ODAC: learning from local to global - an efficient algorithm for integrating time to event healthcare system data

Rui Duan[1]; Chongliang Luo[1]; Martijn J. Schuemie[2]; Yong Chen[1]
[1]University of Pennsylvania; [2]Janssen Research and Development

Abstract

Institutional data integration becomes a trend in EHR-based studies, as it improves the confidence of study results by increasing the sample size and also reduce potential biases that come from disparities in patient population or clinical practice. To overcome the barrier of patient-level data sharing, we propose a privacy-preserving and communication-efficient distributed algorithm for Cox proportional hazard model to study the time to event outcome. The performance of the algorithm is investigated through simulation study and application to datasets from the Observational Health Data Sciences and Informatics (OHDSI). The results demonstrate that the proposed algorithm can provide relatively accurate estimation when studying rare events.

Introduction

- Multicenter analysis confers distinct advantages over single-center studies, including the ability to study rare exposures or outcomes which require larger sample sizes and discover more generalizable findings.
- Direct sharing of patient-level data across institutions is impractical due to considerations on both privacy protection and communication cost. Current distributive networks, such as the **Observational Health Data Sciences and Informatics (OHDSI)**, often rely on sharing summary statistics across sites and synthesizing evidence through meta-analyses.
- However, in the case where the outcomes or exposures are rare, or some of the sites have limited sample size, accuracy of meta-analysis is not guaranteed.
- Distributed computing algorithms, such as the GLORE (Grid Binary Logistic Regression, [1]) or the WebDISCO (a web service for distributed Cox model learning, [2]), require iteratively transferring information across sites, which leads to high communication cost.
- We propose a **One-shot Distributed Algorithm for Cox regression (ODAC)** by extending the surrogate likelihood function approach proposed by Jordan et al. (2018) [3], and the One-shot Distributed Algorithm for Logistic regression (ODAL, [4]). We investigate the performance of ODAC by comparing it to the commonly used meta-analysis and the gold standard estimator where all data are pooled together.

Methods

- Cox proportional hazard regression model.
  Let $X \in R^p$ be the risk factors, T be the time-to-event outcome of interest. $\lambda(t|X) = \lambda_0(t)\exp(\beta^T X)$ where $\lambda_0(t)$ is the baseline hazard function, and β are the regression coefficients, i.e. the log-hazard ratios. δ is the censoring indicator, i.e. $\delta = 1$ indicating an event. The observed data are $\{T_i, \delta_i, x_i\}$ for the i-th subject. For a given time t, $R(t) = \{i;\ T_i \geq t\}$, is the risk set at time t. The Cox partial log likelihood function [] can be constructed as

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

- The data are stored in K clinical sites, $N = \sum_{j=1}^{K} n_j$. For the i-th patient in the j-th site, we observe $\{T_{ij}, \delta_{ij}, x_{ij}\}$. The risk set in site j at time t is $R_j(t) = \{(i,j); T_{ij} \geq t\}$, and $R(t) = \cup_j R_j(t)$. Let $R_{ij} = R(T_{ij})$ and $\mathcal{T}_j$ the set of event times at site j. The overall partial log likelihood in (1) is

$$L(\beta) = \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{n_j} \delta_{ij} \log \frac{\exp(\beta^T x_{ij})}{\sum_{(s,k)\in R_{ij}} \exp(\beta^T x_{sk})} \triangleq \sum_{j=1}^{K} n_j L_j(\beta) \qquad (2)$$

- Assume we only have the patient level data of the "local" site, say site 1. The patient level data of other sites required for (2) are not available. we extend the surrogate likelihood approach developed in [3,4], and propose a distributed algorithm for Cox regression model.

Methods - Continued

- We propose the following surrogate likelihood function, using second-order approximation,

$$\tilde{L}_1(\beta) = L_1(\beta) + \langle \nabla L(\bar{\beta}) - \nabla L_1(\bar{\beta}), \beta \rangle + \frac{1}{2}(\beta - \bar{\beta})^T \{\nabla^2 L(\bar{\beta}) - \nabla^2 L_1(\bar{\beta})\}(\beta - \bar{\beta}), \qquad (3)$$

  where $L_1(\beta)$ is the local likelihood function defined in (2), ∇ and ∇^2 denote the first and second order gradients of a function. $\bar{\beta}$ is an initial value close to the true parameter value, e.g. a fixed-effect meta-analysis estimator.
- $L_1(\beta)$, $\nabla L_1(\bar{\beta})$ and $\nabla^2 L_1(\bar{\beta})$ can be calculated within the local site 1. $\nabla L(\bar{\beta})$ and $\nabla^2 L(\bar{\beta})$ need to be calculated distributively from all site by $\nabla^r L(\bar{\beta}) = \sum_{j=1}^{K} \frac{n_j}{N} \nabla^r L_j(\bar{\beta})$, for $r = 1, 2$. For the Cox partial likelihood, the common denominator term in (2) requires all sites to share
  $U_j(t) = \sum_{i\in R_j(t)} \exp(\bar{\beta}^T x_{ij})$, $W_j(t) = \sum_{i\in R_j(t)} exp(\bar{\beta}^T x_{ij}) x_{ij}$ and $Z_j(t) = \sum_{i\in R_j(t)} exp(\bar{\beta}^T x_{ij}) x_{ij}\, x_{ij}^T$.
  These are all aggregated information and the patients' privacy is protected.
- After constructing the surrogate likelihood function (3) at site 1, we can obtain an estimator by

$$\tilde{\beta}_1 = argmax_\beta\, \tilde{L}_1(\beta), \qquad (4)$$

  and obtain the variance $\tilde{V}_1$.
- If the other sites are also available to be a local site, we can obtain $\tilde{\beta}_j$ for each site j and further aggregate them as

$$\tilde{\beta} = (\textstyle\sum_{j=1}^{K} \tilde{V}_j^{-1})^{-1} \sum_{j=1}^{K} \tilde{V}_j^{-1} \tilde{\beta}_j. \qquad (5)$$

- The algorithm for ODAC is summarized in **Fig. 1**.

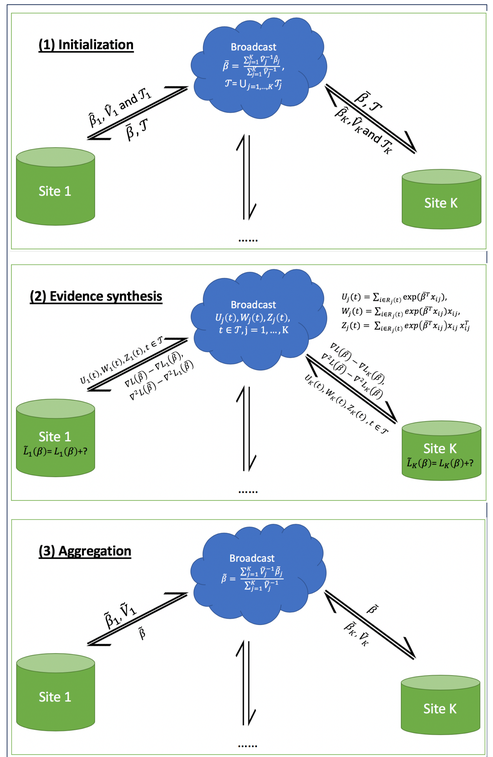

Fig. 1: Schematic illustration of the proposed ODAC algorithm.

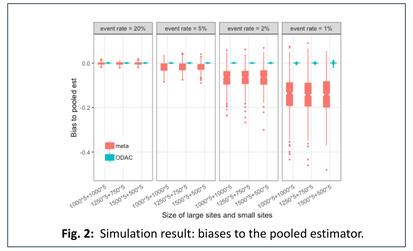

Fig. 2: Simulation result: biases to the pooled estimator.

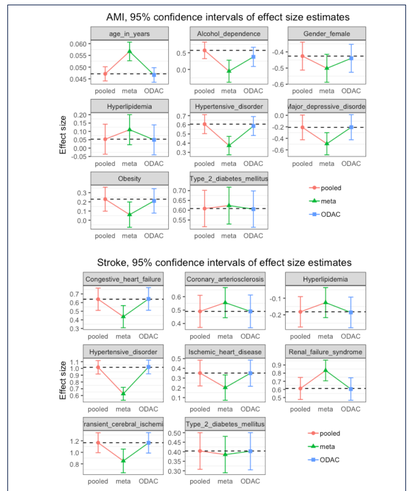

Fig. 3: Effect size estimation of risk factors of AMI and Stroke.

Simulation Design

- A pooled dataset of N=10,000 subjects are generated based on Weibull PH model, where the baseline hazard follows a Weibull-distribution with scale 200 and shape 20.
- We generate 2 covariates from i.i.d. Uniform distribution and the true effects size are set to be β=[-1, 2].
- We set the event rate as 20%, 5%, 2% and 1%.
- The pooled data are distributed to K=10 clinical sites, with 5 large and 5 small sites. We set the size of the large/small sites as 1000/1000, 1250/750, 1500/500. For the most extreme case with event rate 1% and small site sample size 500, each site has 5 events on average.
- We compare 3 methods in terms of the coefficient estimation of β:
  (a) gold standard estimator assuming all the data pooled together;
  (b) fixed-effect meta estimator, i.e. inverse variance weighted average of the estimators from each site separately;
  (c) proposed ODAC estimator.
- The biases of the estimated effect size to the gold standard are in **Fig. 2.**

Application to OHDSI network

We apply ODAC to observational studies of the risk factors of acute myocardial infarction (AMI) and stroke using claims data from 4 different databases at OHDSI [5]. The data sets and number of cases and subjects are listed in **Tab. 1.** The overall event rates for both of the two outcomes are below 1%.

- We study the association between AMI and 8 risk factors, including gender, age, obesity, alcohol dependence, hypertensive disorder, major depressive disorder, type 2 diabetes mellitus and hyperlipidemia.
- We study the association between stroke and 8 risk factors, including congestive heart failure, hypertensive disorder, ischemic heart disease, type 2 diabetes mellitus, coronary arteriosclerosis, renal failure syndrome, transient cerebral ischemia and hyperlipidemia.
- The effect size estimation with 95% confidence intervals are in **Fig. 3**.

Dataset	AMI (cases / subjects)	Stroke (cases / subjects)
Optum (Optum© De-Identified Clinformatics)	360 / 62401	319 / 62348
MDCD (IBM MarketScan® Medicaid)	438 / 60139	451 / 59861
MDCR (IBM MarketScan® Medicare)	1207 / 69788	1402 / 69164
CCAE (IBM MarketScan® Commercial)	155 / 64195	165 / 64222
Total	2160 / 256523	2337 / 255595

Tab. 1. Number of cases and subjects of AMI and Stroke in the 4 data sets at OHDSI.

Conclusions

- When outcome is rare, each site might have limited number of events, which can cause large bias in meta-analysis.
- ODAC can provide fairly accurate estimation of the time to event analysis. In both the simulation and real OHDSI data analyses, the bias to the pooled estimator is almost negligible.
- ODAC algorithm is privacy-preserving since only aggregated information is transferred across sites.
- ODAC is communication efficient since it only requires 3 rounds of broadcasting (including the initialization) among all sites.

Presenting author: Chongliang Luo (chongliang.luo@pennmedicine.upenn.edu)
Coauthors: Rui Duan (ruiduan@upenn.edu), Martijn J. Schuemie (schuemie@ohdsi.org), Yong Chen (ychen123@upenn.edu)

References
1. Wu, Y., et al., *Grid Binary LOgistic REgression (GLORE): building shared models without sharing data.* Journal of the American Medical Informatics Association (2012). 19(5): 758-64.
2. Lu CL, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, Ohno-Machado L. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. Journal of the American Medical Informatics Association (2015);22(6):1212-9.
3. Jordan, M.I., Lee J.D., and Yang Y. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association* (2018): 1-14.
4. 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. InPacific Symposium on Biocomputing. Pacific Symposium on Biocomputing 2019 (Vol. 24, pp. 30-41). NIH Public Access.
5. Hripcsak, George, et al. "Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers." *Studies in health technology and informatics* 216 (2015): 574.