4-22-2022

ODACH: A One-shot Distributed Algorithm for Cox model with heterogeneous multi-center data

Chongliang Luo

Rui Duan

Adam C Naj

Henry R Kranzler

Jiang Bian

See next page for additional authors

Authors

Chongliang Luo, Rui Duan, Adam C Naj, Henry R Kranzler, Jiang Bian, and Yong Chen

scientific reports

Check for updates

OPEN

ODACH: a one-shot distributed algorithm for Cox model with heterogeneous multi-center data

Chongliang Luo[1,2], Rui Duan[3], Adam C. Naj[2,4], Henry R. Kranzler[5], Jiang Bian[6] & Yong Chen[2]✉

We developed a One-shot Distributed Algorithm for Cox proportional-hazards model to analyze Heterogeneous multi-center time-to-event data (ODACH) circumventing the need for sharing patient-level information across sites. This algorithm implements a surrogate likelihood function to approximate the Cox log-partial likelihood function that is stratified by site using patient-level data from a lead site and aggregated information from other sites, allowing the baseline hazard functions and the distribution of covariates to vary across sites. Simulation studies and application to a real-world opioid use disorder study showed that ODACH provides estimates close to the pooled estimator, which analyzes patient-level data directly from all sites via a stratified Cox model. Compared to the estimator from meta-analysis, the inverse variance-weighted average of the site-specific estimates, ODACH estimator demonstrates less susceptibility to bias, especially when the event is rare. ODACH is thus a valuable privacy-preserving and communication-efficient method for analyzing multi-center time-to-event data.

Real-world data (RWD) such as electronic health records (EHRs) and medical claims, are used increasingly to provide evidence-based support for healthcare decision making[1–3]. The past decade has seen an increasing number of clinical research networks that accumulate and promote the use of large collections of RWD for clinical research. For example, the international Observational Health Data Sciences and Informatics (OHDSI) collaborative[4], and the national Patient-Centered Clinical Research Network (PCORnet) in the United States[5], both cover hundreds of millions of patients. These large data consortia provide opportunities to integrate RWD from various healthcare organizations. Multicenter analyses using RWD from these clinical research networks have expanded rapidly because of improved generalizability from more representative population samples and increased statistical power to detect modest associations between exposures and outcomes.

Despite the benefits of multicenter analyses, two major challenges exist for multi-site data integration. First, the direct sharing of patient-level data across institutions may be prohibited, as individual patient-level data are protected by privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States or the European Union's General Data Protection Regulation (GDPR). Hence many research networks such as OHDSI and PCORnet have adopted a federated model in which patient-level data are stored at local institutions and often only aggregated information are shared across sites[5–7]. Second, data from different sites are often heterogeneous with respect to patient characteristics, data quality, and other unobserved site-specific features. Assuming that a common statistical model is appropriate across all sites may result in biased estimation and poor prediction.

Within these clinical research networks, the abundance of EHRs containing data on patients at multiple time points is especially useful for survival analyses, which model the time to a specific outcome or event of interest. To conduct multicenter survival analyses without sharing patient-level data, a common and convenient approach

[1]Division of Public Health Sciences, Washington University School of Medicine in St. Louis, St. Louis, MO, USA. [2]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. [3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [4]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [5]Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania and the VISN 4 MIRECC, Crescenz VAMC, Philadelphia, PA, USA. [6]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, USA. ✉email: ychen123@upenn.edu

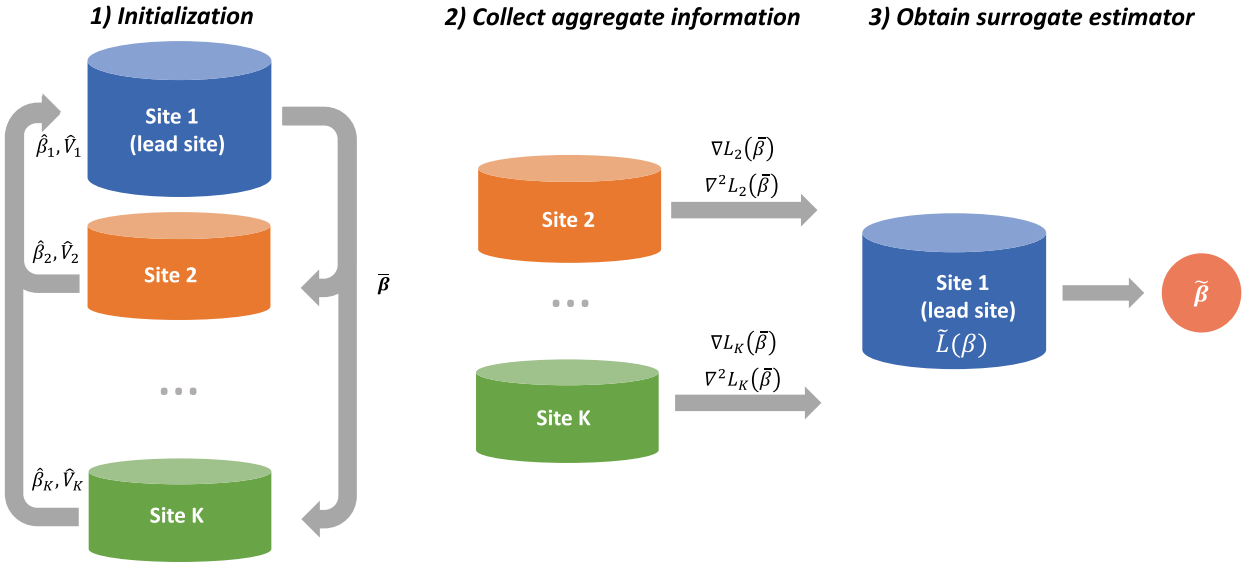

Figure 1. Schematic illustration of the ODACH algorithm. The first step is initialization, where each site reports the local estimation of the log hazard ratio ($\hat{\beta}_j$) and the corresponding variance estimate ($\hat{V}_j$). The lead site then computes the initial estimate $\bar{\beta}$ as the weighted average of all local estimates and sends it back to each site. In the second step, each site calculates and shares the local gradients $\nabla L_j(\bar{\beta})$ and $\nabla^2 L_j(\bar{\beta})$. In the third step, the lead site constructs a surrogate likelihood function $\tilde{L}(\beta)$ with these gradients and obtains the surrogate estimate β.

is meta-analysis, where a weighted average of the local estimates from each site is used. However, when the outcomes or exposures are rare, or the samples at some sites are small, the accuracy of the meta-analysis may be low[8,9]. To obtain more accurate results under these conditions, distributed algorithms have been developed, such as the WebDISCO (a web service for distributed Cox model learning)[10]. Despite providing identical results to that from pooling individual-level data ("lossless"), this algorithm is communication intensive due to its iterative nature, which requires multiple rounds of communications across sites. To balance communication efficiency and estimation accuracy, Shu et al.[11] proposed a lossless one-shot algorithm for a stratified Cox model that can include only one binary covariate in the model. Huang and Huo[12] proposed a distributed one-step estimator to improve the accuracy of meta-analysis estimator. Wang et al.[13] proposed a "divide-and-conquer approach," aiming to reduce the computational burden when the sample size is extremely large. Duan et al.[9] proposed a One-shot Distributed Algorithm for Cox model (ODAC) based on the surrogate likelihood approach that relies on patient-level data from a single site and aggregated data from other sites. This algorithm requires aggregated data from only two iterations but obtains estimates close to those resulting from the inclusion of patient-level data from all sites.

Most of these approaches are based on the Cox proportional-hazards model, with a few accounting for between-site heterogeneity. Specifically, in multicenter survival analyses, baseline hazard functions and the distribution of covariates are likely to differ across sites as patients often come from different sub-populations varying in racial/ethnic compositions across geographic regions. Ignoring the heterogeneity across sites could lead to biased estimated associations. Here we propose a distributed algorithm that accounts for site-level heterogeneities in covariate distributions and baseline hazard functions, the One-shot Distributed Algorithm for Cox model with Heterogeneity (ODACH). Compared to the previously described ODAC, which assumes a common baseline hazard function across sites, ODACH assumes heterogeneous baseline hazard functions, and is therefore more flexible and practical in real-world settings. Moreover, unlike ODAC, the use of a constructed surrogate likelihood means that ODACH does not require an extra round of communication regarding the risk set in each site, improving communication efficiency. We illustrate in a simulation study and in a real-world multicenter opioid use disorder study that our proposed algorithm is both a 'one-shot' approach and highly accurate (i.e., demonstrates less bias).

Results

A one-shot distributed algorithm for cox model with heterogeneity. The proposed ODACH algorithm constructs a surrogate log-likelihood function to approximate the log-likelihood function of the stratified Cox model, which is commonly used to account for site-specific baseline hazards when analyzing multi-site time-to-event outcomes. We provide a schematic illustration of the ODACH algorithm in Fig. 1.

ODACH can reduce estimation bias in multicenter survival analyses. We used a simulation study to demonstrate the bias-reduction property of the proposed ODACH algorithm in multicenter survival analyses, especially when the outcome is rare. We generated time-to-event outcomes that are associated with two covariates. The pooled data are evenly distributed to $K = 10$ clinical sites. Details of the data generation are in the "Methods" section. We applied three approaches to estimate the HRs of the two covariates on the time-to-

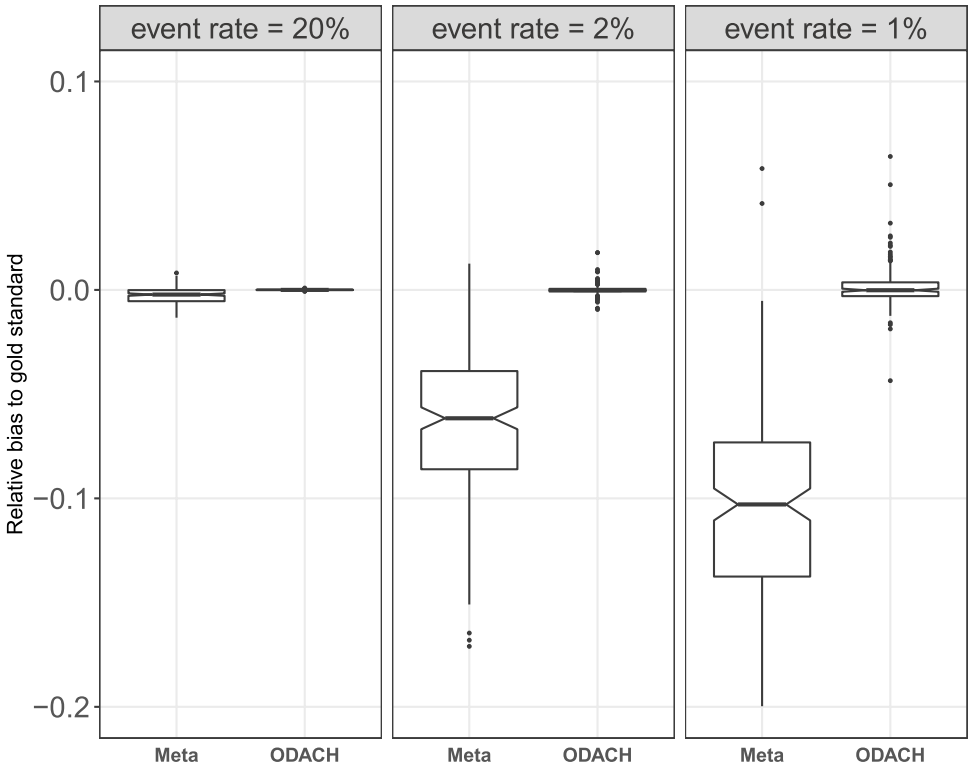

Figure 2. Boxplot of bias relative to the gold standard (stratified Cox model on the pooled dataset across all sites). The two methods compared in the plot are meta-analysis (meta) and One-shot Distributed Algorithm for Cox model with Heterogeneous baseline hazards (ODACH). The event rate varies from 20 to 1% and under each setting the boxplots are based on 200 replications of the simulation. The true effect size is 1.

event outcome, i.e., pooled stratified Cox regression, meta-analysis, and the proposed ODACH method. Because the pooled Cox regression estimator can be considered a gold standard, the relative bias of meta-analysis and ODACH estimates to the pooled estimate are compared to demonstrate the advantage of ODACH.

Results of the simulation show that ODACH achieves better estimation performance than the meta-analysis estimator, especially for a rare event. Figure 2 shows that ODACH yields relative biases close to 0, meaning that it provides results almost identical to those of the pooled estimator, i.e., by stratified Cox model on the pooled dataset across all sites. As the event becomes rarer, the meta-analysis estimator is observed to have a larger bias. For example, when the event rate is 1%, the average relative bias is around − 11% for the meta-analysis estimator, but is negligible for the ODACH estimator. Moreover, the variation of the meta-analysis estimator is much larger than that of the ODACH estimator.

The OneFlorida opioid use disorder study. We demonstrate the use and advantage of the proposed ODACH method by studying the association of time to an opioid use disorder (OUD) diagnosis with risk factors (e.g., patients' demographic and clinical characteristics) using RWD from the OneFlorida Clinical Research Consortium. A detailed description of the data and the risk factors are in the "Methods" section.

Figure 3 shows the estimated log HRs of the seven risk factors from the pooled analysis (stratified Cox model), meta-analysis, and the proposed ODACH analysis and their 95% confidence intervals (CIs). The ODACH provides HR estimates that are nearly identical to the pooled estimates for all of the risk factors. As a comparison, meta-analysis estimates have substantial biases relative to the pooled estimator, especially for the effects of age, smoking status, and pain history. For example, the estimated log HR of pain history is 0.554 from the pooled analysis, 1.097 from the meta-analysis, and 0.491 from the ODACH estimator. The relative bias is 98.0% for the meta-analysis estimator and − 11.4% for the ODACH estimator. Moreover, the quantitatively larger biases of meta-analysis estimates may lead to qualitatively different statistical significance. For example, with a significance threshold $\alpha = 0.05$, the effect of pain history on time to OUD diagnosis is considered statistically significant ($p = 0.001$) per the meta-analysis estimator, but not statistically significant per either the pooled analysis ($p = 0.061$) or the ODACH estimator ($p = 0.111$).

Discussion

We developed a privacy-preserving One-shot Distributed Algorithm for the Cox model to analyze Heterogeneous multicenter time-to-event data (ODACH). The proposed surrogate likelihood approach approximates the log partial likelihood of the stratified Cox model that uses patient-level data from all of the sites. The simulation study and application to the real-world OneFlorida OUD study both show that the surrogate estimation yields

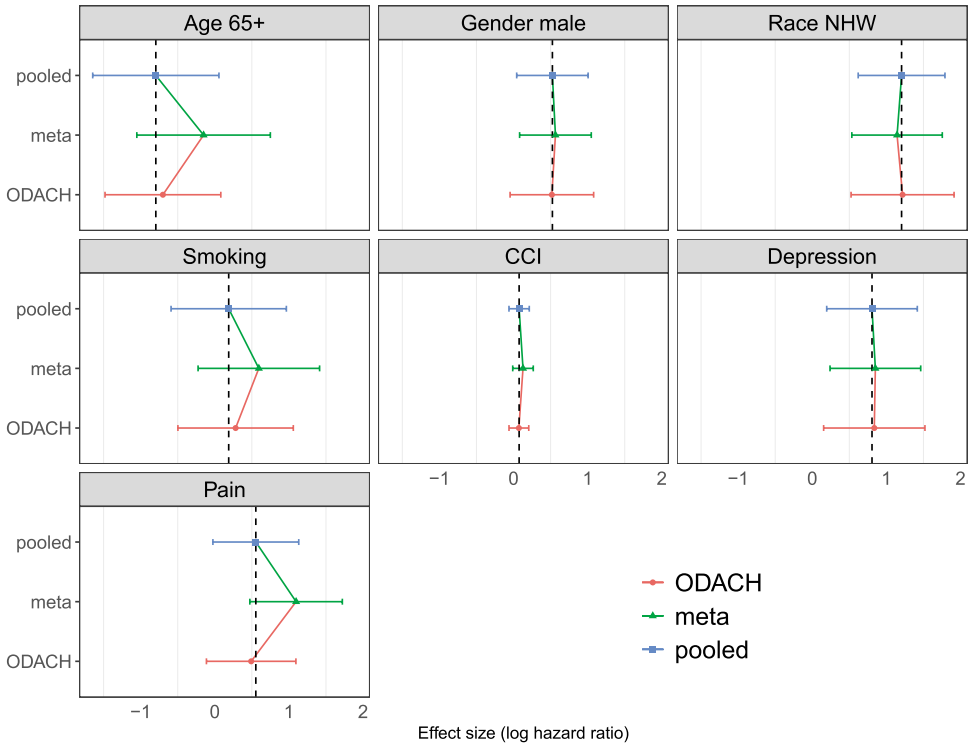

Figure 3. Comparison of estimation biases by meta-analysis and ODACH in the opioid use disorder study. Presented are the estimated log hazard ratios (HRs) with 95% confidence intervals for risk factors for opioid use disorder using pooled analysis (blue), meta-analysis (green), and One-shot Distributed Algorithm for multicenter Cox proportional hazards model with heterogeneous hazard (ODACH) (red). The analyses used data of N = 14,015 patients from five clinical sites in the OneFlorida Clinical Research Consortium.

results that are closer than the typical meta-analysis approach to the pooled analysis results, especially when the event is rare. As suggested by a reviewer, simulation results comparing more approaches are deferred in the Supplementary Information. Compared to the existing One-shot Distributed Algorithm for Cox model (ODAC), ODACH allows baseline hazard functions and covariate distributions to be site specific, and hence it is more flexible in its application.

RWD play an increasing role in generating real-world evidence to support healthcare decision making. Observational RWD such as those from EHRs and medical claims contain longitudinal information, which enables time-to-event modeling such as through the Cox proportional-hazards model, one of the most commonly used models for time-to-event analysis in observational studies that evaluate treatment effects and identify risk factors. In multicenter studies, when sharing patient-level data across databases is not possible, the individual estimates from each database are integrated through a meta-analysis approach. Our proposed distributed algorithm could provide a better alternative to the commonly used meta-analysis, with particular benefits in the case of rare events. The algorithm is implemented in the R package "*pda*"[14]. A demo example is available at https://github.com/Penncil/ODACH.

There are several directions for future work. For instance, time-varying covariates or time-varying effects are sometimes encountered in time-to-event analyses[15,16]. Under these conditions, because the Cox model relaxes the usual proportional hazards assumption but requires additional data for accurate estimation[17,18], the development of a distributed algorithm for the Cox model with time-varying covariates or time-varying effects in multi-center studies would be desirable. Moreover, in certain settings, other survival models such as the accelerated failure time (AFT) model[19] are more appropriate than the Cox model. A distributed algorithm for the AFT model is currently under investigation and will be reported in the future. In addition, because sources of heterogeneity other than baseline hazard functions or distributions might exist, such as missing data patterns and site-specific effect sizes, robust methods for handling different types of heterogeneity[20–22] are needed to avoid potentially misleading results.

Methods

The ODACH algorithm. Suppose that we have study subjects from K different clinical sites and denote n_j to be the number of subjects in the j-th site. We denote the total number of subjects as $N = \sum_{j=1}^{K} n_j$. For the i-th subject in the j-th site, we observe $\{T_{ij}, \delta_{ij}, x_{ij}\}$, where T_{ij} is the observed time to event, x_{ij} is a p-dimensional covariate variable, and $\delta_{ij} = 0$ indicates censoring and $\delta_{ij} = 1$ indicates an event. The Cox proportional hazard model describes that the hazard of the i-th subject in the j-th site having the event at time t as $\lambda\left(t|x_{ij}\right) = \lambda_j(t)\exp\left(\beta^T x_{ij}\right)$. We assume that the log hazard ratio (HR) β is the same across all sites, i.e., there

are common effects of the covariates on the time-to-event across sites. The stratified log Cox partial likelihood function is

$$L(\beta) = \frac{1}{N}\sum_{j=1}^{K} n_j L_j(\beta), \tag{1}$$

where $L_j(\beta)$ is the log Cox partial likelihood function for the j-th site,

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

where $R_j(t)$ is the risk set in site j at time t defined as $R_j(t) = \{i;\ T_{ij} \geq t\}$, which contains all of the subjects in site j who have not experienced an event or been censored at time t. The common effect β can be estimated by maximizing (1), i.e., $\widehat{\beta} = argmax_\beta L(\beta)$. We call this the pooled estimator, as it requires all of the data to be pooled together.

In practice, it is often difficult to transfer patient-level data across sites, hence the pooled estimate $\widehat{\beta}$ can be hard to obtain. Inspired by the previously-developed surrogate likelihood approach[8,9,23], we aimed to construct a proxy of the stratified Cox partial likelihood function (1), using only summary-level information from other sites. We assume we have access only to the patient-level data from a lead site (e.g., the first site). The ODACH surrogate likelihood function is constructed as

$$\tilde{L}(\beta) = L_1(\beta) + \langle \nabla L(\overline{\beta}) - \nabla L_1(\overline{\beta}), \beta\rangle + \frac{1}{2}(\beta - \overline{\beta})^T\{\nabla^2 L(\overline{\beta}) - \nabla^2 L_1(\overline{\beta})\}(\beta - \overline{\beta}), \tag{3}$$

where $L_1(\beta)$ is the log-likelihood function of the lead site, and ∇ and ∇^2 denote the first and second order gradients of a function (explicit forms of $\nabla L_j(\overline{\beta})$, $\nabla L(\overline{\beta})$, $\nabla^2 L_j(\overline{\beta})$ and $\nabla^2 L(\overline{\beta})$ can be found in the Supplementary Materials). $\overline{\beta}$ is an initial value that is close to the true value of β, e.g. the inverse variance-weighted average of the estimates obtained by fitting a Cox model at each site,

$$\overline{\beta} = \left(\sum_{j=1}^{K} \widehat{V}_j^{-1}\right)^{-1} \sum_{j=1}^{K} \widehat{V}_j^{-1}\widehat{\beta}_j, \tag{4}$$

where $\widehat{\beta}_j = \underset{\beta}{\text{argmax}} L_j(\beta)$ is the estimator of the Cox model fitted on data at the j-th site, and $\widehat{V}_j$ is the estimated variance of $\widehat{\beta}_j$. The surrogate estimator is thus obtained by maximizing (3), i.e., $\tilde{\beta} = argmax_\beta \tilde{L}(\beta)$.

Intuitively, the surrogate likelihood function (3) modifies the likelihood function $L_1(\beta)$ of the lead site to approximate the stratified likelihood (1), with the modification being the first- and second-order terms, i.e., $\langle \nabla L(\overline{\beta}) - \nabla L_1(\overline{\beta}), \beta\rangle$ and $\frac{1}{2}(\beta - \overline{\beta})^T\{\nabla^2 L(\overline{\beta}) - \nabla^2 L_1(\overline{\beta})\}(\beta - \overline{\beta})$. By sharing the second-order gradients, our method allows each site to have different covariate distributions. In the construction of the surrogate likelihood function (3), $\nabla^r L(\overline{\beta})$ can be calculated distributively by $\nabla^r L(\overline{\beta}) = \frac{1}{N}\sum_{j=1}^{K} n_j \nabla^r L_j(\overline{\beta})$, for $r = 1, 2$. Because $\nabla L_1(\overline{\beta})$ and $\nabla^2 L_1(\overline{\beta})$ are available from the lead site, it only requires other collaborative sites to calculate and transfer $\nabla L_j(\overline{\beta})$ and $\nabla^2 L_j(\overline{\beta}), j = 2, \ldots, K$. As these gradients are all aggregated information, patient-level information is protected. We summarize the ODACH algorithm in the box below.

Note that we assume the first site is the lead site when constructing the surrogate likelihood. In practice, if any site can serve as the lead site, we recommend using the largest site for this purpose. Alternatively, after the derivatives $\nabla^r L_j(\overline{\beta}), r = 1, 2, j = \ldots, K$ have been shared across sites, each site can serve as the lead site and obtain its own surrogate estimate. These surrogate estimates can be further synthesized to obtain more accurate estimation.

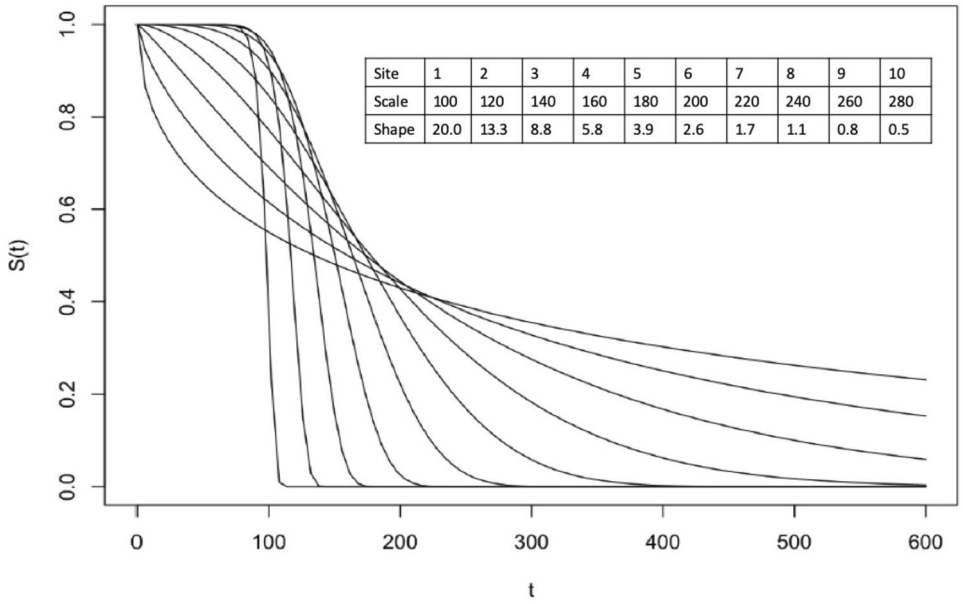

Figure 4. The baseline survival functions of the 10 sites in the simulated data. The varying hazard functions are Weibull functions with scale and shape parameters as listed.

Box 1. Pseudo-code of the ODACH algorithm.

Algorithm ODACH
(1) Initialization
In Site k = 1 to K,
do
Fit a Cox regression model and obtain the local estimate $\hat{\beta}_k$ and the variance estimate $\hat{V}_k$
broadcast $\hat{\beta}_k$, $\hat{V}_k$.
end
(2) Aggregated data communication
In Site k = 1 to K,
do
obtain $\bar{\beta}$ using (4)
calculate and broadcast the gradients $\nabla L_j(\bar{\beta})$, and $\nabla^2 L_j(\bar{\beta})$
end
(3) Local surrogate estimator
In the leading site k = 1
do
construct the surrogate likelihood $\tilde{L}(\beta)$ in (3)
obtain $\tilde{\beta}$ by maximizing $\tilde{L}(\beta)$.
Return $\tilde{\beta}$**.**

Simulation study. We evaluated the performance of the proposed ODACH estimator using simulated multi-site time-to-event data. A pooled dataset of N = 5000 subjects was generated based on a Weibull proportional hazards model, where the baseline hazard follows a Weibull distribution with varying scale and shape parameters. Specifically, the scale parameters range from 100 to 280 and are equally spaced. The shape parameters range from 20 to 0.5, spaced equally in the logarithmic scale (see Fig. 4 for an illustration). We generated two covariates from uniform distributions and the true log HRs were set to be $\beta = (-1, 1)$. We set the event rate (number of cases over number of subjects) as 20%, 2% or 1% by generating censoring times following appropriate distributions. The pooled data were evenly distributed to K = 10 clinical sites, with 500 subjects in each site. We applied three approaches to estimate the HRs of the two covariates on the time-to-event outcome, i.e., pooled stratified Cox regression, meta-analysis, and the proposed ODACH method. Because the pooled Cox regression estimator can be considered a gold standard, the relative bias of meta-analysis and ODACH estimates to the pooled estimate are compared to demonstrate the advantage of ODACH. The simulation was replicated 200

Site	Site 1	Site 2	Site 3	Site 4	Site 5
Total, N (%)	4078 (100)	3354 (100)	2367 (100)	2296 (100)	1920 (100)
Age ≥ 65 years, N (%)	602 (14.8)	562 (16.8)	464 (19.6)	433 (18.9)	229 (11.9)
Male, N (%)	1560 (38.3)	1142 (34)	972 (41.1)	799 (34.8)	530 (27.6)
NHW, N (%)	2510 (61.5)	1643 (49)	234 (9.9)	1406 (61.2)	889 (46.3)
Current smoker, N (%)	714 (17.5)	61 (1.8)	1 (0)	297 (12.9)	99 (5.2)
CCI, mean (S.D.)	0.86 (1.64)	0.69 (1.39)	0.97 (1.82)	0.79 (1.56)	0.75 (1.35)
Depression, N (%)	415 (10.2)	196 (5.8)	232 (9.8)	262 (11.4)	155 (8.1)
Pain, N (%)	636 (15.6)	392 (11.7)	252 (10.6)	248 (10.8)	385 (20.1)
OUD, N (%)	19 (0.5)	15 (0.4)	11 (0.5)	11 (0.5)	12 (0.6)

Table 1. Characteristics of the patients from five *OneFlorida* clinical sites. *NHW* non-Hispanic White, *CCI* Charlson comorbidity index, *OUD* opioid use disorder.

times. For simplicity of illustration, we present only the results for the estimation of coefficient β_2, as the results for the other coefficient are similar.

The OneFlorida opioid use disorder study. We evaluated the use and advantage of the proposed ODACH method by studying the association of time to an opioid use disorder (OUD) diagnosis with risk factors using RWD from the OneFlorida Clinical Research Consortium. A total of 14,015 subjects were sampled from five clinical sites and followed for 3 years after their index opioid prescription for chronic non-cancer pain (CNCP) and the time to the diagnosis of OUD was recorded. A summary of the patients' age (65 + vs. 18–65), gender (male vs. female), race (Non-Hispanic White (NHW) vs. others), smoking status (current smoker vs. others), CCI (Charlson comorbidity index[24], a weighted score of comorbid conditions), depression, and pain history measured at the index date are shown in Table 1. The rates of OUD are < 1% at all sites.

Use of experimental animals, and human participants. The use of human subject HIPAA limited data set was approved by the University of Florida (UF) Institute Review Board (IRB) under the protocol number IRB202001100. The University of Florida Federalwide Assurance number is FWA00005790. The study protocol has been reviewed by the UF IRB in accordance with the institutional and federal guidelines. Both Waivers of Informed Consent and HIPAA Waiver of Authorization were granted by the Institutional Review Board of the University of Florida.

Received: 17 July 2021; Accepted: 28 February 2022
Published online: 22 April 2022

References

1. Shore, N. Accelerating the use of electronic health records in physician practices. *N. Engl. J. Med.* **362**, 192–195 (2010).
2. Sherman, R. E. *et al.* Real-world evidence—What is it and what can it tell us. *N. Engl. J. Med.* **375**(23), 2293–2297 (2016).
3. Friedman, C. P., Wong, A. K. & Blumenthal, D. Achieving a nationwide learning health system. *Sci. Transl. Med.* **2**(57), 57cm29. https://doi.org/10.1126/scitranslmed.3001456 (2010).
4. Hripcsak, G. *et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for observational researchers. *Stud. Health Technol. Inform.* **216**, 574–578 (2015).
5. Fleurence, R. L. *et al.* Launching PCORnet, a national patient-centered clinical research network. *J. Am. Med. Inform. Assoc.* **21**(4), 578–582. https://doi.org/10.1136/amiajnl-2014-002747 (2014).
6. Schuemie, M. J., Hripcsak, G., Ryan, P. B., Madigan, D. & Suchard, M. A. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proc. Natl. Acad. Sci. U. S. A.* **115**(11), 2571–2577. https://doi.org/10.1073/pnas.1708282114 (2018).
7. Duke, J. D. *et al.* Risk of angioedema associated with levetiracetam compared with phenytoin: Findings of the observational health data sciences and informatics research network. *Epilepsia* **58**(8), e101–e106. https://doi.org/10.1111/epi.13828 (2017).
8. Duan, R. *et al.* Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *J. Am. Med. Inform. Assoc.* **27**(3), 376–385 (2020).
9. Duan, R. *et al.* Learning from local to global-an efficient distributed algorithm for modeling time-to-event data. *J. Am. Med. Inform. Assoc.* **27**(7), 1028–1036 (2020).
10. Lu, C.-L. *et al.* WebDISCO: A web service for distributed cox model learning without patient-level data sharing. *J. Am. Med. Inform. Assoc.* **22**(6), 1212–1219. https://doi.org/10.1093/jamia/ocv083 (2015).
11. Shu, D., Yoshida, K., Fireman, B. H. & Toh, S. Inverse probability weighted Cox model in multi-site studies without sharing individual-level data. *Stat. Methods Med. Res.* **29**(6), 1668–1681 (2020).
12. Huang, C. & Huo, X. A distributed one-step estimator. *Math. Program.* **174**(1), 41–76 (2019).
13. Wang, Y. *et al.* A fast divide-and-conquer sparse Cox regression. *Biostatistics* **22**(2), 381–401 (2021).
14. Luo, C. *et al.* pda: Privacy-Preserving Distributed Algorithms (v 1.2-4). *Github*. https://github.com/Penncil/pda. (Accessed on Mar 20, 2021).
15. Therneau, T., Crowson, C. & Atkinson, E. Using time dependent covariates and time dependent coefficients in the cox model. *Surviv Vignettes.* **2**, 3 (2017).
16. Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E. & Groothuis-Oudshoorn, C. G. M. Time-varying covariates and coefficients in Cox regression models. *Ann. Transl. Med.* **6**(7), 121 (2018).

17. Cai, Z. & Sun, Y. Local linear estimation for time-dependent coefficients in Cox's regression models. *Scand. Stat. Theory Appl.* **30**(1), 93–111. https://doi.org/10.1111/1467-9469.00320 (2003).
18. Tian, L., Zucker, D. & Wei, L. J. On the Cox model with time-varying regression coefficients. *J. Am. Stat. Assoc.* **100**(469), 172–183. https://doi.org/10.1198/016214504000000845 (2005).
19. Wei, L. J. The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. *Stat. Med.* **11**(14–15), 1871–1879. https://doi.org/10.1002/sim.4780111409 (1992).
20. Duan, R., Ning, Y. & Chen, Y. Heterogeneity-aware and communication-efficient distributed statistical inference. *Biometrika* **109**(1), 67–83. https://doi.org/10.1093/biomet/asab007 (2022).
21. Luo, C. *et al.* DLMM as a lossless one-shot algorithm for collaborative multi-site distributed linear mixed models. *Nature Communications* **13**(1), 1–10 (2022).
22. Tong, J. *et al.* Robust-ODAL: Learning from heterogeneous health systems without sharing patient-level data. *Pac Symp Biocomput.* **25**, 695–706 (2020). PMID: 31797639. PMCID: PMC6905508.
23. Jordan, M. I., Lee, J. D. & Yang, Y. Communication-efficient distributed statistical inference. *J. Am. Stat. Assoc.* **114**(526), 668–681. https://doi.org/10.1080/01621459.2018.1429274 (2019).
24. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *J. Chronic Dis.* **40**(5), 373–383 (1987).

Acknowledgements

This work was supported partially through a Patient-Centered Outcomes Research Institute (PCORI) Project Program Award (ME-2019C3-18315). All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee. Dr. Naj acknowledges NIH support from RF1 AG061351, U54 AG052427, and R01 AG054060.

Author contributions

R.D. and Y.C. conceived the original idea. R.D. and C.L developed the methodology and algorithm. C.L. conducted the numerical analyses and wrote the main manuscript. A.N., H.K. and J.B. improved the application study. All authors reviewed the manuscript.

Competing interests

Dr. Kranzler is a scientific advisory board member for Dicerna Pharmaceuticals, Sophrosyne Pharmaceuticals, and Enthion Pharmaceuticals; a consultant for Sobrera Pharmaceuticals; the recipient of research funding and medication supplies for an investigator-initiated study from Alkermes; a member of the American Society of Clinical Psychopharmacology's Alcohol Clinical Trials Initiative (ACTIVE Group), which over the past three years was supported by Alkermes, Dicerna, Ethypharm, Lundbeck, Mitsubishi, and Otsuka; and named as an inventor on PCT patent application #15/878,640 entitled: "Genotype-guided dosing of opioid agonists," filed January 24, 2018. All other authors have no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-09069-0.

Correspondence and requests for materials should be addressed to Y.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.