Distributed Learning from Electronic Health Records Across Multiple Sites for Zero-Inflated Count Outcomes

Mackenzie Edmondson[a], Chongliang Luo[a], Rui Duan[b], Mitchell Maltenfort[c], Justine Shults[a], Patrick Ryan[d], Christopher Forrest[c], Yong Chen[a]

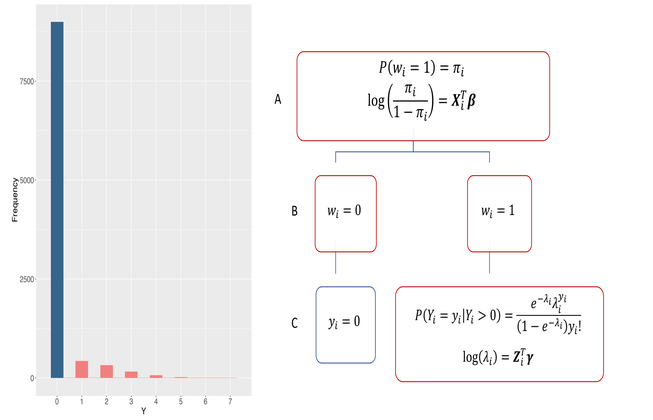
a. Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA
b. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA
c. Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA
d. Janssen Research and Development, Titusville, NJ

Background

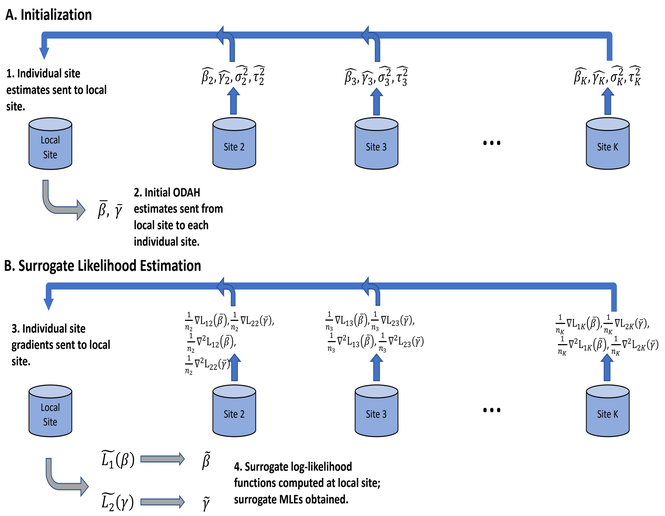
- Privacy concerns often prevent patient-level data sharing in multi-site studies, increasing popularity of privacy-preserving analysis methods.
- Meta-analysis is commonly used in OHDSI studies but has been shown to produce inaccurate parameter estimates in some settings (Duan et al. 2019).
- Count outcomes common in observational health data studies, many of which are zero-inflated and overdispersed.
- To our knowledge, no existing approach for modeling zero-inflated count data in distributed manner.
- We offer **ODAH**, a privacy-preserving, communication-efficient algorithm for performing distributed Poisson-Logit hurdle regression to model multi-site zero-inflated count outcomes.

Methods

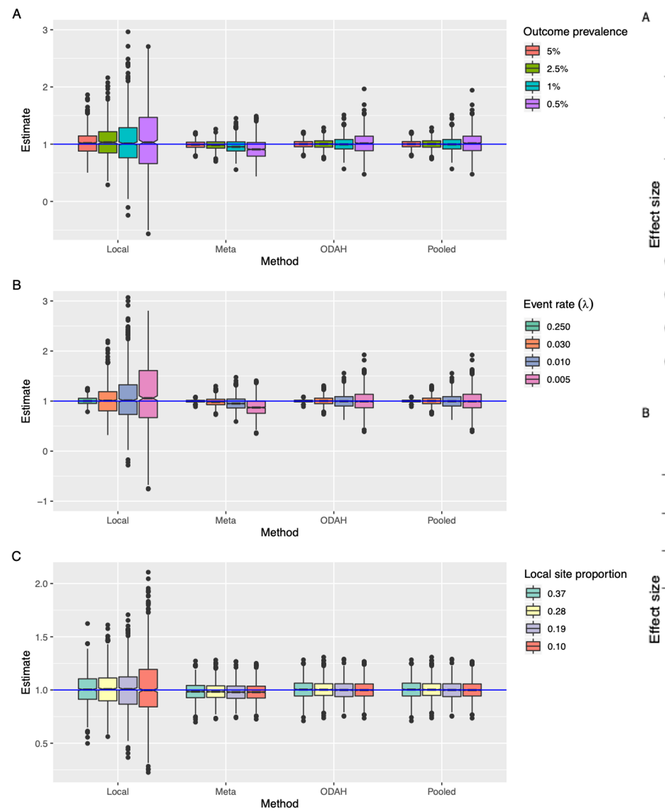
- Poisson-Logit Hurdle model: two-part (logistic regression, zero-truncated Poisson regression) count model.
- **Why use a hurdle model rather than a zero-inflated model?**
  - Two-stage sequential model improves inference and clinical interpretation in many applications
  - Logistic and zero-truncated Poisson components are information orthogonal (improves computational efficiency)

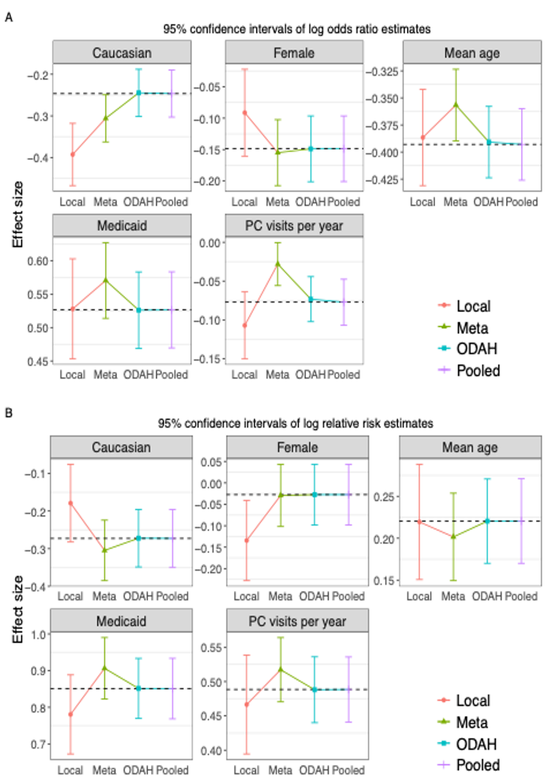


- In distributed data setting, assume we only have access to patient-level data at local site and aggregate information from all other (collaborating) sites.
- ODAH maximizes **surrogate log-likelihood function** (Jordan et al. 2019) to obtain parameter estimates, approximating estimates from pooled data analysis (gold standard, access to all patient-level data).
- Aggregate information from collaborating sites: first- and second- order gradients of log-likelihood function.
- Two non-iterative rounds of communication: initialization (meta-analysis) and surrogate likelihood estimation (sending gradients to local site).
- Evaluated ODAH in simulations (settings featuring low outcome prevalence and event rate) and data analysis (modeling total avoidable hospitalizations given EHR variables at 6 Children's Hospital of Philadelphia (CHOP) primary care sites.



Results





Simulations (left panel):

- No discernable difference between methods in estimating logistic component coefficients. Results shown are for zero-truncated Poisson component.
- Across all simulation settings, ODAH estimates exhibited relative bias to pooled estimates of less than 0.1%.
- Meta-analysis estimates exhibited bias up to 12.7%, with greater bias for smaller event rates (across settings left to right in panel B, event rate (λ in untruncated Poisson distribution) decreases from 0.25 to 0.005.

CHOP Data Analysis (right panel):

- ODAH relative bias to pooled estimates ranged from 0.08% to 5.02% for the logistic component and less than 0.5% for the zero-truncated Poisson component.
- Meta-analysis relative bias ranged from 4.15 to 63.6% for the logistic component and from 5.89% to 11.7% for the zero-truncated Poisson component.

Conclusions

- Through extensive simulations and a real-world EHR application, our method (ODAH) consistently produced parameter estimates comparable to and sometimes better than those produced by meta-analysis.
- ODAH's utility especially evident in settings featuring severely zero-inflated count outcome and very low event rate.
- We believe ODAH is a worthy alternative to meta-analysis when modeling multi-site zero-inflated count outcomes, and we look forward to collaborating with the PLP and PLE teams to have our method compatible with the OMOP CDM and integrated for future use within the OHDSI network.

Contact: macjohn@pennmedicine.upenn.edu