Advancing Interpretable Regression Analysis for Binary Data: A Novel Distributed Algorithm Approach

Jiayi (Jessie) Tong, Assistant Professor
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
December 3, 2024
OHDSI Community Call

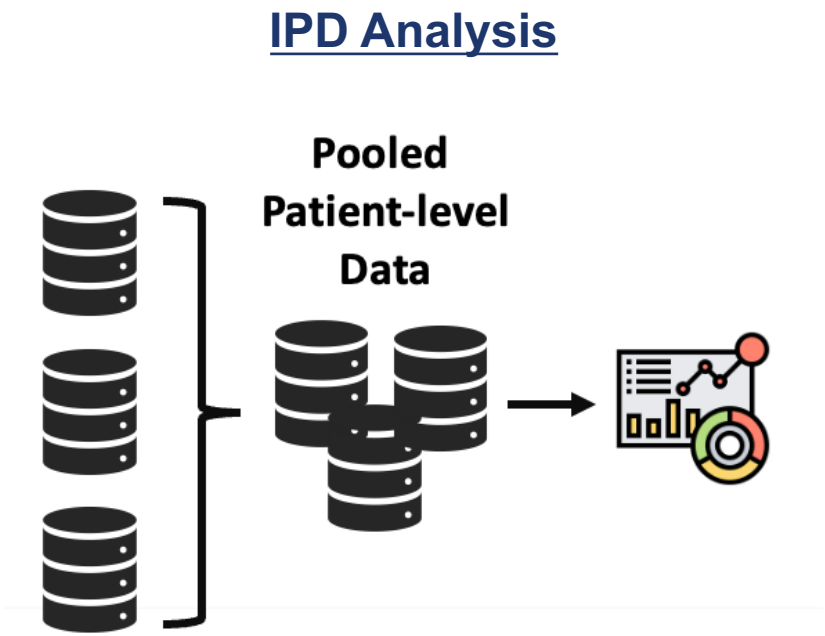
Background

- **Rare diseases -> rare binary outcome**, the sparse data problem is a significant challenge.

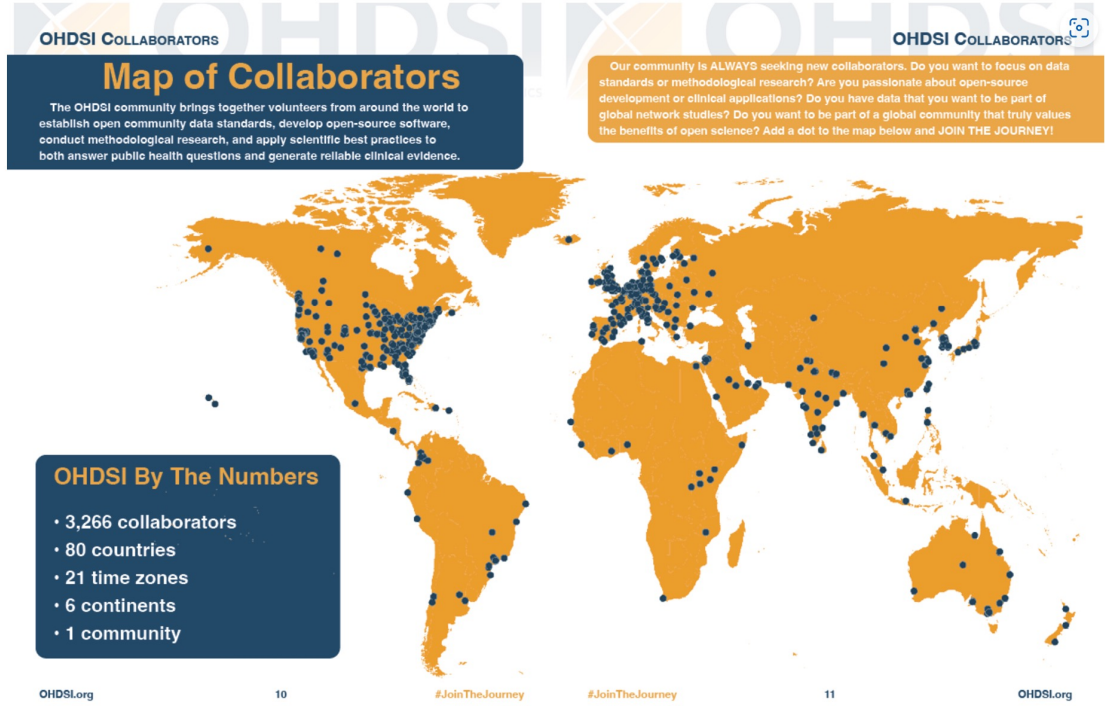
- The **lack of sufficient cases** (e.g., patients with disease) in the data leads to **biased estimates** of the effect of a treatment or a medication in observational studies.

Background

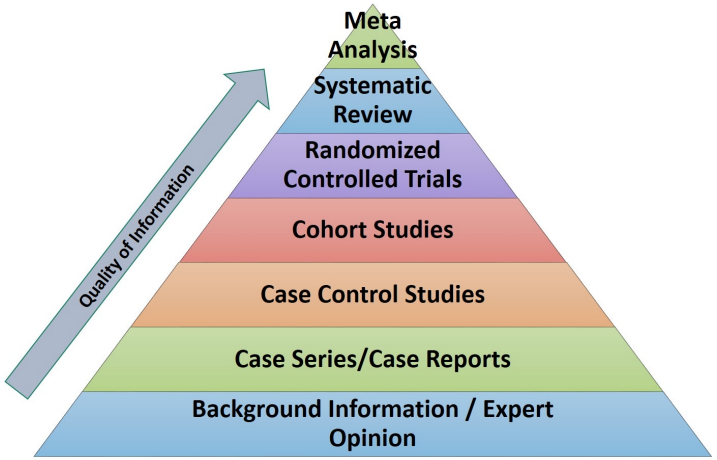
- Multi-site individual patient-level data (IPD) analysis increases the number of cases



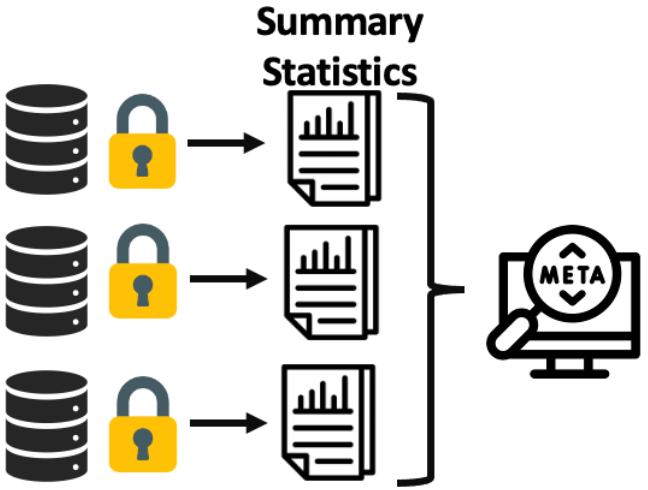
- However, it is not feasible when the individual-level data from different studies cannot be shared.



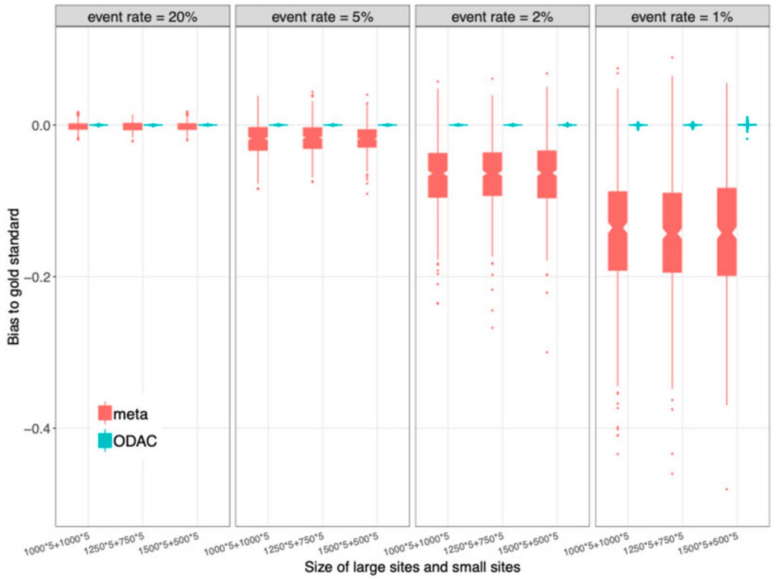
Meta-analysis Method



Meta-analysis
(Divide and conquer)



Meta-analysis has low accuracy for rare data



Federated Learning Algorithms

Privacy-Preserving Federated Learning Algorithms

- Enables fitting statistical models in a federated manner
- Requires summary statistics, instead of IPD
- Ensures data privacy and security

Federated Learning Algorithms for Rare Binary Data

Existing Federated Learning Algorithms for Binary Data

Logistic Regression Model -> Odds Ratio (OR)

The choice between **Relative Risk (RR)** and OR has been debated in the literature, with RR being preferred in most prospective studies due to **its collapsibility** and **better interpretation**.

Poisson Regression Model -> Relative Risk (RR)

American Journal of Epidemiology

Vol. 159, No. 7
Printed in U.S.A.
DOI: 10.1093/aje/kwh090

A Modified Poisson Regression Approach to Prospective Studies with Binary Data

Guangyong Zou[1,2]

[1] Robarts Clinical Trials, Robarts Research Institute, London, Ontario, Canada.
[2] Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada.

Proposed Method

ODAP-B: **O**ne-shot **D**istributed Algorithm of Modified **P**oisson Regression for **B**inary Data

- The surrogate likelihood (SL) approach (Jordan et al 2018 JASA)

For some given initial value $\bar{\beta}$, consider Taylor expansion for the multi-site likelihood

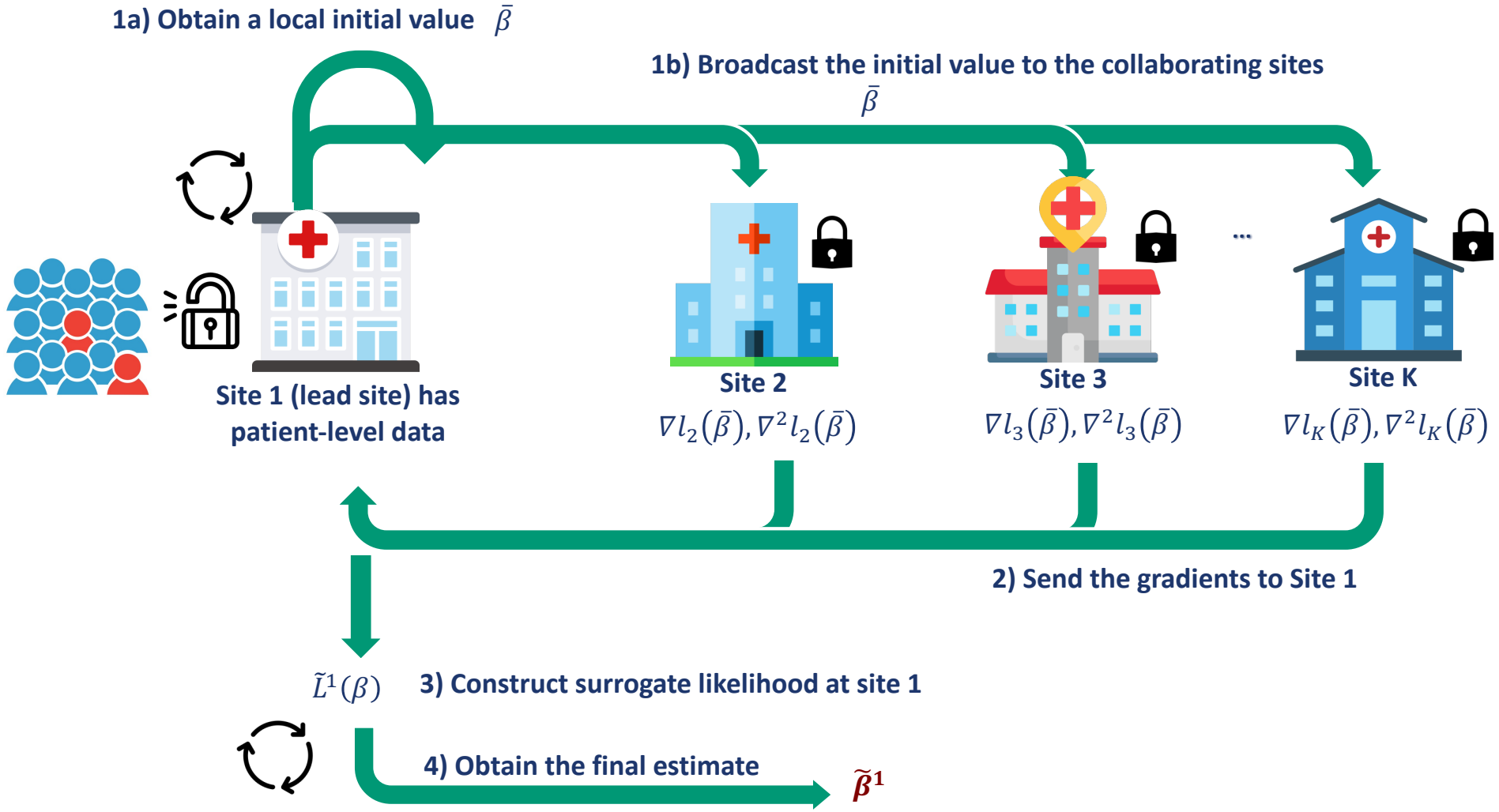
$$L(\beta) = L(\bar{\beta}) + \nabla L(\bar{\beta})^T(\beta - \bar{\beta}) + \sum_{t=2}^{\infty} \frac{1}{t!} \nabla^t L(\bar{\beta})(\beta - \bar{\beta})^{\otimes t}$$

For some given initial value $\bar{\beta}$, consider Taylor expansion for the local likelihood

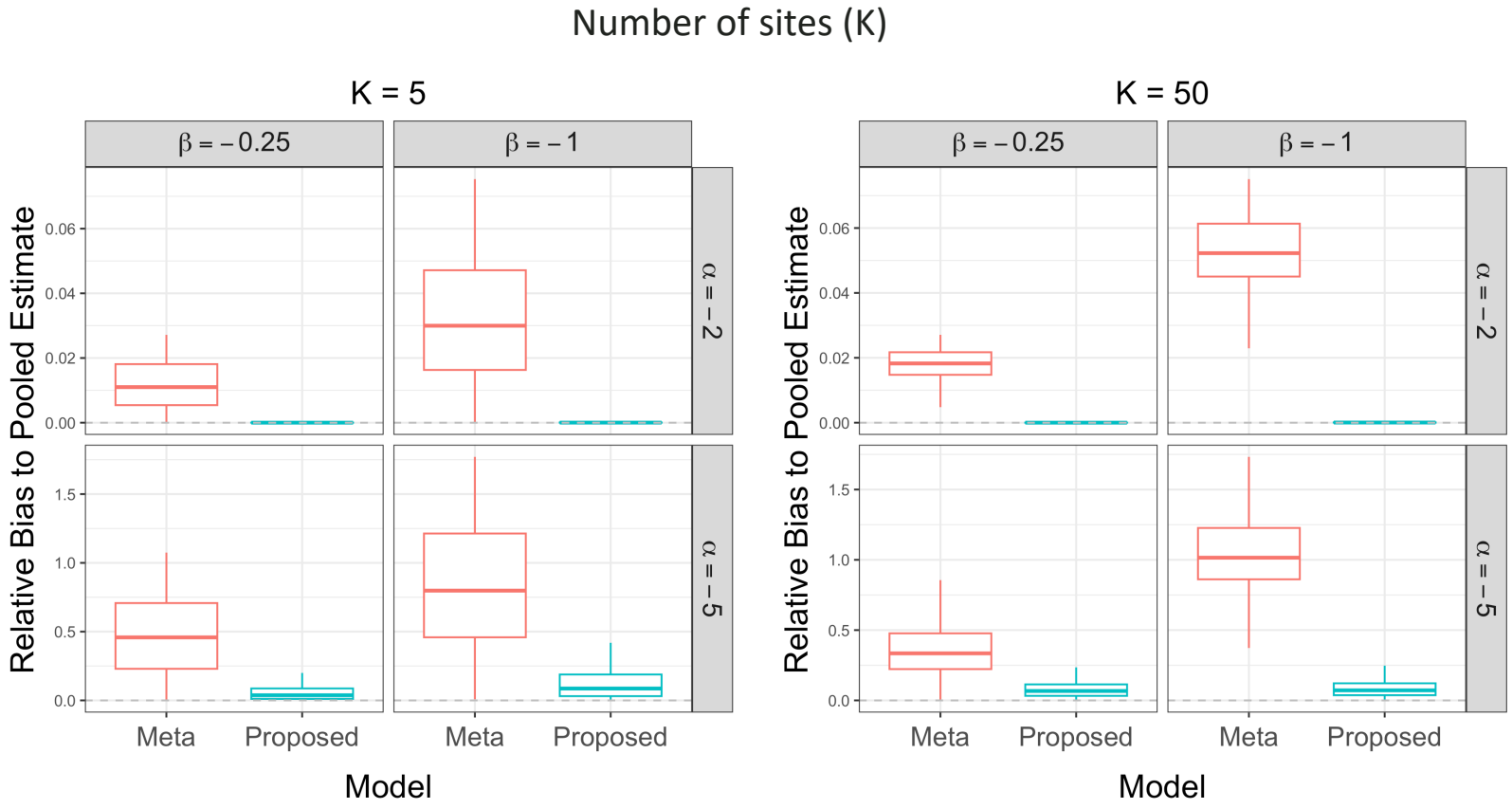
$$L_1(\beta) = L_1(\bar{\beta}) + \nabla L_1(\bar{\beta})^T(\beta - \bar{\beta}) + \sum_{t=2}^{\infty} \frac{1}{t!} \nabla^t L_1(\bar{\beta})(\beta - \bar{\beta})^{\otimes t}$$

Surrogate likelihood $\tilde{L}^1(\beta) = L_1(\beta) + \{\nabla L(\bar{\beta}) - \nabla L_1(\bar{\beta})\}^T \beta$

$$\nabla L(\bar{\beta}) = \frac{1}{K} \sum \nabla L_j(\bar{\beta}); \quad \tilde{\beta} = \operatorname{argmax}_\beta \tilde{L}(\beta)$$


1a) Obtain a local initial value $\bar{\beta}$
1b) Broadcast the initial value to the collaborating sites
$\bar{\beta}$
Site 1 (lead site) has patient-level data
Site 2
$\nabla l_2(\bar{\beta}), \nabla^2 l_2(\bar{\beta})$
Site 3
$\nabla l_3(\bar{\beta}), \nabla^2 l_3(\bar{\beta})$
...
Site K
$\nabla l_K(\bar{\beta}), \nabla^2 l_K(\bar{\beta})$
2) Send the gradients to Site 1
$\tilde{L}^1(\beta)$
3) Construct surrogate likelihood at site 1
4) Obtain the final estimate
$\tilde{\boldsymbol{\beta}}^{\mathbf{1}}$

Simulation Study



Real-world Data Application

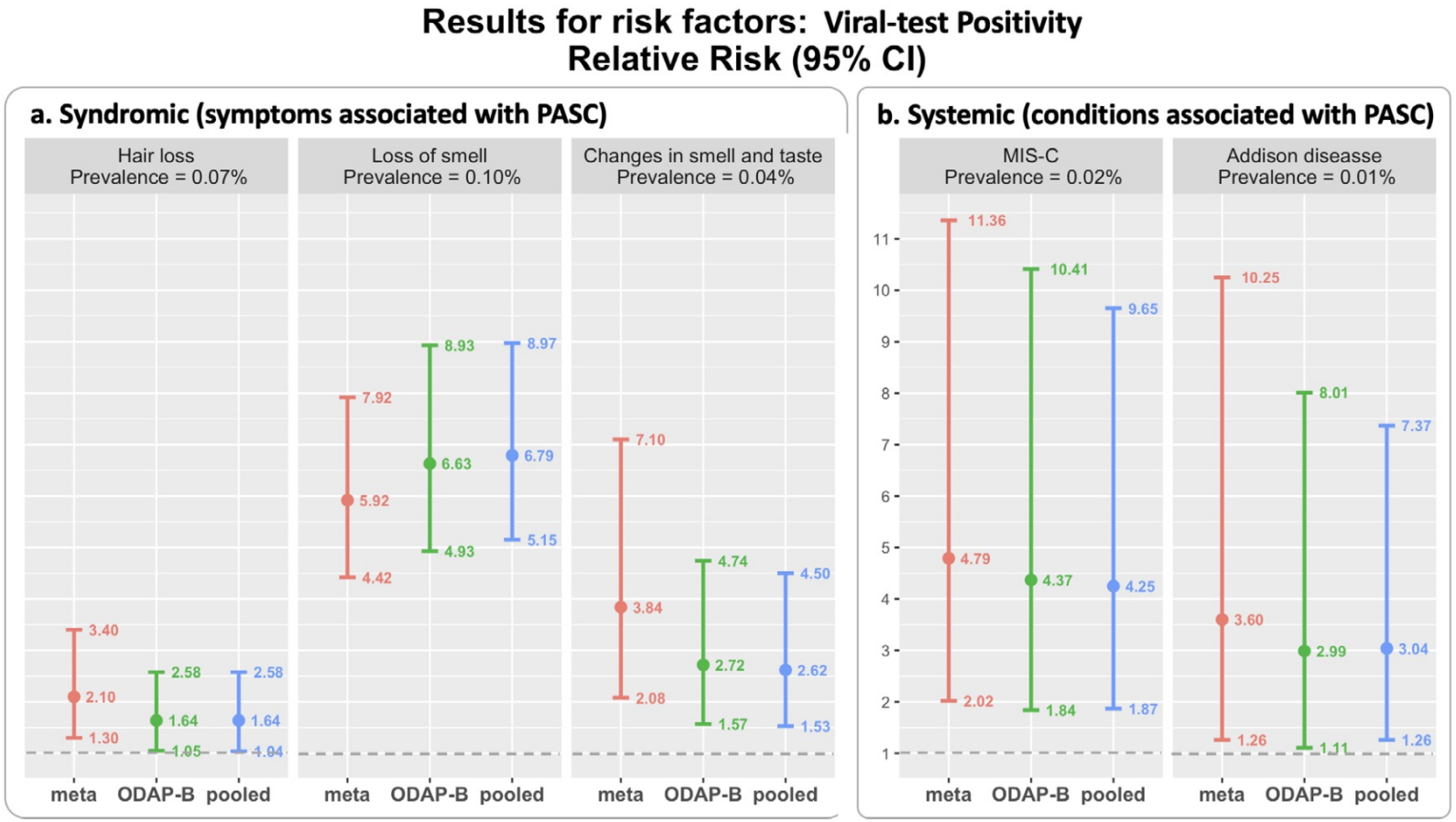
Investigate the relationship between **COVID-19 viral** (SARS-CoV-2 polymerase chain reaction [PCR] or antigen) **test positivity** and the **symptoms and conditions associated with Long-COVID** in **children**

- Use case 1: Centralized data from PEDSnet
  - Nine children's hospitals across the nation
  - Sample size: 184,501

- Use case 2: Decentralized datasets
  - 12 sites: PEDSnet (9), OHDSI (2), OneFlorida (3)
  - Sample size: 452,160

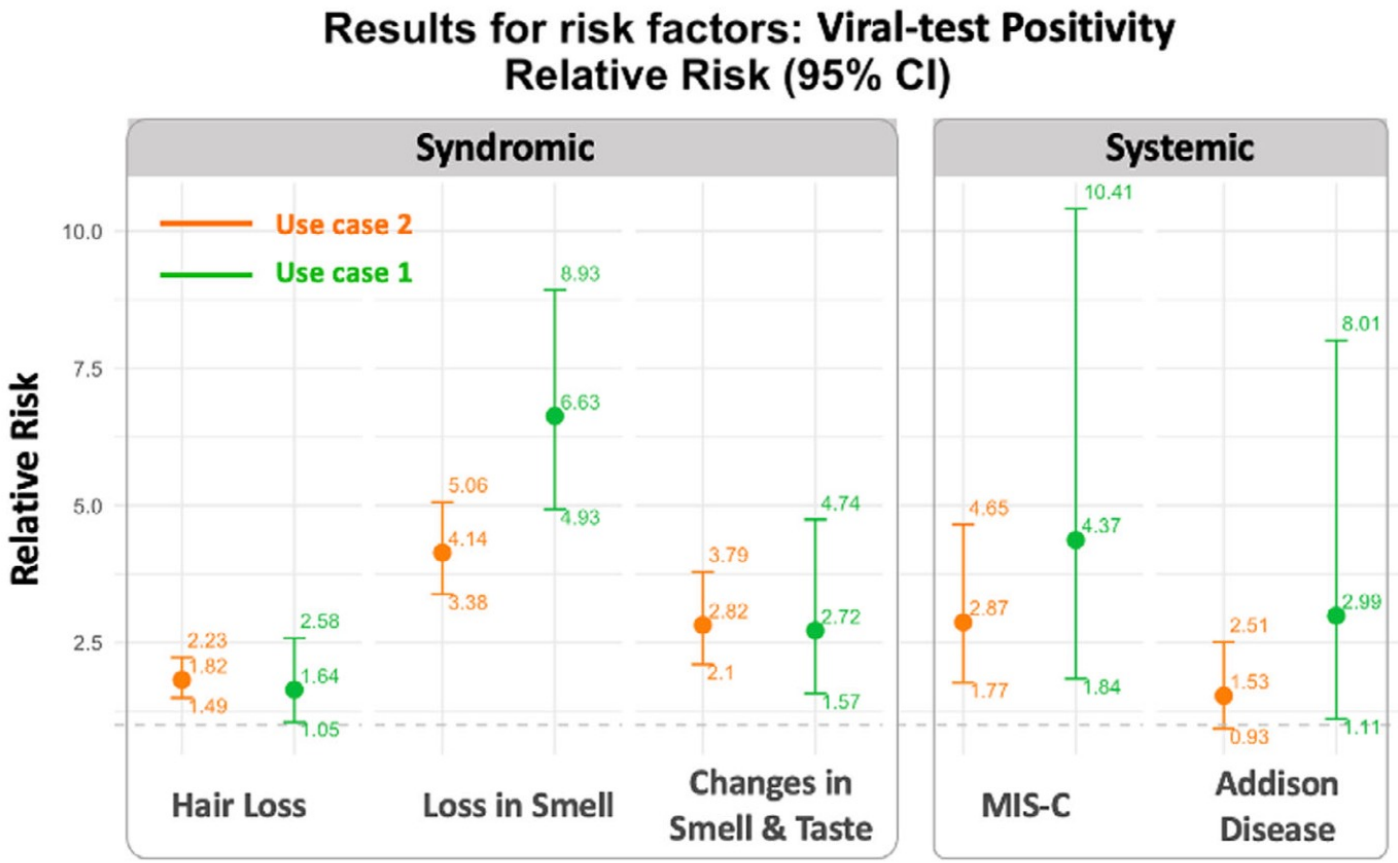
Confounding variables adjusted in the regression model:
age at cohort entrance, sex (male vs. female), race, COVID-19 testing location, diagnosis date of the outcome (i.e., PASC conditions), Pediatric Medical Complexity Algorithm (PMCA).

Use Case 1 Result



MIS-C: a multisystem inflammatory syndrome in children

Use Case 1 & Use Case 2 Results Comparison



Summary

- ODAP-B is an effective federated learning algorithm for **Poisson regression** to study **rare binary outcome**.
- ODAP-B provides inference on **adjusted relative risk** with robust variance estimator.
- ODAP-B is **easy to implement** and **applicable** to analyze multi-site data

PDA R Package: 13300+ downloads since 2020

PDA Github Page: https://github.com/Penncil/pda

PDA
Privacy-preserving Distributed Algorithms

PDA website: https://pdamethods.org/

PDA-OTA

PDA-OTA: https://pda-ota.pdamethods.org/

Penn security office certified

Thank you