PDA-OTA: Privacy-preserving Distributed Algorithms Over the Air, an OHDSI journey

Authors: Yong Chen[1], Jiayi Tong[1], Chongliang Luo[2], Lu Li[1], Yiwen Lu[1], Hai-Shuo Shu[1]

1. **Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA**
2. **Division of Public Health Sciences, Department of Surgery, Washington University in St. Louis, St. Louis, MO**

Background

Motivated by OHDSI as a next generation open science research consortium for evidence generation using federated real-world data, over the last 5 years, we have been developing a tool set of communication-efficient and heterogeneity-aware distributed algorithms, as well as user-centered software and web-based secure data sharing infrastructure tailored for OHDSI users. At this year's OHDSI annual symposium, we are ready to formally release our software package – PDA[1,2], and its user centered communication system for distributed learning – PDA-OTA[3,4].

With the increasing availability of real-world data including EHR data and claims data, it is important to effectively integrate and generate evidence from multiple data sources to improve the generalizability and reproducibility of scientific discovery. However, practical challenges remain in evidence generation using real-world data, including data privacy, high dimensionality of features, non-random missingness, and heterogeneity across different datasets.

The overarching goal of PDA is to facilitate efficient multi-institutional data analysis without sharing individual patient-level data (IPD), while addressing the aforementioned challenges. PDA enables a broad range of multi-site analyses, including association studies, predictive modeling, causal inference, and counterfactual analyses. Currently PDA includes distributed algorithms for binary outcomes including logistic regression[5–8] and modified Poisson regression[9], time to event outcomes including Cox model[10,11], count outcomes including Poisson model[12,13] and hurdle model [10], (Generalized) linear mixed models[14–16], penalized regressions for high dimensional features[17], and general heterogeneity-aware distributed inference[18].

Our PDA framework (**Figure 1**) was designed with the following features:

- Communication-efficient: only requires the collaborating sites to send aggregated data to a coordinating site once or few times (i.e., non-iterative).
- Privacy-preserving: only requires sharing of aggregated data/summary statistics
- Heterogeneity-aware: properly accounts for between-site heterogeneity
- Accurate: provides accurate estimation of the parameters of interest, especially for the analysis of rare diseases

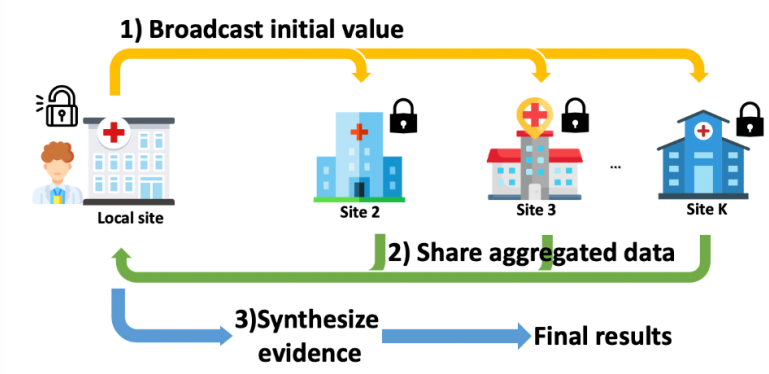

Figure 1. Schematic diagram of the proposed PDA framework

PDA algorithms have been applied to studies of long COVID among children[9], characterizing impacts of risk factors of still birth[5,6,19], opioid use disorder (OUD)[11,17,20], pediatric avoidable hospitalization[12], serious adverse event of colorectal cancer[12], and trajectories of Alzheimer's disease (AD)[21], hospitalization of COVID patients[12–14,16], mortality of COVID patients[8], risk factors of acute myocardial infarction (AMI)[10], kidney graft failure[15,22] using data from **more than 30 million patients.**

Methods

Algorithms under PDA: we have developed different algorithms tailored for different types of outcomes, including:

- ODAL/Robust-ODAL[5,6,19]: One-shot Distributed Algorithm for binary outcomes with Logistic regression model
- ODAC/ODACH[10,11]: One-shot Distributed Algorithm for time to event outcomes with Cox regression model with Heterogeneous multi-site data
- ODAP[13]: One-shot Distributed Algorithm for count outcomes with Poisson regression model
- ODAH[12]: One-shot Distributed Algorithm for zero inflated count outcomes with Hurdle regression model
- ODAP-B[9]: One-shot Distributed Algorithm for binary outcomes with modified Poisson regression model
- Heterogeneity-aware and communication-efficient distributed statistical inference[18]
- DLMM[14]: Lossless Distributed Linear Mixed Model for continuous outcomes
- dPQL[16]: Lossless Distributed Penalized Quasi-likelihood algorithm for Generalized Linear Mixed Model for outcomes in exponential family
- dCLR[8]: One-shot Distributed Algorithm for Conditional Logistic Regression Model for binary outcome, accounting for heterogeneous baseline characteristics of patients across sites
- ADAP[17]: One-Shot Distributed Algorithm for fitting Penalized regression model for high-dimensional heterogeneous data for outcomes in exponential family

- dGEM[15]: Decentralized algorithm for Generalized mixed Effect Models with the Application in Hospital Profiling for outcomes in exponential family
- dGEM-disparity[22]: Decentralized algorithm for Generalized mixed Effect Models for Disparity quantification for outcomes in exponential family

Implementation and platform for secure sharing of aggregated data: To implement the algorithms under PDA framework and facilitate collaborative studies, we developed a web-based software for secured sharing of aggregated data for multi-site studies using our privacy-preserving distributed algorithms. We termed this software as PDA-OTA, which stands for PDA over the air. PDA-OTA is a unified platform that facilitates national and international collaborations requiring secure sharing of aggregated data across collaborating sites. PDA-OTA synchronizes project status, offers cloud-based SFTP, and generates model-specific tasks for streamlined implementations. It provides a user-centered platform for two types of users: the project lead and project participants. PDA-OTA also allows users to invite participating sites to collaborate, upload aggregated data, track project status, receive automated email notifications, and generate project summaries automatically.

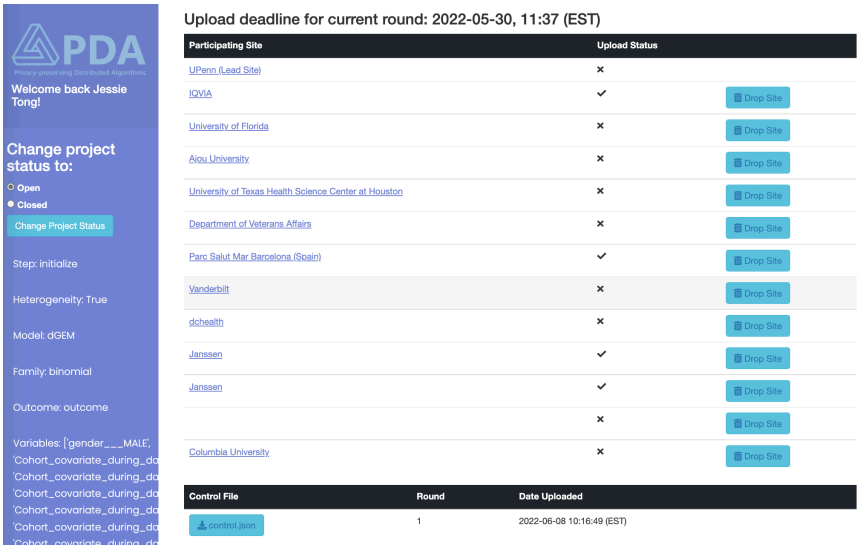

Figure 2. PDA-OTA platform

Results

The algorithms under PDA framework have been applied to a variety of studies by collaborating with national and international partners (**Figure 3**):

- long COVID among children[9]
- risk factors of still birth[5,6,19]
- opioid use disorder (OUD)[11,17,20]
- pediatric avoidable hospitalization[12]
- serious adverse event of colorectal cancer[12]

- trajectories of Alzheimer's disease (AD)[21]
- hospitalization of COVID patients[12–14,16]
- mortality of COVID patients[8]
- risk factors of acute myocardial infarction (AMI)[10]
- kidney graft failure[15,22]

Figure 3. Our partners

Conclusion

- PDA provides ***a solution for next generation data sharing for collaborative modeling***
- PDA includes major algorithms for association studies and predictive modeling, and will soon include ***distributed cluster analysis, federated causal inference,*** **and** ***federated transfer learning*** that are tailored for OHDSI studies
- Please subscribe to our twitter (**@PennCIL_lab**) and YouTube channel (Penncil Lab) (https://www.youtube.com/channel/UCGGz4o-1kMNY23k4xdA64Ew) for more updates on PDA and PDA-OTA.

References/Citations

1. CRAN - Package pda. https://cran.r-project.org/web/packages/pda/index.html.
2. PDA website. https://pdamethods.org/.
3. PDA-OTA. https://pda-ota.pdamethods.org/login.
4. Penncil/pda-ota. https://github.com/Penncil/pda-ota.
5. Duan, R., Boland, M. R., Moore, J. H. & Chen, Y. ODAL: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. in *Biocomputing 2019* 30–41 (WORLD SCIENTIFIC, 2018). doi:10.1142/9789813279827_0004.
6. Tong, J. *et al.* Robust-ODAL: Learning from heterogeneous health systems without sharing patient-level data. in *Pacific Symposium on Biocomputing* vol. 25 695–706 (World Scientific Publishing Co. Pte Ltd, 2020).
7. Duan, R. *et al.* Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *Journal of the American Medical Informatics Association* **27**, 376–385 (2020).

8. Tong, J. *et al.* Distributed learning for heterogeneous clinical data with application to integrating COVID-19 data across 230 sites. *npj Digital Medicine* **5**, 1–8 (2022).
9. Penncil/ODAP-B. https://github.com/Penncil/ODAP-B.
10. Duan, R. *et al.* Learning from local to global-an efficient distributed algorithm for modeling time-to-event data. *Journal of the American Medical Informatics Association* **27**, 1028–1036 (2020).
11. Luo, C. *et al.* ODACH: a one-shot distributed algorithm for Cox model with heterogeneous multi-center data. *Sci Rep* **12**, 1–8 (2022).
12. Edmondson, M. J. *et al.* An efficient and accurate distributed learning algorithm for modeling multi-site zero-inflated count outcomes. *Sci Rep* **11**, 1–17 (2021).
13. Edmondson, M. J. *et al.* Distributed Quasi-Poisson Regression Algorithm for Modeling Multi-Site Count Outcomes in Distributed Data Networks. *Journal of Biomedical Informatics* 104097 (2022).
14. Luo, C. *et al.* DLMM as a lossless one-shot algorithm for collaborative multi-site distributed linear mixed models. *Nat Commun* **13**, 1–10 (2022).
15. Penncil/dGEM. https://github.com/Penncil/dGEM.
16. Luo, C. *et al.* dPQL: a lossless distributed algorithm for generalized linear mixed model with application to privacy-preserving hospital profiling. *Journal of the American Medical Informatics Association* ocac067 (2022) doi:10.1093/jamia/ocac067.
17. Penncil/ADAP. https://github.com/Penncil/ADAP.
18. Duan, R., Ning, Y. & Chen, Y. Heterogeneity-aware and communication-efficient distributed statistical inference. *Biometrika* (2021) doi:10.1093/biomet/asab007.
19. Duan, R., Boland, M. R., Moore, J. H. & Chen, Y. ODAL: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. *Pac. Symp. Biocomput.* **24**, 30–41 (2019).
20. Tong, J. *et al.* Identifying Clinical Risk Factors for Opioid Use Disorder using a Distributed Algorithm to Combine Real-World Data from a Large Clinical Data Research Network. *AMIA Annual Symposium Proceedings* **2020**, 1220 (2020).
21. Duan, R. *et al.* Leverage Real-world Longitudinal Data in Large Clinical Research Networks for Alzheimer's Disease and Related Dementia (ADRD). *AMIA Annu Symp Proc* **2020**, 393–401 (2020).
22. Penncil/dGEM-disparity. https://github.com/Penncil/dGEM-disparity.