ODAP-B: A One-shot Distributed Algorithm for Modified Poisson Regression for Prospective Studies with Binary Data

Authors: Lu Li[a], Jiayi Tong[a], Jenna Reps[b,c], Suchitra Rao[d], Mackenzie Edmondson[a], Vitaly Lorman[f], Hanieh Razzaghi[e], Haitao Chu[f], Christopher B. Forrest[e], Yong Chen[a]

a. Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

b. Janssen Research and Development, Titusville, NJ, USA

c. Observational Health Data Sciences and Informatics (OHDSI), New York, NY

d. Department of Pediatrics, University of Colorado School of Medicine and Children's Hospital Colorado, Aurora, CO

e. Applied Clinical Research Center, Children's Hospital of Philadelphia, Philadelphia, PA

f. Division of Biostatistics, University of Minnesota, Minneapolis, MN

Background

When analyzing a relatively rare binary outcome, the sparse data problem is a significant challenge[1]. The lack of sufficient cases (e.g., patients with disease) in the data leads to a biased estimation of the effect of a treatment in observational studies[2,3]. One strategy to tackle this problem is to collaborate via consortia such as OHDSI, where multiple contributors can contribute to their (aggregated) data to increase the number of cases. Often patient-level data cannot be shared due to privacy concerns.

The logistic regression model is a natural choice for modeling binary data and provides an odds ratio (OR) for the effect of an intervention or strength of an association by comparing the exposed and unexposed groups. The relative risk (RR) is another metric reporting the effect magnitude of the two groups. The choice between RR and OR is a long-standing debate[4-8] and RR is preferred over OR for most prospective studies due to collapsibility, especially when the outcome is not rare[9]. Poisson regression is usually recommended to estimate the adjusted relative risk directly[10] for binary data as it can be used to approximate the binomial distribution when the sample size is large, and probability is small. Zou proposed a modified Poisson regression with a sandwich error term, which allows the direct estimation of the adjusted relative risk with robust variance estimation even when the Poisson model is misspecified for the binary outcome[11].

Distributed/federated learning algorithms are needed, but, under OHDSI, needs to be devised to be communication efficient as iterative communications are impractical. We consider a modified Poisson regression for binary outcomes: a key advantage is that we obtain estimates of relative risks which has collapsibility.

We proposed a one-shot distributed learning algorithm for the modified Poisson regression for binary data, which has similar communication costs as a traditional meta-analysis and only require sharing of aggregated data. In particular, we are interested in the investigation of rare binary data with the Poisson regression model[9]. Without requiring individual-level data, the proposed distributed algorithm transfers aggregated data across sites once to obtain the estimates of relative risk using the Poisson regression. Along with the consistent estimates of the intervention effects, the sandwich estimation offers a robust variance estimation of the estimated relative risk.

Methods

We propose a one-shot distributed modified Poisson regression approach for binary data and refer to it as ODAP-B. The workflow of the proposed algorithm is presented in **Figure 1**. Assume that there are K sites in total. Using methods developed by Jordan et. al[12] and adapted to the clinical setting by Duan et al.[13,14], we implemented a surrogate likelihood approach, by constructing the surrogate log-likelihood function whereby only local site patient-level data and gradients from other sites are needed. This procedure *preserves the patient-level privacy* in the data integration process by only transferring aggregated data across sites *once*. The main steps are:

1. At each site, fit a Poisson regression and obtain the initial estimate $\widehat{\boldsymbol{\theta}}_k$ and variance $\hat{\sigma}_k^2$.
2. (Optional step) obtain meta-analysis estimator $\overline{\boldsymbol{\theta}}$, using initial estimates across k sites.
3. At each collaborative site, calculate the first two gradients of Poisson likelihood at $\overline{\boldsymbol{\theta}}$.
4. For each collaborative site, share the gradients to lead site; at lead site, construct surrogate likelihood function, and obtain the ODAP-B estimator.

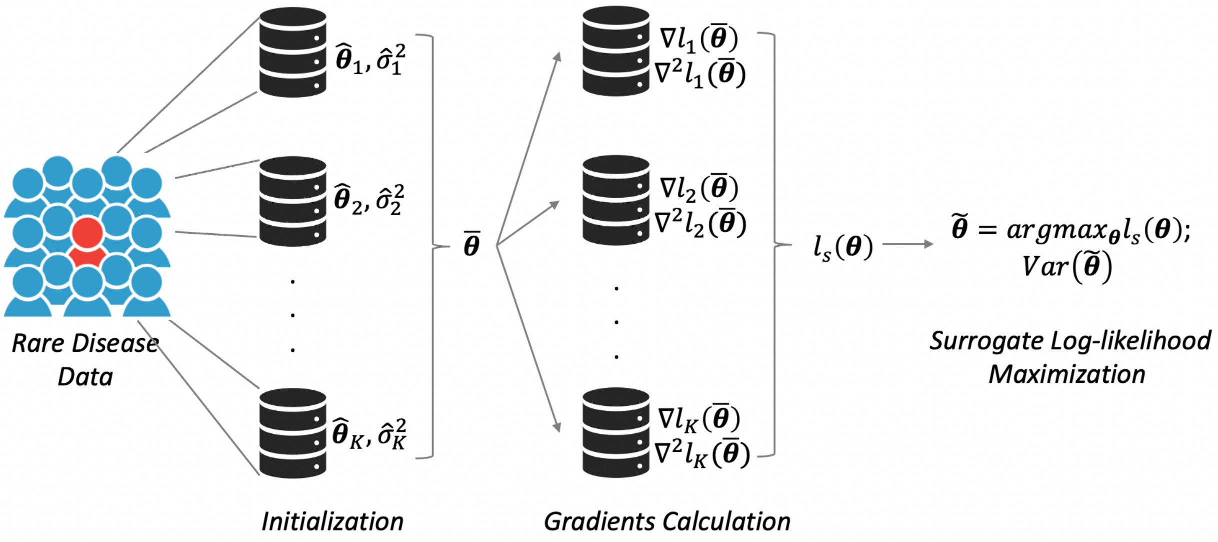
$\widehat{\boldsymbol{\theta}}_1, \hat{\sigma}_1^2$
$\widehat{\boldsymbol{\theta}}_2, \hat{\sigma}_2^2$
$\widehat{\boldsymbol{\theta}}_K, \hat{\sigma}_K^2$
$\overline{\boldsymbol{\theta}}$
$\nabla l_1(\overline{\boldsymbol{\theta}})$
$\nabla^2 l_1(\overline{\boldsymbol{\theta}})$
$\nabla l_2(\overline{\boldsymbol{\theta}})$
$\nabla^2 l_2(\overline{\boldsymbol{\theta}})$
$\nabla l_K(\overline{\boldsymbol{\theta}})$
$\nabla^2 l_K(\overline{\boldsymbol{\theta}})$
$l_s(\boldsymbol{\theta})$
$\widetilde{\boldsymbol{\theta}} = argmax_{\boldsymbol{\theta}} l_s(\boldsymbol{\theta});$
$Var(\widetilde{\boldsymbol{\theta}})$
Rare Disease Data
Initialization
Gradients Calculation
Surrogate Log-likelihood Maximization

Figure 1. Schematic illustration of the proposed ODAP-B method. For the rare disease data, each site calculates the initial estimate $\widehat{\boldsymbol{\theta}}_k$ and variance $\hat{\sigma}_k^2$ in the initialization step. Then the meta-estimate $\overline{\boldsymbol{\theta}}$ is obtained and transferred to all sites for the gradients calculation. Each site calculates the first and second gradients with the initial value and local data, and then transfers the gradients back to lead site or local site for the construction of surrogate likelihood function.

To evaluate the proposed method, we conducted a simulation study where we compared the performance of the proposed method to that of meta-analysis in terms of the relative bias to pooled estimates. We set the total number of sites to be 5 or 50, each with a sample size of 500. We consider the setting where a binary outcome is associated with four variables, including three binary predictors (e.g., medication, sex, chronic condition) and one continuous predictor (e.g., age) sampled from the "age" distribution in real world data. The exposure of interest (e.g., medication) was generated from a Bernoulli distribution with probability 0.3. We fit the following model:

$$log(P(Y = 1|\boldsymbol{X}, \boldsymbol{Z})) = \alpha + \beta_1 X_1 + \gamma_1 Z_1 + \gamma_2 Z_2 + \gamma_3 Z_3,$$

where the true values of the parameters are set as $\alpha = -2\ or\ -5$ to mimic a relatively common disease and rare disease, respectively. $\beta_1 = -0.25\ or\ -1$, and $\gamma_1 = \gamma_2 = \gamma_3 = -0.1$. The simulation was conducted with 1000 replications.

We also applied the method to examine the effect of COVID-19 viral (SARS-CoV-2 polymerase chain reaction (PCR) or antigen) test-positivity on the symptoms and conditions associated with the post-acute sequelae of SARS-CoV-2 (PASC) in children by using data from PEDSnet[15], a national clinical research network of large pediatric medical centers. We conducted the analysis to estimate the relative risk and 95% confidence interval (CI) of the risk factor, viral-positivity (yes versus no) for each of the five outcomes, including three syndromic and two systemic features. Our key finding is that when the outcome is relatively rare (prevalence = 6.7%), ODAP-B has much smaller bias and substantially more efficiency than the meta-analysis estimates.

Results

Boxplots summarizing results from our simulation studies are presented in Figure 2.

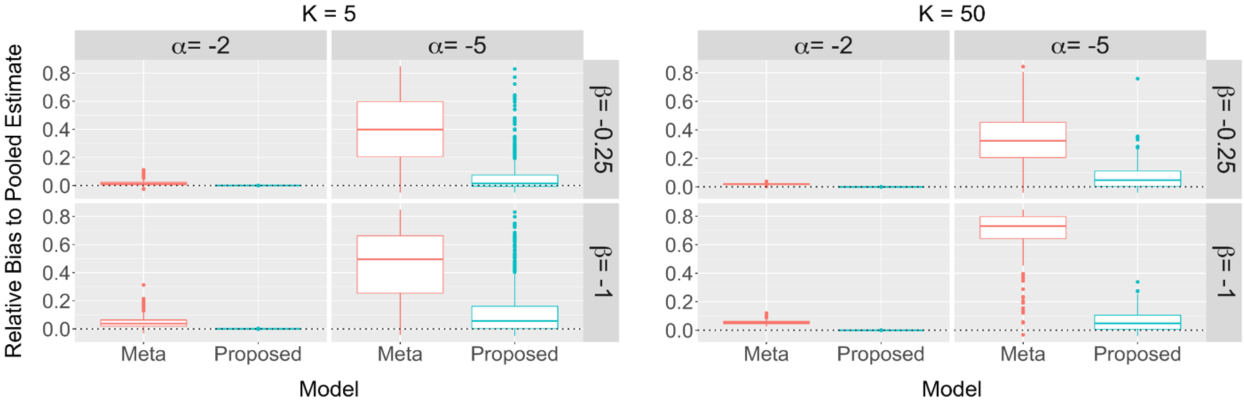

Figure 2. Comparison between the meta-analysis method (left box, red) and the proposed ODAP-B method (right box, cyan) in terms of relative bias of $\beta_1$when the total number of sites is 5 (left panel) and 50 (right panel).

Boxplots summarizing results from our real-world data application are presented in Figure 3.

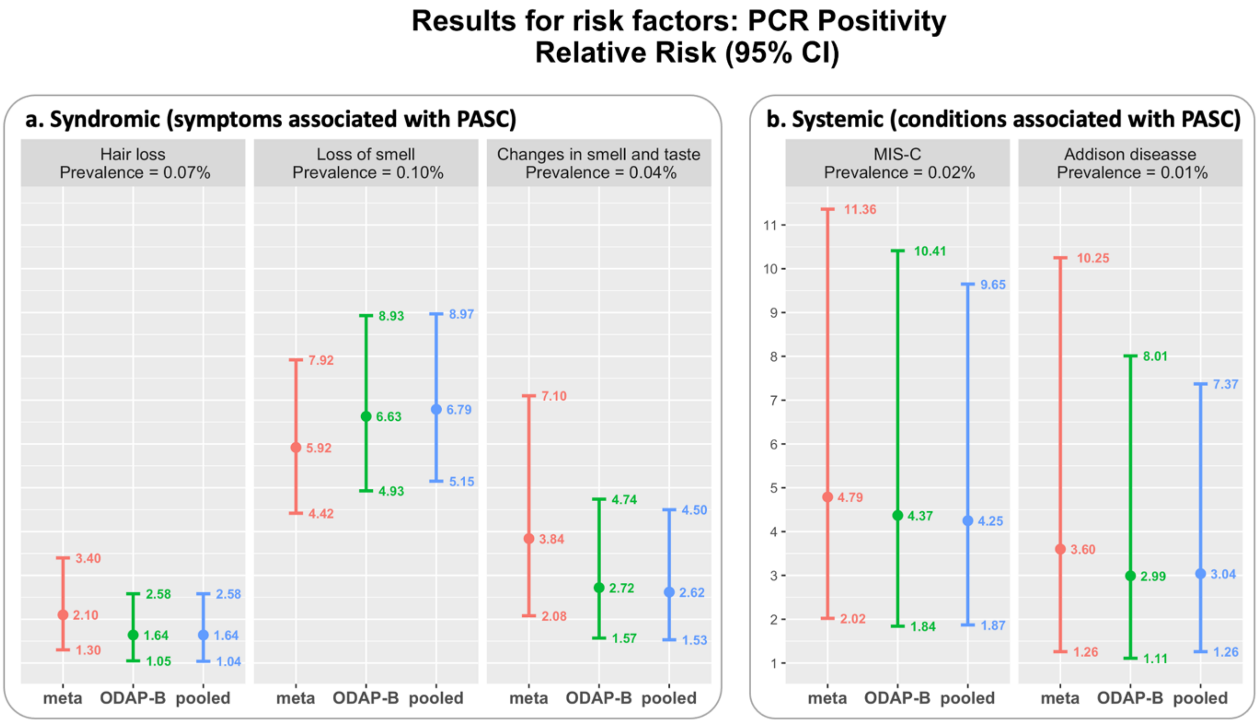

Figure 3. Comparison between the estimates of the risk factor, PCR positivity, with pooled method (blue), proposed ODAP-B (green), and meta-analysis method (red) using the real-world data on post-acute sequelae of SARS-CoV-2 infection (PASC) in 184,501 children across eight national clinical sites.

A real-world application with distributed data from three sites: Children's Hospital of Philadelphia, Johnson & Johnson (OHDSI), and University of Florida (ongoing) using ODAP-B is still ongoing.

Conclusion

We proposed the ODAP-B method for the analysis of rare binary outcomes. The proposed approach provides estimated relative risk (RR) with efficient sandwich variance estimates to analyze sparse binary data. We believe that ODAP-B is a significant contribution to this new generation of distributed research networks. As a powerful tool in modeling the risk factors of binary outcomes, the ODAP-B method facilitates the collaborative environment by providing accurate estimation, privacy-preserving feature, and efficient communication.

References/Citations

1. Greenland S. Noncollapsibility, confounding, and sparse-data bias. Part 2: What should researchers make of persistent controversies about the odds ratio? Journal of Clinical Epidemiology. 2021;139:264–8.
2. Richardson DB, Cole SR, Ross RK, Poole C, Chu H, Keil AP. Meta-analysis and sparse-data bias. Am J Epidemiol. 2021;190(2):336–40.
3. Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. bmj. 2016;352.

4. Suhail A D, Furuya-Kanamori L, Xu C, Lin L, Chivese T, Thalib L. Questionable utility of the relative risk in clinical research: A call for change to practice. 2020;
5. Xiao M, Chu H, Cole S, Chen Y, MacLehose R, Richardson D, et al. Odds Ratios are far from" portable": A call to use realistic models for effect variation in meta-analysis. arXiv preprint arXiv:210602673. 2021;
6. Furuya-Kanamori L, Xu C, Chivese T, Lin L, Musa OAH, Hindy G, et al. The odds ratio is “portable” but not the relative risk: Time to do away with the log link in binomial regression. 2021;
7. Xiao M, Chen Y, Cole S, MacLehose R, Richardson D, Chu H. Is OR “portable” in meta-analysis? Time to consider bivariate generalized linear mixed model. J Clin Epidemiol. 2022;142:280.
8. Greenland S, Pearl J, Robins JM. Confounding and collapsibility in causal inference. Statistical science. 1999;14(1):29–46.
9. Chu H, Cole SR. Estimation of risk ratios in cohort studies with common outcomes:" a bayesian approach". Epidemiology. 2010;855–62.
10. Zocchetti C, Consonni D, Bertazzi PA. Estimation of prevalence rate ratios from cross-sectional data. Int J Epidemiol. 1995;24(5):1064–5.
11. Zou G. A modified poisson regression approach to prospective studies with binary data. Am J Epidemiol. 2004;159(7):702–6.
12. Jordan MI, Lee JD, Yang Y. Communication-efficient distributed statistical inference. J Am Stat Assoc. 2018;
13. Duan R, Boland MR, Liu Z, Liu Y, Chang HH, Xu H, et al. Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. J Am Med Inform Assoc. 2019;
14. Duan R, Luo C, Schuemie MJ, Tong J, Liang JC, Chang HH, et al. Learning from local to global - an efficient distributed algorithm for modeling time-to-event data. biorxiv.orgPaperpile.
15. Forrest CB, Margolis PA, Bailey LC, Marsolo K, Del Beccaro MA, Finkelstein JA, et al. PEDSnet: a National Pediatric Learning Health System. J Am Med Inform Assoc. 2014;21(4):602–6.