Expanding the reach of EHR through data integration
Yong Chen¹, Rui Duan², Chongliang Luo¹, Jiayi Tong¹, Mackenzie Edmondson¹, Hua Xu³, Patrick Ryan⁴, Martijn Schuemie⁵
¹ University of Pennsylvania; ² Harvard University; ³ UTHSC; ⁴ Columbia University; ⁵ Janssen Research and Development LLC

Research Question
Integration of EHR data across institutes
- Electronic Health Record (EHR) data are playing an increasing role in generating real-world evidence (RWE) to support clinical decision making.
- Multi-site study using EHR data from clinical research networks is increasing popular due to larger sample size and broader population.
- Sharing individual patient data (IPD) across clinical sites is logistically challenging due to privacy concerns.
- Distributed algorithms for various models have been proposed in the literature, but most of them require iterative communications across sites.

How PDA works
1. Initial building of a predictive model
2. Transfer initial model coefficients to other sites
3. Collaborating Data Site 2
4. Calculate aggregated data within each site
5. (4) Transfer aggregated data to the lead data site
6. Refine predictive model based on aggregated data from sites 2, ..., K
7. Deliverables: refined predictive model

Tasks for lead data site:
1. Train initial predictive model and obtain model coefficients \( \hat{\beta} \)
2. Send \( \hat{\beta} \) to collaborating sites 2, ..., K
3. Incorporate outputs (aggregated data) from sites to refine initial predictive model

Tasks for each collaborating data site k:
1. Run function from lead data site
2. Input 1 is the initial model coefficients \( \hat{\beta} \) from lead data site
3. Input 2 is data from collaborating data site k
4. Output: a p x 1 vector, rounded to the 2nd decimal (p is the # of predictors in the predictive model)

Use cases & Collaborations
- OHDSI’s global research community
- OHDSI – drug adverse events
- PEDSNet – pediatric Crohn’s disease
- OneFlorida – Alzheimer’s disease/opioid use disorder

Available of PDA
https://github.com/Penncl/pda

Proposed Solution: PDA
- A toolbox of Privacy-preserving Distributed Algorithms that conduct distributed learning and inference for various models.
- Aims to facilitate efficient multi-institutional data analysis without sharing IPD.
- Compared to existing distributed algorithms, has the following features:
  - **Accurate**: provide estimates on par with the pooled estimator
  - **Safe**: only require aggregated data (AD)
  - **Fast or NICE**: Non-Iterative and Communication-Efficient
  - **Heterogeneity-aware**: handles between-site heterogeneity
- PDA outperforms meta-analysis methods in many settings such as pharmacovigilance applications

Specific Methods Developed & Validated
Binary outcomes:
- ODAL: One-shot Distributed Algorithm for Logistic Regression Model [Duan et al., 2019, 2020]
- ODAL-Robust, ODAL-H: One-shot Distributed Algorithm to Handle Heterogeneity across Clinical Sites [Tong et al. 2019; Tong et al.]

Continuous outcomes:
- ODALMM: One-shot Distributed Algorithm for Linear Mixed Model [Luo et al.]

Time-to-event outcomes:
- ODAC: One-shot Distributed Algorithm for Cox Proportional Hazards Model [Duan et al., 2020]

Count outcomes:
- ODAP: One-shot Distributed Algorithm for Poisson model [Edmondson et al.]
- ODAX: One-shot Distributed Algorithm to Handle Zero-inflated Counts using Hurdle Model [Edmondson et al.]

Selected Publications
4. Tong et al. Robust-ODAL: Learning from heterogeneous health systems without sharing patient-level data. InPSB 2020

Funding
- National Library of Medicine
- National Institute of Allergy and Infectious Diseases
- PCORI