Expanding the reach of EHR through data integration

Yong Chen[1], *Rui Duan*[2], *Chongliang Luo*[1], *Jiayi Tong*[1], *Mackenzie Edmondson*[1], *Hua Xu*[3], *Patrick Ryan*[4], *Martijn Schuemie*[5]

[1] University of Pennsylvania; [2] Harvard University; [3] UTHealth; [4] Columbia University; [5] Janssen Research and Development LLC

OHDSI | PennCIL A Computing · Inference · Learning lab at University of Pennsylvania | DEPARTMENT of BIOSTATISTICS EPIDEMIOLOGY & INFORMATICS | Penn Medicine

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.

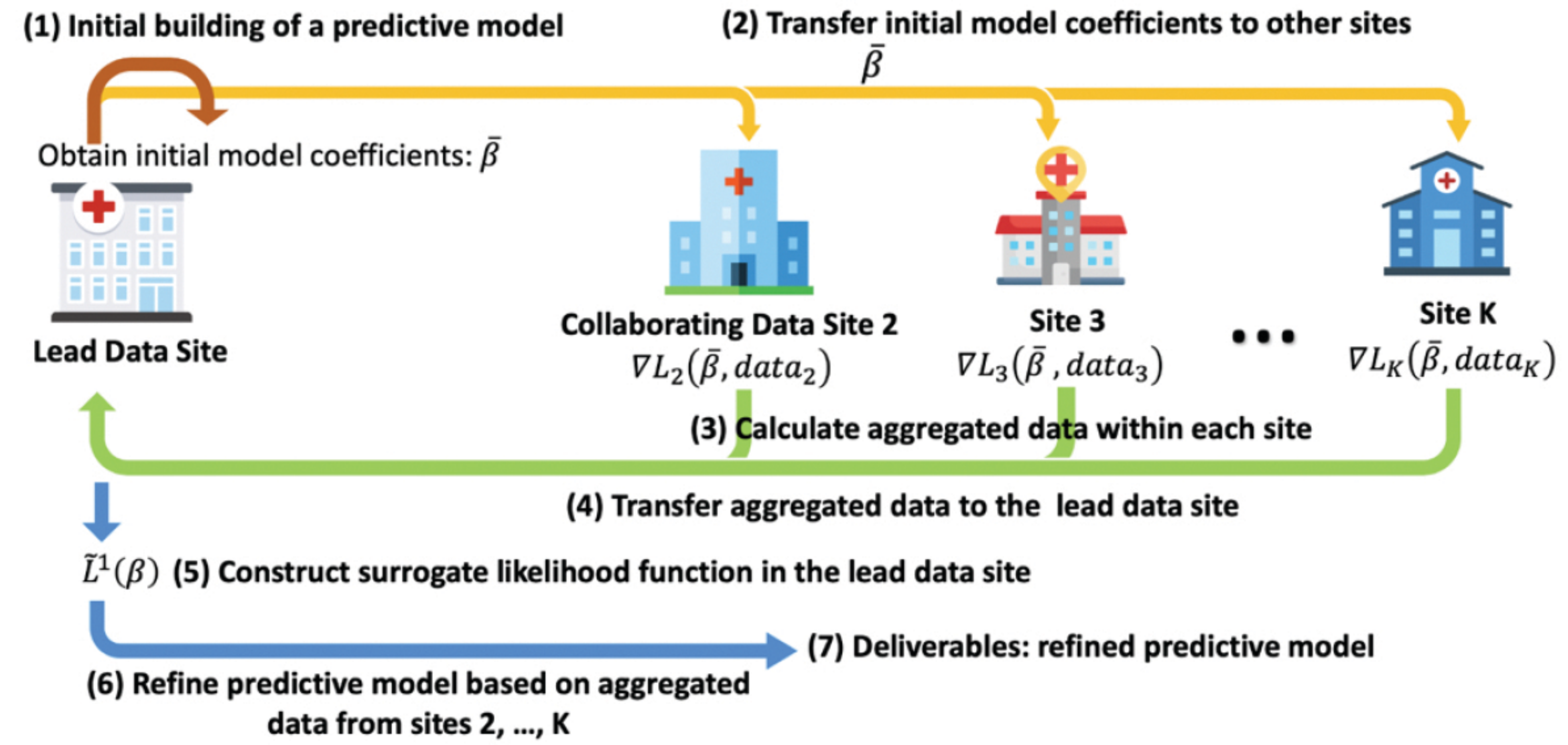
Proposed Solution: PDA

- 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

Funding

NIH National Library of Medicine | NIH National Institute of Allergy and Infectious Diseases | pcori

How PDA works



Tasks for **lead data site**:	Task for each **collaborating data site k**:
(1) Train initial predictive model and obtain model coefficients $\bar{\beta}$ **(2)** Send $\bar{\beta}$ to collaborating sites $2, \cdots, K$ **(5)** Incorporate outputs (aggregated data) from sites to refine initial predictive model **(6)** Refine predictive model	**(3)** Run function from lead data site "∇L_k(Input 1, Input 2)" **(4)** Send output of the function to lead data site • Input 1 is the initial model coefficients $\bar{\beta}$ from lead data site Input 2 is data from collaborating data site k • Output: a $p \times 1$ vector, rounded to the 2nd decimal (p is the # of predictors in the predictive model)

Availability of PDA

https://github.com/Penncil/pda

PennCIL | Penncil

pda

PDA: Privacy-preserving Distributed Algorithm

C++ 1 0 1 0

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.]
- ODAH: One-shot Distributed Algorithm to Handle Zero-inflated Counts using Hurdle Model [Edmondson et al.]

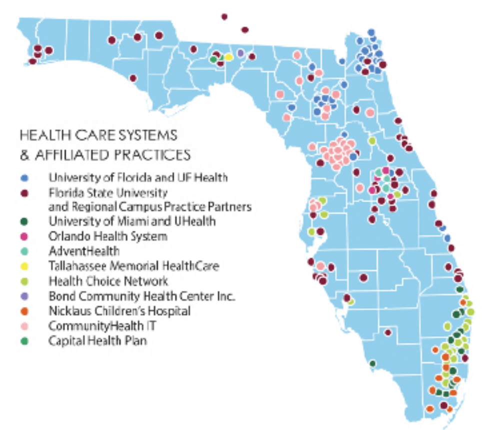
Use cases & Collaborations

OHDSI's global research community

OHDSI – drug adverse events

Nemours | CHOP The Children's Hospital of Philadelphia | Nationwide Children's | Boston Children's Hospital | PEDSnet | Cincinnati Children's | Children's Hospital · St. Louis | Children's Hospital Colorado | Seattle Children's

PEDSnet– pediatric Crohn's disease



OneFlorida – Alzheimer's disease/opioid use disorder

Selected Publications

1. Duan et al. ODAL: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. *InPSB 2019.*

2. Duan et al. Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *JAMIA 2020 Mar.*

3. Duan et al. Learning from local to global: An efficient distributed algorithm for modeling time-to-event data. *JAMIA 2020 July.*

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