5 with 750 samples, and (3) 5 sites with 1500 samples and 5 with 500 samples. Under each setting, the boxplots are based on 1000 replications of the experiment.

the risk factors, while meta-analysis estimates have substantial biases compared with the pooled estimator. The differences in estimated effect sizes can lead to potentially different conclusions in the investigation of risk factors. For example, the estimated log HR of alcohol dependence for AMI changed signs when comparing the pooled estimate (0.571; 95% CI, 0.322-0.820) to the meta-analysis estimate (−0.053; 95% CI, −0.361 to 0.255). In the HR scale, the pooled estimates was 1.770 (95% CI, 1.380-2.271) and the meta-analysis estimate was 0.948 (95% CI, 0.697-1.290). Furthermore, the meta-analysis estimates were sometimes qualitatively different from the pooled analysis and ODAC estimates. For example, both the pooled and the ODAC estimates suggested that alcohol dependence is significantly associated with the time to AMI, while the meta-analysis estimate showed it is not a significant risk factor. Similar contrasts were also present regarding risk factors such as hyperlipidemia, obesity, and depression. For all the risk factors of stroke, the relative bias of ODAC was <2%, whereas the relative bias of meta-analysis estimates were as high as 109%. This real-world example suggests that the ODAC estimator is preferred over meta-analysis in the investigation of risk factors of rare time-to-event outcomes.

### Numerical experiments

Our numerical study shows that ODAC has better performance than the meta-analysis estimator especially when the event is rare. The conclusion is supported by Figure 4, which shows the boxplots of the bias to the pooled estimator for different event rates and sample sizes. We observe that for all scenarios, ODAC obtains relative biases

close to 0, meaning that it provides almost identical results to the pooled estimator. When the event rate is fixed and the sample size changes, the performance of the ODAC and meta-analysis estimators stay unchanged. When the event rate decreases, meta-analysis estimator is observed to have larger bias. When event rate is 1%, the average bias of the meta-analysis estimator is around −0.14, while the largest absolute bias of ODAC is 0.01. In addition to the difference in magnitude of bias, the variation of the meta-analysis estimator is much larger compared with that of the ODAC estimator.

### DISCUSSION

In this article, we developed and evaluated ODAC, a distributed, privacy-preserving algorithm to fit a Cox model in a federated multicenter setting. ODAC does not require iterative communication across sites, which minimizes the communication cost, and therefore is easy to implement in large research networks such as OHDSI and PCORnet as well as smaller networks such as those within PCORnet, such as PEDSnet (A National Pediatric Learning Health System)<sup>23,24</sup> and OneFlorida.<sup>25</sup>

ODAC was shown to be significantly more accurate than conventional meta-analysis, particular when event rates were low, 1% or below; even though the communication cost (defined as the quantity of numbers needed to be transferred or shared) is greater than that of meta-analysis. Specifically, for a model with  $p$  parameters, the meta-analysis requires each site to share 2 numbers (a point estimate and a standard error for each parameter), while ODAC requires transferring  $(M_i + 1)(p^2 + p + 1)$  numbers from each site,

where  $M_t$  is the number of unique event time points across all sites. Usually,  $M_t$  is not very large in the EHR data setting, as the event time is often measured in days. For example, for a 1-year follow-up, the largest possible  $M_t = \min(365, n_e)$ , where  $n_e$  is the total number of events. When studying extremely rare events,  $n_e$  is likely to be much smaller than the number of days during the follow-up period. As a consequence, when the number of covariates in the model is not extremely large, transferring  $O(M_t p^2)$  numbers is not considered a burden in multicenter analysis. Yet, the improvement in estimation accuracy is substantial, compared with meta-analysis, particularly when the event rate is low. Compared with iterative methods such as WebDISCO, the communication cost of ODAC is low.

Implementing ODAL in distributive networks such as OHDSI is easy, as it does not require iterative communications. The initialization step is fairly standard in a multicenter analysis where each site shares their local estimates. After the initialization, each site is required to share the intermediate quantities  $U_k$ ,  $Z_k$ ,  $W_k$  to all the other sites. These quantities all have closed forms and can be obtained using prewritten code or software packages distributed across the network. Both OHDSI and PCORnet have the necessary infrastructure—ARACHNE (A distributed OHDSI research network and study workflow orchestration)<sup>26</sup> and the PCORnet Query Tool<sup>27</sup> based on PopMedNet (an open-source application used to facilitate multisite health data networks)<sup>28</sup> to distribute these analytical codes. The surrogate partial likelihood estimator can also be obtained using an optimization function that is commonly found in prewritten code or software packages. The code for ODAC is available from the authors upon request.

The impact of this work is clear, on several dimensions. First, ODAC affords researchers the ability to conduct Cox analyses across many sites in a federated network. Second, ODAC provides a means to preserve the privacy and confidentiality of patients in the context of research conducted on federated data networks. Rather than require individual patient-level data from all network sites, such data are required from only 1 site. The other sites a proxied by aggregate data from each of the other sites. Third, federated data networks are essential to the study of rare diseases. ODAC estimators were shown to be superior to meta-analysis across a range of sampling regimes including rare outcomes.

Our ODAC algorithm is based on the pooled analysis, which is to fit a unified Cox regression model on the combined dataset. Therefore it essentially requires the data are homogenous distributed across sites. In the future, we plan to extend our method to handle heterogeneity across clinical sites by allowing site-specific effects and covariates. We are also developing open-source software packages to support studies that wish to use ODAC for multicenter analysis in existing distributed networks such as OHDSI and PCORnet. One key task is to implement ODAC according to the networks' Observational Medical Outcomes Partnership CDM for OHDSI and the PCORnet CDM to alleviate the potentially complicated preprocessing procedures. We believe that ODAC is a significant contribution to the fast-growing distributed research networks with enormous patient-level RWD who are facing data sharing issues and privacy concerns, and ultimately will help facilitate collaborative efforts to investigate risk factors for time to event outcomes in healthcare systems. Some of the numerous applications in which it may be a particularly powerful tool involve use in modeling risk factors for disease progression and treatment effectiveness for uncommon and rare diseases. Because it minimizes communication costs, it will also be an excellent tool for international studies that require time-to-event modeling of outcome.

## CONCLUSION

The proposed ODAC algorithm for multicenter Cox model is privacy-preserving and noniterative. Through a simulation study and a real-world use case using OHDSI data, ODAC was shown to have higher estimation accuracy compared with the meta-analysis, especially for studying rare events.

## FUNDING

This work was supported in part by the National Institutes of Health grants 1R01LM012607, 1R01AI130460, 5R01LM010098, UL1TR001427, R21AG061431, R01CA246418, and 1R01AI116794, a grant from the Patient-Centered Outcomes Research Institute (PCORI) for the PEDSnet Clinical Research Infrastructure (RI-CRN-2020-007), and the Cancer Informatics and eHealth Core program at the University of Florida Health Cancer Center.

## AUTHOR CONTRIBUTIONS

RD, CL, and YC designed methods and experiments; MJS provided the dataset from the Observational Health Data Sciences and Informatics; CL and JT conducted simulation experiments; MJS conducted data analysis; all authors interpreted the results and provided instructive comments; RD, CL, and YC drafted the main article. All authors have approved the article.

## SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

## CONFLICT OF INTEREST STATEMENT

None declared.

## 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 (23): 2293–7.
2. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. *Sci Transl Med* 2010; 2 (57): 57cm29.
3. Hagar Y, Albers D, Pivovarov R, et al. Survival analysis with electronic health record data: experiments with chronic kidney disease. *Stat Anal Data Mining* 2014; 7 (5): 385–403.
4. Ranganath R, Perotte A, Elhadad N, Blei D. Deep survival analysis. arXiv preprint arXiv: 1608.02158, 2016.
5. Cox DR. Regression models and life tables. *J R Stat Soc Series B Stat Methodol* 1972; 34 (2): 187–202.
6. 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.
7. Fleurence RL, Curtis LH, Califf RM, et al. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc* 2014; 21 (4): 578–82.
8. 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: analysis from the observational health data sciences and informatics initiative. *JAMA Netw Open* 2018; 1 (4): e181755.
9. 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 (3): 275–88.
10. Duke JD, Ryan PB, Suchard MA, et al. Risk of angioedema associated with levetiracetam compared with phenytoin: findings of the observational health data sciences and informatics research network. *Epilepsia* 2017; 58 (8): e101–6.



11. 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 (27): 7329–36.
12. Chen Y, *et al.* Regression cubes with lossless compression and aggregation. *IEEE Trans Knowl Data Eng* 2006; 18 (12): 1585–99.
13. Wu Y, Jiang X, Kim J, *et al.* Grid Binary LOGistic REGression (GLORE): building shared models without sharing data. *J Am Med Inform Assoc* 2012; 19 (5): 758–64.
14. 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 (6): 1212–9.
15. Wang J, Kolar M, Srebro N, Zhang T. Efficient distributed learning with sparsity. *Proc Int Conf Mach Learn* 2017; 70: 3636–45.
16. Jordan MI, Lee JD, Yang Y. Communication-efficient distributed statistical inference. *J Am Stat Assoc* 2019; 114 (526): 668–81.
17. 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.
18. 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 (4): 621–6.
19. 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 (1): 54–60.
20. Anand SS, Islam S, Rosengren A, *et al.* Risk factors for myocardial infarction in women and men: insights from the INTERHEART study. *Eur Heart J* 2008; 29 (7): 932–40.
21. Lanas F, Avezum A, Bautista LE, *et al.* Risk factors for acute myocardial infarction in Latin America: the INTERHEART Latin American study. *Circulation* 2007; 115 (9): 1067–74.
22. Kokotailo RA, Hill MD. Coding of stroke and stroke risk factors using international classification of diseases, revisions 9 and 10. *Stroke* 2005; 36 (8): 1776–81.
23. Forrest CB, Margolis PA, Bailey LC, *et al.* PEDSnet: a national pediatric learning health system. *J Am Med Inform Assoc* 2014; 21 (4): 602–6.
24. Forrest CB, Margolis P, Seid M, *et al.* PEDSnet: how a prototype pediatric learning health system is being expanded into a national network. *Health Aff (Millwood)* 2014; 33 (7): 1171–7.
25. Shenkman E, Hurt M, Hogan W, *et al.* OneFlorida Clinical Research Consortium: linking a clinical and translational science institute with a community-based distributive medical education model. *Acad Med* 2018; 93 (3): 451–5.
26. Observational Health Data Sciences and Informatics. **OHDSI Network Research 2019**. In: *The Book of OHDSI*. Chapter 20. New York, NY: OHDSI. <https://ohdsi.github.io/TheBookOfOhdsi/NetworkResearch.html> Accessed October 13, 2019.
27. PCORnet Distributed Research Network Data-Driven Common Model. <https://pcorntest.org/data-driven-common-model/> Accessed October 13, 2019.
28. PopMedNet. 2012; <https://www.popmednet.org/> Accessed October 13, 2019.