Distributed Learning from Multi-Site Observational Health Data for Zero-Inflated Count Outcomes

Mackenzie J. Edmondson[a], Chongliang Luo[a], Rui Duan[a], Mitchell Maltenfort[b], Zhaoyi Chen[c,d], Justine Shults[a], Jiang Bian[c,d], Patrick B. Ryan[e], Christopher B. Forrest[b], Yong Chen[a]

a. Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
b. Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA, USA
c. Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, USA
d. Cancer Informatics Shared Resource, University of Florida Health Cancer Center, Gainesville, FL, USA
e. Janssen Research and Development, Titusville, NJ, USA

Keywords: Distributed algorithm, Electronic health records, Hurdle regression, Data privacy, Zero-inflation

Corresponding Author:

Yong Chen
Ychen123@upenn.edu
Associate Professor
Department of Biostatistics, Epidemiology and Informatics
Perelman School of Medicine, University of Pennsylvania

Abstract

Background: Multi-site studies facilitate the study of rare outcomes or exposures through integrating patient information from several distinct care sites. Due to patient privacy concerns, sharing of patient-level information among collaborating sites is often prohibited, suggesting a need for privacy-preserving data analysis methods. Several such methods exist, but have been shown to sometimes result in biased estimation or require extensive communication among sites.

Objective: We present a communication-efficient, privacy-preserving method for performing distributed regression on Electronic Health Records (EHR) data across multiple sites for zero-inflated count outcomes. Our approach is motivated by two real-world data problems: modeling frequency of serious adverse events and examining risk factors associated with pediatric avoidable hospitalization.

Methods: We use hurdle regression, a two-part (logistic-Poisson) regression model, to characterize the effects of risk factors on zero-inflated count outcomes. Further, we develop a one-shot algorithm for performing hurdle regression (ODAH) across multiple sites, using individual patient data at one site and aggregated data from all other sites to approximate the complete data log likelihood. We evaluate ODAH through extensive simulations and an application to EHR data from the Children's Hospital of Philadelphia (CHOP) and the OneFlorida Clinical Research Consortium. We compare ODAH estimates to those from meta-analysis and pooled analysis (the gold standard in which all patient data are pooled together).

Results: In simulations, ODAH estimates exhibited bias relative to the gold standard of less than 0.1% across several settings. In contrast, meta-analysis estimated exhibited relative bias up to

12.7%, largely dependent on the event rate. When applying ODAH to CHOP data, relative biases for estimates in both components of the hurdle model were less than 5.1%, while meta-analysis estimates exhibited relative bias as high as 63.6%. When analyzing OneFlorida data, ODAH relative biases were less than 10% for eight of the ten log relative risks estimated, while meta-analysis estimates again showed substantially greater bias.

Conclusions: Our simulations and real-world applications suggest ODAH is a promising method for performing privacy-preserving distributed learning on EHR data when modeling zero-inflated count outcomes.

Word Count: 4,794

Introduction

Background

Electronic health records (EHRs) comprise patient data including demographics, diagnoses, procedures, medication, and laboratory tests, among additional patient-level information. EHRs are a viable alternative to data traditionally collected in clinical trials, which can be both expensive and time-consuming [1]. Count data, consisting of non-negative integers denoting the observed frequency of a discrete event, are common in EHRs; length of stay, number of hospitalizations, and number of physician visits are all popular count outcomes in studies using observational health data [2-3]. Count outcomes are typically modeled with either Poisson or Negative Binomial regression.

Medical count data are frequently susceptible to zero-inflation, where the number of zero counts observed far exceeds that expected in either a true Poisson or Negative Binomial distribution. Examples often involving zero-inflation include medications prescribed according to insurance claims during a given period and number of laboratory tests ordered during an

emergency department visit [3]. In scenarios with zero-inflated count outcomes, it is sometimes postulated that the processes generating zero and positive counts are systematically different, defined by distinct distributions. For these cases, data can be represented with a hurdle model, which consist of two components: one for estimating the probability of a non-zero count (typically a logistic regression), and another for estimating a count given that it is greater than zero (typically a zero-truncated count model, commonly Poisson or Negative Binomial regression) [2]. The sequential nature of the hurdle model allows for each component's set of parameters to be estimated separately, benefitting computational efficiency.

A chief concern regarding EHR use is patient privacy. EHRs contain sensitive patient-level identifiers, often preventing data sharing across institutions. This leads to many clinical data analyses performed at single sites; these analyses are often underpowered, with results specific to a certain sub-population due to choice of site.

In light of the limitations of single-site analyses, many have stressed the importance of multi-site studies [4-6]. Multi-site studies allow for integration of clinical information from several sites by using distributed data networks (DDNs). DDNs, made up of several health care institutions, are designed to assist analyses of medical product safety and comparative effectiveness research, among many other uses [6]. By integrating patient information from several sources, DDNs facilitate the study of rare outcomes or exposures, often featuring larger and more inclusive samples from the target population. An example of a successful DDN is the Observational Health Data Sciences and Informatics (OHDSI) consortium, whose primary purpose is to develop open-source tools to be shared across multiple centers for use in collaborative observational health data research [7]. Another successful DDN is PEDSnet (pedsnet.org), a multi-site network made up of eight large pediatric health systems [8]. PEDSnet

contains patient information for over 6 million children, offering substantial opportunity to enhance the quality of pediatric EHR research.

Ideally, one would analyze data in a multi-site study by pooling all patient-level data together at a central site prior to analysis. This is not always possible due to concerns regarding patient privacy and confidentiality; the Health Insurance Portability and Accountability Act of 1996 (HIPAA) established a privacy rule to regulate use of protected health information (PHI) often found in EHRs, requiring de-identification of PHI prior to secondary use in biomedical research [9]. De-identified PHI have proven to be susceptible to re-identification, causing concern among patients [10-11].

Existing Methods for Privacy-Preserving Data Analysis

There are several established methods for performing privacy-preserving data analysis aside from using de-identified data alone. A common approach for analyzing data in multi-site studies is to use meta-analysis, where only aggregate established measures are shared across sites. Meta-analysis is a long-standing popular choice for privacy-preserving analysis, notably in several OHDSI studies [12-14]. While relatively easy to implement, meta-analysis has been shown to result in biased or imprecise effect estimates in the context of rare outcomes or exposures, as well as with smaller sample sizes [15]. Another favorable option is to use distributed regression, primarily designed to break up computationally-intensive tasks into smaller, more manageable tasks. Each of the smaller tasks is computed in parallel at separate centers without sharing raw data. One example of a distributed algorithm is WebDISCO (Web service for DIStributed COx model learning), which fits the Cox proportional hazards model in a distributed fashion across several sites using only aggregated statistics [16]. Another example is GLORE (Grid binary LOgistic REgression), a distributed algorithm for performing logistic

regression. Like WebDISCO, GLORE avoids use of patient-level data by only using aggregated information from model fittings at individual sites [17].

A main limitation of many distributed regression methods is the extensive communication required between sites until convergence is reached while estimating model parameters [16-17]. Methods such as WebDISCO and GLORE may require several iterations of an algorithm across sites, which can be time-consuming and computationally expensive. Duan et al. recently proposed ODAL (One-shot Distributed Algorithm for performing Logistic regression) and ODAC (One-shot Distributed Algorithm to fit a multicenter Cox proportional hazards model), communication-efficient alternatives to GLORE and WebDISCO, respectively [15,18]. Like the aforementioned methods, ODAL and ODAC preserve privacy by avoiding data sharing at the observation level, but offer an advantage in terms of efficiency, requiring only one or two rounds of communication between centers.

Goal of This Study

We build upon the framework of communication-efficient distributed algorithms and present an algorithm for hurdle regression, ODAH (One-shot Distributed Algorithm for performing Hurdle regression), for modeling zero-inflated count data. We evaluate ODAH through an extensive simulation study before applying our method to two real-world data use cases: analyzing risk factors of pediatric avoidable hospitalization using data from the Children's Hospital of Philadelphia and modeling serious adverse event frequency for colorectal cancer patients using data from the OneFlorida Clinical Research Consortium.

Methods

Poisson-Logit Hurdle Model

A hurdle model is an altered count model in which the processes of generating zero and positive counts are not constrained to be the same, designed to cope with zero-inflated count outcomes [2]. To derive the hurdle model, we consider the two processes making up the model independently. First, we model the proportion of zero counts with a Bernoulli process using a logit link. Let $w_1, w, \dots, w_n \in \{0,1\}$ be independent realizations of a binary response variable W, such that $P(w_i = 1) = \pi_i$ and $P(w_i = 0) = 1 - \pi_i$. The logistic model of the probability π_i is modeled as a linear combination of explanatory variables $\boldsymbol{X}$ and regression coefficients $\boldsymbol{\beta}$:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \boldsymbol{X}_i^T \boldsymbol{\beta}. \tag{1}$$

Next, positive counts are modeled using a zero-truncated Poisson model. Let $y_1, y_2, \dots, y_n \in \{0,1,2, \dots\}$ be independent realizations of a count variable Y. Assume $P(Y_i = 0) = P(w_i = 0) = 1 - \pi_i$, and $P(Y_i > 0) = P(w_i = 1) = \pi_i$. Thus, π_i can interpreted as the probability that the "hurdle is crossed", resulting in a non-zero count. In the context of zero-inflated counts, we assume $P(y_i = 0)$ is much greater than $P(y_i > 0)$.

For observations where the realization from the logistic model is 1, positive counts follow a zero-truncated Poisson distribution such that $P(Y_i = y_i | Y_i > 0) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{(1-e^{-\lambda_i})y_i!}$. Thus, we can write the mixture probability mass function of the Poisson hurdle model as

$$P(Y_i = y_i) = \begin{cases} 1 - \pi_i, & y_i = 0 \\ \pi_i \dfrac{e^{-\lambda_i}\lambda_i^{y_i}}{(1 - e^{-\lambda_i})y_i!}, & y_i = 1, 2, 3, \dots \end{cases}.$$

Modeling the rate parameter λ_i using a log link, we can express the log of λ_i as a linear combination of explanatory variables $\boldsymbol{Z}$ and regression coefficients $\boldsymbol{\gamma}$:

$$\log(\lambda_i) = \boldsymbol{Z}_i^T \boldsymbol{\gamma}.$$

(Figure 1 here.)

We write the log-likelihood of the Poisson hurdle model as $L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = L_1(\boldsymbol{\beta}) + L_2(\boldsymbol{\gamma})$, with

$$L_1(\boldsymbol{\beta}) = \sum_{Y_i > 0} X_i^T \boldsymbol{\beta} - \sum_{i=1}^{n} \log\left(1 + e^{X_i^T \boldsymbol{\beta}}\right) \tag{2}$$

and

$$L_2(\boldsymbol{\gamma}) = \sum_{Y_i > 0} \left(-e^{Z_i^T \boldsymbol{\gamma}} + Y_i Z_i^T \boldsymbol{\gamma} - \log\left(1 - e^{-e^{Z_i^T \boldsymbol{\gamma}}}\right) - \log(Y_i!)\right). \tag{3}$$

Note that this factors into two components such that $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are separable; the Hessian matrix is block diagonal, so $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are information orthogonal. Thus, there will not be any loss of information in estimating each set of parameters separately. This property is useful in the context of distributed regression, reducing computational complexity.

While less common than traditional regression models for count data, hurdle models have been used successfully in various health contexts with substantial zero inflation. For instance, Negative Binomial – Logit hurdle models were utilized to estimate risk of vaccine adverse events for clinical trial participants, as well as to estimate cigarette and marijuana use among youth e-cigarette users [19, 20]. Hurdle regression has also been used in other specialized contexts, such as in estimating spatiotemporal patterns of emergency department use and quantifying association between preventive dental behaviors and caries prevalence [21, 22]. Contrary to zero-inflated Poisson or Negative Binomial regression models, hurdle models have only one source of zero counts, indicating that all individuals in the study sample are at risk of a particular outcome.

This is more appropriate in many clinical settings, offering improved interpretation of estimated model coefficients.

Parameter Estimation using Distributed Hurdle Regression: ODAH

Suppose we have clinical data stored in K sites, where the jth site has a sample size n_j and the total sample size across sites is $N = \sum_{j=1}^{K} n_j$. Let Y_{ij} and X_{ij} denote the count outcome and covariate vector for subject i in site j, respectively. We can write the log likelihood functions for the combined data as

$$L_{1N}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{n_j} X_{ij}^T\boldsymbol{\beta} - \log\left(1 + e^{X_{ij}^T\boldsymbol{\beta}}\right)$$

and

$$L_{2N}(\boldsymbol{\gamma}) = \frac{1}{N}\sum_{j=1}^{K}\sum_{Y_{ij}>0}\left(-e^{Z_{ij}^T\boldsymbol{\gamma}} + Y_{ij}Z_{ij}^T\boldsymbol{\gamma} - \log\left(1 - e^{-e^{Z_{ij}^T\boldsymbol{\gamma}}}\right) - \log(Y_{ij}!)\right)$$

In the context of distributed inference, we assume that we do not have access to the combined data. We only have access to data at one of the K sites (the *lead site*, with site index j = 1), as well as aggregate information from the other sites (the *collaborating sites*). Using methods developed by Jordan et al. and later adapted to the clinical data setting by Duan et al., we construct a *surrogate log likelihood function*, which approximates the complete data log likelihood using patient-level data from the lead site and aggregate information from the collaborating sites [15,18,23]. The aggregate information used in our work is the set of first- and second-order gradients of the log likelihood function at the K – 1 collaborating sites. The surrogate log likelihood function for each component of the hurdle model can be expressed as

$$\widetilde{L_1}(\boldsymbol{\beta}) = L_{11}(\boldsymbol{\beta}) + \{\nabla L_{1N}(\overline{\boldsymbol{\beta}}) - \nabla L_{11}(\overline{\boldsymbol{\beta}})\}\boldsymbol{\beta} + \frac{1}{2}(\boldsymbol{\beta} - \overline{\boldsymbol{\beta}})^2\{\nabla^2 L_{1N}(\overline{\boldsymbol{\beta}}) - \nabla^2 L_{11}(\overline{\boldsymbol{\beta}})\} \quad (4)$$

and

$$\widetilde{L_2}(\boldsymbol{\gamma}) = L_{21}(\boldsymbol{\gamma}) + \{\nabla L_{2N}(\overline{\boldsymbol{\gamma}}) - \nabla L_{21}(\overline{\boldsymbol{\gamma}})\}\boldsymbol{\gamma} + \frac{1}{2}(\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma}})^2\{\nabla^2 L_{2N}(\overline{\boldsymbol{\gamma}}) - \nabla^2 L_{21}(\overline{\boldsymbol{\gamma}})\}, \quad (5)$$

where $\overline{\boldsymbol{\beta}}$ and $\overline{\boldsymbol{\gamma}}$ are initial estimates for the algorithm. Here, $L_{11}(\boldsymbol{\beta})$ and $L_{21}(\boldsymbol{\gamma})$ are log-likelihoods computed using patient-level data at the lead site for the logistic and zero-truncated components, respectively. The terms

$$\nabla^g L_{1N}(\overline{\boldsymbol{\beta}}) = \frac{1}{N}\sum_{j=1}^{K} n_j \nabla^g L_{1j}(\overline{\boldsymbol{\beta}})$$

and

$$\nabla^g L_{2N}(\overline{\boldsymbol{\gamma}}) = \frac{1}{N}\sum_{j=1}^{K} n_j \nabla^g L_{2j}(\overline{\boldsymbol{\gamma}})$$

are weighted averages of first-order (g = 1) or second-order (g = 2) gradients at each site, and $\nabla^g L_{11}(\overline{\boldsymbol{\beta}})$ and $\nabla^g L_{21}(\overline{\boldsymbol{\gamma}})$ are first-order or second-order gradients calculated at the lead site for the logistic and Poisson components of the hurdle model, respectively, evaluated at $\overline{\boldsymbol{\beta}}$ and $\overline{\boldsymbol{\gamma}}$. Explicit formulations of the first- and second-order gradients for each component of the hurdle model are available in the Supplement (Equations S.1 – S.4). The ODAH estimators are then defined as

$$\widetilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, \tilde{L}_1(\boldsymbol{\beta})$$

and

$$\widetilde{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmax}}\, \tilde{L}_2(\boldsymbol{\gamma}).$$

Well-chosen $\overline{\boldsymbol{\beta}}$ and $\overline{\boldsymbol{\gamma}}$ will increase the accuracy of $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\gamma}}$, respectively. In this work, $\overline{\boldsymbol{\beta}}$ and $\overline{\boldsymbol{\gamma}}$ are estimates computed from performing a fixed-effects meta-analysis using all *K* sites, or inverse-variance weighted sums of estimates from the *K* studies, i.e. for $\overline{\boldsymbol{\beta}}$ (with $\overline{\boldsymbol{\gamma}}$ similar),

$$\overline{\boldsymbol{\beta}} = \frac{\sum_{j=1}^{K} \widehat{\boldsymbol{\beta}_j}\, \boldsymbol{\omega}_j}{\sum_{j=1}^{K} \boldsymbol{\omega}_j}, \qquad \text{with } \boldsymbol{\omega}_j = \frac{1}{\widehat{\boldsymbol{\sigma}_j^2}}.$$

This requires each site to send point and variance estimates to the lead site to initiate the algorithm. Alternatively, one could use lead site maximum likelihood estimates $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\gamma}}_1$ obtained via fitting the hurdle model of interest at the lead site. This has been shown to work well when the lead site is largely representative of the entire sample across sites and requires one less round of communication among sites [18].

When using a meta-analysis estimate to initiate ODAH, two non-iterative rounds of communication are necessary for transferring information across sites; thus, our approach is considered a *one-shot* approach for performing distributed inference. ODAH requires each collaborating site to first fit the hurdle model of interest using its own data before sending parameter point and variance estimates to the lead site. A user at the local site can then initiate ODAH by, following its own hurdle model fitting, computing initial estimates using meta-analysis before sending these estimates to the collaborating sites for computing gradients. These gradients are then sent to the lead site to construct the surrogate log likelihood function. Using only gradients and patient-level data from the lead site, we obtain parameter estimates calculated from maximizing each surrogate likelihood function with respect to the parameter of interest. The ODAH algorithm is outlined in detail below.

Algorithm:

Input: Patient-level data $X = \{X_{i1}\}, Y = \{Y_{i1}\}$ from the lead site, as well as parameter estimates $\widehat{\boldsymbol{\beta}_j}$, $\widehat{\boldsymbol{\gamma}_j}$, $\widehat{\boldsymbol{\sigma}_j^2}$, and $\widehat{\boldsymbol{\tau}_j^2}$, first order-gradients ($\frac{1}{n_j}\nabla L_{1j}(\overline{\boldsymbol{\beta}})$ and $\frac{1}{n_j}\nabla L_{2j}(\overline{\boldsymbol{\gamma}})$) and second-order gradients ($\frac{1}{n_j}\nabla^2 L_{1j}(\overline{\boldsymbol{\beta}})$ and $\frac{1}{n_j}\nabla^2 L_{2j}(\overline{\boldsymbol{\gamma}})$) from coordinating sites, where i, j denote the observation and clinical site indices, respectively.
Output: Surrogate maximum likelihood estimators $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\gamma}}$.

Initialization:
1: At site $j = 1, \ldots, K$, **do**
Fit hurdle model and obtain point estimates $\widehat{\boldsymbol{\beta}_j}$ and $\widehat{\boldsymbol{\gamma}_j}$, as well as variance estimates $\widehat{\boldsymbol{\sigma}_j^2}$ and $\widehat{\boldsymbol{\tau}_j^2}$ of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$, respectively. Send $\widehat{\boldsymbol{\beta}_j}$, $\widehat{\boldsymbol{\gamma}_j}$, $\widehat{\boldsymbol{\sigma}_j^2}$, and $\widehat{\boldsymbol{\tau}_j^2}$ to the lead site.
end

2: At lead site, compute initial estimates $\overline{\boldsymbol{\beta}}$ and $\overline{\boldsymbol{\gamma}}$ using meta-analysis. Send to sites $j = 2, \ldots, K$.
3: At site sites $j = 2, \ldots, K$, **do**
Calculate first order-gradients ($\frac{1}{n_j}\nabla L_{1j}(\overline{\boldsymbol{\beta}})$ (S.1) and $\frac{1}{n_j}\nabla L_{2j}(\overline{\boldsymbol{\gamma}})$ (S.3)) and second-order gradients ($\frac{1}{n_j}\nabla^2 L_{1j}(\overline{\boldsymbol{\beta}})$ (S.2) and $\frac{1}{n_j}\nabla^2 L_{2j}(\overline{\boldsymbol{\gamma}})$ (S.4)). Send to lead site.
end

Surrogate Likelihood Construction/Maximization:
1: At lead site, compute surrogate log likelihoods $\widetilde{L_1}(\boldsymbol{\beta})$ (4) and $\widetilde{L_2}(\boldsymbol{\gamma})$ (5).
2: At lead site, obtain $\widetilde{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \tilde{L}_1(\boldsymbol{\beta})$ and $\widetilde{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} \tilde{L}_2(\boldsymbol{\gamma})$.
3: **Return** $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\gamma}}$.

(Figure 2 here)

Simulation Study

To evaluate ODAH empirically in a controlled setting, we conducted a simulation study to primarily compare the performance of ODAH to that of meta-analysis, which does not incorporate any subject-level data. We additionally compare the performance ODAH to that when using only lead site data and all subject-level data (*pooled* – the gold standard).

In our simulations, a count outcome Y was associated with two risk factors, X_1 and X_2. X_1 was generated using a truncated Normal distribution emulating the number of primary care visits per year for each patient in the Children's Hospital of Philadelphia (CHOP) data ($X_1 \sim N(3,2), X_1 \in (0,18)$), while X_2 was generated using a Bernoulli distribution with the probability of success representing that of public insurance use among CHOP patients in our sample ($X_2 \sim Bern(0.33)$). Our covariate of interest was X_2, with X_1 assumed to be a confounder. Y was generated from the Poisson-Logit hurdle described above. The sets of covariates making up each component of the model are the same; $\boldsymbol{\beta} = \{\beta_0, \beta_1, \beta_2\}$ and $\boldsymbol{\gamma} = \{\gamma_0, \gamma_1, \gamma_2\}$ are each 3 x 1 vectors of regression coefficients.

Motivated by our rare-event real-world data applications, we primarily sought to examine how varying levels of low outcome prevalence and event rate affect the performance of ODAH relative to pooled analysis. We explored four rare-event prevalence settings while holding event rate constant at 0.03 (event rate in CHOP data): 5%, 2.5% (near CHOP data prevalence), 1%, and 0.5%. To evaluate the effect of event rate on method performance, we explored additional event rates of 0.25, 0.01, and 0.005 while holding outcome prevalence constant at 2.5%.

In all settings, we fixed the number of sites $K = 10$ and total population size $N = 200{,}000$. In settings outcome prevalence or event rate vary, we set $n_1 = n_2 = \cdots = n_{10}$, so all sites had the same number of observations. We also explored the effect of the lead site being larger than the

collaborating sites, setting lead site sizes at 38,000 (collaborating site size 18,000), 56,000 (collaborating site size 16,000), and 74,000 (collaborating site size 14,000). All ten unique simulation settings explored are summarized in Table 1.

For each setting, we evaluated estimation accuracy using lead site data, meta-analysis, ODAH, and pooled data across 1,000 simulations in terms of relative bias to the pooled estimate. In all settings, we assume true coefficient values $\{\beta_1, \gamma_1\} = -1$ and $\{\beta_2, \gamma_2\} = 1$.

(Table 1 here)

2.4 Real-World Data Applications

To examine the performance of ODAH on real world data, we applied our algorithm to two real-world use cases featuring zero-inflated count outcomes.

Children's Hospital of Philadelphia: Pediatric Avoidable Hospitalization

The CHOP system provides care to about 400,000 children per year and includes a large, multi-state outpatient network, as well as one of the largest inpatient facilities for pediatric patients residing in the greater Philadelphia region. Data for this study were extracted from the CHOP EHR system for outpatient, emergency department, and inpatient visits for patients with at least two primary care facility visits from January 2009 to December 2017.

To mimic a scenario in which different sites do not have access to patient-level information at other sites, we assigned patients to the primary care site they attended most often during the study period and carried out analysis as if patient-level information could not be shared across primary care sites. In total, patients were assigned to 27 different primary care sites; we selected six of these sites to illustrate our method, made up of 70,818 patients (Table 2). The largest site of these six, Site 4, was chosen to be the lead site.

For this study, we sought to examine risk factors associated with pediatric avoidable hospitalization (AH); about one-third of pediatric healthcare costs are associated with hospital admissions, the majority of which are unplanned [24]. Unplanned hospitalizations associated with a diagnosis treatable at the primary care level are considered avoidable [25]. By studying which risk factors are most strongly associated with AH, hospital systems can identify patient subpopulations for which primary care should be improved, ideally leading to an overall reduction in hospital costs or admissions [26]. Because pediatric hospitalization is uncommon, integrating data across hospital systems can lead to more robust inference, increasing power to detect differences in rates of AH among patients.

To evaluate ODAH, we modeled total number of AHs given a collection of EHR variables: gender, race (Caucasian or other), mean age (across all visits), primary care visits per year, and insurance type (public or private). While the majority of patients who experience an AH in these data only experience one, 22% experience more than one, suggesting an advantage of using Poisson regression over logistic regression alone (Figure 3). This, combined with substantial zero-inflation, makes Poisson-Logit hurdle regression an appropriate method for modeling these data. The probability of at least one AH is estimated using a logistic regression model. For patients with at least one AH, the total number of hospitalizations is estimated using a zero-truncated Poisson regression model.

(Table 2 here)

(Figure 3 here)

OneFlorida Clinical Research Consortium: Serious Adverse Event Frequency

The OneFlorida Clincal Research Consortium contains robust longitudinal and linked patient-level real-world data of around 15 million Floridians, making up over 50% of the Florida

population. OneFlorida data includes records from Medicaid & Medicare claims, cancer registry data, vital statistics, and EHRs from its clinical partners. OneFlorida data are centralized in a HIPAA limited dataset (i.e., dates are not shifted and 9-digit zip codes of patients' residencies are available) that contains detailed patient and clinical variables, including demographics, encounters, diagnoses, procedures, vitals, medications, and labs, following the PCORnet Common Data Model (CDM). The OneFlorida data undergo rigorous quality checks at its data coordinating center, the University of Florida, and a privacy-preserving record linkage process is used to deduplicate records of same patients coming from different health care systems within the network [27]. Figure 4 shows the geographic locations of OneFlorida partners.

(Figure 4 here)

To apply ODAH to OneFlorida data, we identified a study population of patients with colorectal cancer (CRC) and who use FOLFIRI, an FDA-approved standard of care first line chemotherapy treatment in patients with metastatic CRC, as their CRC treatment. We focused on assessing drug safety in terms of the occurrences of serious adverse events (SAEs). To define an SAE, we followed the FDA definition of SAEs and the Common Terminology Criteria for Adverse Events (CTCAE) v 5.0, and the number of SAEs were counted for each patient within 180 days after first FOLFIRI prescription [28]. We removed the chronic conditions that occurred before prescription. A set of covariates and risk factors for all patients were extracted from patients' medical records for this analysis, including patients' demographic variables (age, race, Hispanic ethnicity status, and gender) on the day of CRC diagnosis. We also calculated each patient's Charlson comorbidity index (CCI) using their medical history.

Since OneFlorida data are centralized, we were able to both carry out analysis as if patient-level information could not be shared across clinical sites (as was done in our CHOP application)

as well as fit a hurdle regression model using pooled analysis, which served as the gold standard. In total, our analysis included 660 patients from three clinical sites, with Site 3 being the largest and serving as the local site (Table 3). To evaluate ODAH using these data, we modeled serious adverse event frequency given the extracted clinical information noted above for each patient.

(Table 3 here)

As in our simulations, we evaluated method performance by calculating relative bias to the pooled estimate for lead site analysis, meta-analysis, and ODAH. To estimate variance of ODAH parameter estimates, we used the inverse of the Hessian matrix produced when optimizing the surrogate log likelihood function of each hurdle model component.

Results

Simulation Results

Figure 5 depicts simulation results from evaluating method performance across all scenarios described in Table 2. Across settings, there was no discernable difference in method performance for estimating β_2, the regression coefficient associated with X_2 in the logistic component of the hurdle model. We therefore present the simulation results for estimating γ_2, leaving β_2 estimation results for the Supplement. Due to select iterations resulting in outlying estimates when using lead site analysis, the median bias for the lead site estimate across iterations is reported rather than the mean.

(Figure 5 here)

When lead site size and event rate were fixed at $20{,}000$ and $\lambda = 0.03$, respectively, we varied outcome prevalence to see how each method performed relative to pooled analysis, the gold standard (Figure 4A). In all prevalence levels examined, ODAH performed nearly as well as

pooled analysis, with negligible difference in the estimate's bias and variance; bias in the ODAH estimate relative to the pooled estimate was less than 0.1% for each prevalence level. Conversely, meta-analysis bias relative to the pooled estimate increased with decreasing prevalence, ranging from 0.97% (5% prevalence) to 10.4% (0.5% prevalence). Lead site analysis exhibited the largest variance of all methods; bias relative to the pooled estimate ranged from 0.79% (5% prevalence) to 2.77% (0.5% prevalence).

When lead site size and outcome prevalence were fixed at 20,000 and 2.5%, respectively, we varied event rate to examine its impact on estimating γ_2 in a low prevalence setting (Figure 4B). For all methods, variance of estimates decreased with increasing event rate. ODAH and meta-analysis estimates were nearly identical to pooled estimates when events rates were set to $\lambda = 0.25$ and 0.03, exhibiting negligible bias relative to the pooled estimate (ODAH bias < 0.1%, meta-analysis bias < 1.9%). When the event rate was set to $\lambda = 0.01$ and 0.005, ODAH again exhibited negligible relative bias (< 0.1%) but meta-analysis exhibited larger bias relative to the pooled estimate (4.57% and 12.7%, respectively). Lead site analysis exhibited the largest variance of all methods examined, maintaining relatively low relative bias to the pooled estimate when $\lambda = 0.25$, 0.03 and 0.01 (< 1.1%) but larger bias when λ = 0.005 (5.31%).

When examining the effect of increasing lead site size while fixing outcome prevalence and event rate at 2.5% and $\lambda = 0.03$, respectively, there was not substantial evidence for lead site size affecting ODAH or meta-analysis performance relative to pooled analysis (Figure 4C). Variance of lead site analysis estimates decreased with increasing lead site size.

3.2. Data Analysis Results

Avoidable Hospitalization (CHOP) Results

Figure 6 depicts results of our analysis of CHOP EHRs to estimate AH. Regression coefficient estimates for each covariate in the fitted hurdle model are shown along with their corresponding 95% confidence interval.

(Figure 6 here)

Log odds ratio estimates (corresponding to the logistic component of the hurdle model) when using ODAH were relatively close to the pooled estimates, with relative bias to pooled estimates ranging from 0.08% (insurance covariate) to 5.02% (primary care visits per year covariate). Meta-analysis estimates were more biased, with relative bias ranging from 4.15% (gender covariate) to 63.6% (primary care visits per year covariate).

Log relative risk estimates (corresponding to the Poisson component of the hurdle model) were nearly identical when using ODAH and pooled analysis. Meta-analysis performed similarly to ODAH across all coefficients, but ODAH always achieved the smaller relative bias to pooled estimates. ODAH relative bias was < 0.50% for all covariates, while meta-analysis relative bias ranged from 5.89% (PC visits per year) to 11.7% (race).

SAE (OneFlorida) Results

Results from using ODAH to model SAE frequency in colorectal cancer patients using data from OneFlorida are shown in Figure 7, displayed similarly to the CHOP AH results. In this application, we again see our method performing well in terms of relative bias to pooled estimation. For four of the five log odds ratios estimated in the logistic component of the hurdle model, relative biases produced by ODAH were less than 7%. The lone exception, the gender coefficient, reflected greater bias due to its near-zero effect size (reflecting an odds ratio of 1). Similar results were observed in the zero-truncated Poisson component, with relative biases to the pooled estimates less than 10% for four of the five estimated log relative risks. The age

coefficient had higher relative bias, again due to negligible effect size. In both components, meta-analysis tended to do poorer relative to pooled estimation. The largest difference in estimation can be seen in the coefficients reflecting association of SAE frequency with Hispanic ethnicity, where relative bias was 71% in the logistic component and 276% in the Poisson component (compared to 5.3% and 1.8% for ODAH, respectively).

(Figure 7 here)

Discussion

We introduced a non-iterative, privacy-preserving algorithm for performing distributed hurdle regression with zero-inflated count outcomes. As demonstrated by simulations and a real-world EHR application, our method consistently produced parameter estimates comparable to and sometimes better than those produced by meta-analysis. Our method's utility is especially evident in settings featuring a count outcome with severe zero-inflation and very low event rate, as we demonstrated the tendency of only meta-analysis to produce biased estimates under these circumstances.

There are many advantages to using our method for performing privacy-preserving data analysis. By using a form of distributed regression, our approach is generally well-suited for multi-site studies which are on-going. The surrogate likelihood method takes advantage of patient-level data still being accessible by collaborating sites, allowing collaborators to engage in limited inter-site communication to produce less biased results than would be obtained via meta-analysis, which is better suited for studies that are already completed. Further, most existing distributed regression techniques require iterative communication among sites to produce accurate estimates. ODAH requires two rounds of non-iterative communication between the local site and all other sites before surrogate likelihood functions can be maximized to obtain

accurate, precise parameter estimates. This is particularly advantageous in big data settings, where iterative procedures have a high computational burden in terms of memory and processing time. Additionally, due to the separability of hurdle model components, each component's likelihood function can be maximized independently, reducing computational complexity.

Our simulation results suggest that lead site size relative to total population size does not have a discernable effect on any method performance outside of analysis only using data at the lead site. However, since the surrogate log likelihood function only uses individual-level data stored in the lead site, we recommend that the lead site is as large as possible; this ensures the surrogate likelihood is a close approximation to the complete data log likelihood.

In terms of limitations when using ODAH, the main limitation is the requirement of relative homogeneity among the data to be analyzed. This is an implication of the surrogate likelihood construction, which approximates the complete data log likelihood in part by using a sample-size-weighted sum of gradients. This implicitly assumes that study data are independent and identically distributed across all sites, which is not the case in real-world settings. As evidenced by Figure 8, geographical heterogeneity among the patient population can occur in the covariates, with some locations having substantially different demographic makeups than others. We recommend those who implement ODAH ensure patient demographics are largely similar across institutions, or alternatively perform subgroup analysis for several relatively homogeneous subsets of institutions. Additionally, the Poisson component of our method does not currently account for over-dispersion in the outcome; in our real data application, we did not find strong evidence of dispersion. Finally, there were discrepancies when comparing simulation and data analysis results in terms of bias in the hurdle model's logistic component estimates. We suspect this is due to simulated data not fully capturing the true distribution of the CHOP data,

namely covariate imbalance. For example, 52% of patients that had at least one AH used public insurance, compared to 32% of patients who did not have an AH. We seek to address these limitations in future work.

(Figure 8 here)

Work continues to be done in our group on constructing methods for performing non-iterative, privacy-preserving distributed inference. We seek to eventually cover a wide array of outcome types, namely binary, count, and time-to-event outcomes. As we continue to develop a collection of privacy-preserving algorithms, we believe ODAH is worthy of consideration when one seeks to perform distributed regression on zero-inflated count outcome data.

Funding

This work is support in part by National Institutes of Health Grants 1R01LM012607 (ME, CL, RD and YC) and 1R01AI130460 (ME, CL, RD and YC).

This study is part of the Pediatric Big Health Data initiative funded by the State of Pennsylvania and led by the Children's Hospital of Philadelphia, University of Pennsylvania, and the Urban Health Collaborative at Drexel University. We would like to thank the investigators of the Pediatric Big Health Data initiative for their contributions. These individuals include: Christopher B. Forrest, MD, PhD; L. Charles Bailey, MD, PhD; Shweta P. Chavan, MSEE; Rahul A. Darwar, MPH; Jillian Benedetti, MPH; Daniel Forsyth; Chén C. Kenyon, MD, MSHP; Ritu Khare, PhD; Mitchell G. Maltenfort, PhD; Aaron J. Masino, PhD, ME; Xueqin Pang, PhD; Ting Qian, PhD; Hanieh Razzaghi, MPH; Justine Shults, PhD; Levon H. Utidjian, MD, MBI from the Children's Hospital of Philadelphia; Ana V. Diez Roux, MD, PhD, MPH; Amy H. Auchincloss, PhD, MPH; Elizabeth A. Campbell, MSPH; Kimberly Daniels, MS; Anneclaire J. De Roos, PhD, MPH; J. Felipe Garcia-Espana, MS, PhD; Irene Headen, PhD, MS; Félice Lê-Scherban, PhD, MPH; Steven Melly, MS, MA; Yvonne L. Michael, ScD, SM; Jeffrey Moore, MS; Kari Moore, MS; Abigail E. Mudd, MPH; Leah Schinasi, PhD, MSPH; and Yuzhe Zhao, MS from Drexel University and, Yong Chen, PhD; John H. Holmes, PhD; Rebecca A. Hubbard, PhD; A. Russell Localio, JD, MPH, PhD from the University of Pennsylvania.

This work was supported by a grant from the Commonwealth Universal Research Enhancement (C.U.R.E) program funded by the Pennsylvania Department of Health - 2015 Formula award - SAP #4100072543.

Conflicts of Interest

None declared.

Abbreviations

EHR: electronic health record
ODAH: One-shot Distributed Algorithm for performing Hurdle regression
CHOP: Children's Hospital of Philadelphia
DDN: distributed data network
OHDSI: Observational Health Data Sciences and Informatics
PHI: protected health information
AH: avoidable hospitalization
OR: odds ratio
RR: relative risk

REFERENCES

1. Center for Drug Evaluation and Research. Submitting Documents Using Real-World Data and Real-World Evidence. *FDA Med Bull*. Available: www.fda.gov/regulatory-information/search-fda-guidance-documents/submitting-documents-using-real-world-data-and-real-world-evidence-fda-drugs-and-biologics-guidance

2. Cameron, A. Colin, & Trivedi, P. K. (1998). *Regression analysis of count data.* Cambridge, UK: Cambridge University Press.

3. Du J, Park YT, Theera-Ampornpunt N, McCullough JS, Speedie SM. The use of count data models in biomedical informatics evaluation research. *J Am Med Inform Assoc*. 2012;19(1):39-44.

4. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. *Sci Transl Med.* 2010;2: 57cm29.

5. Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: A practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. Medical Care 48:S45-S51, 2010.

6. Fleurence, R.L., et al., Launching PCORnet, a national patient-centered clinical research network. Journal of the American Medical Informatics Association, 2014. 21(4): p. 578- 582.

7. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc.* 2012;19: 54–60.

8. 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: 602–606.

9. April Moreno Arellano, Wenrui Dai, Shuang Wang, Xiaoqian Jiang, and Lucila Ohno-Machado. Privacy Policy and Technology in Biomedical Data Science. Annual Review of Biomedical Data Science 2018 1:1, 115-129.

10. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. *J Am Med Inform Assoc*. 2010;17(2):169-177. doi:10.1136/jamia.2009.000026.

11. D. McGraw, Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data. J. Am. Med. Inform. Assoc (2012).10.1136/amiajnl- 2012-000936.

12. Vashisht R, Jung K, Schuler A, Banda JM, Park RW, Jin S, Li L, Dudley JT, Johnson KW, Shervey MM, Xu H. Association of hemoglobin A1c levels with use of sulfonylureas, dipeptidyl peptidase 4 inhibitors, and thiazolidinediones in patients with type 2 diabetes treated with metformin: analysis from the observational health data sciences and informatics initiative. JAMA network open. 2018 Aug 3;1(4):e181755-.

13. Duke JD, Ryan PB, Suchard MA, Hripcsak G, Jin P, Reich C, Schwalm MS, Khoma Y, Wu Y, Xu H, Shah NH. Risk of angioedema associated with levetiracetam compared with phenytoin: Findings of the observational health data sciences and informatics research network. Epilepsia. 2017 Aug;58(8):e101-6.

14. Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, Suchard MA, Schuemie MJ, DeFalco FJ, Perotte A, Banda JM. Characterizing treatment pathways at scale using the OHDSI network. Proceedings of the National Academy of Sciences. 2016 Jul 5;113(27):7329-36.

15. Duan, R., Luo, C., Schuemie, M. J., Tong, J., Liang, C. J., Chang, H. H., Boland, M. R., Bian, J., Xu, H., Holmes, J. H., Forrest, C. B., Morton, S. C., Berlin, J. A., Moore, J. H., Mahoney, K. B., & Chen, Y. (2020). Learning from local to global: An efficient distributed algorithm for modeling time-to-event data. *Journal of the American Medical Informatics Association : JAMIA*, *27*(7), 1028–1036. https://doi.org/10.1093/jamia/ocaa044

16. Lu, C. L., Wang, S., Ji, Z., Wu, Y., Xiong, L., Jiang, X., & Ohno-Machado, L. (2015). WebDISCO: a web service for distributed cox model learning without patient-level data sharing. *Journal of the American Medical Informatics Association*, *22*(6), 1212-1219.

17. Wu Y, Jiang X, Kim J, Ohno-Machado L. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. *J Am Med Inform Assoc*. 2012;19: 758–764.

18. Duan, R., Boland, M.R., Liu, Z., Liu, Y., Chang, H.H., Xu, H., Chu, H., Schmid, C.H., Forrest, C.B., Holmes, J.H. and Schuemie, M.J., 2020. Learning from electronic health

records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *Journal of the American Medical Informatics Association*, *27*(3), pp.376-385.

19. Rose CE, Martin SW, Wannemuehler KA, Plikaytis BD. On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. J Biopharm Stat. 2006;16(4):463-481.

20. Pittman B, Buta E, Krishnan-Sarin S, O'Malley SS, Liss T, Gueorguieva R. Models for analyzing zero-inflated and overdispersed count data: an application to cigarette and marijuana use [published online ahead of print, 2018 Apr 18]. Nicotine Tob Res. 2018;10.1093/ntr/nty072.

21. Neelon B, Chang HH, Ling Q, Hastings NS. Spatiotemporal hurdle models for zero-inflated count data: Exploring trends in emergency department visits. Stat Methods Med Res. 2016;25(6):2558-2576.

22. Hofstetter H, Dusseldorp E, Zeileis A, Schuller AA. Modeling Caries Experience: Advantages of the Use of the Hurdle Model. Caries Res. 2016;50(6):517-526.

23. Jordan MI, Lee JD, Yang Y. Communication-efficient distributed statistical inference. Journal of the American Statistical Association. 2019 Apr 3;114(526):668-81.

24. Bui AL, Dieleman JL, Hamavid H, Birger M, Chapin A, Duber HC, Horst C, Reynolds A, Squires E, Chung PJ, Murray CJ. Spending on children's personal health care in the United States, 1996-2013. JAMA pediatrics. 2017 Feb 1;171(2):181-9.

25. Lu, S., & Kuo, D. Z. (2012). Hospital charges of potentially preventable pediatric hospitalizations. *Academic pediatrics*, *12*(5), 436-444.

26. Maltenfort MG, Chen Y, Forrest CB. Prediction of 30-day pediatric unplanned hospitalizations using the Johns Hopkins Adjusted Clinical Groups risk adjustment system. PloS one. 2019;14(8).

27. Bian J, Loiacono A, Sura A, et al. Implementing a hash-based privacy-preserving record linkage tool in the OneFlorida clinical research network. Jamia Open 2019; 2: 562–569.

28. CFR - Code of Federal Regulations Title 21 [Internet]. [cited 2020 Mar 6]. Available from: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=314.8

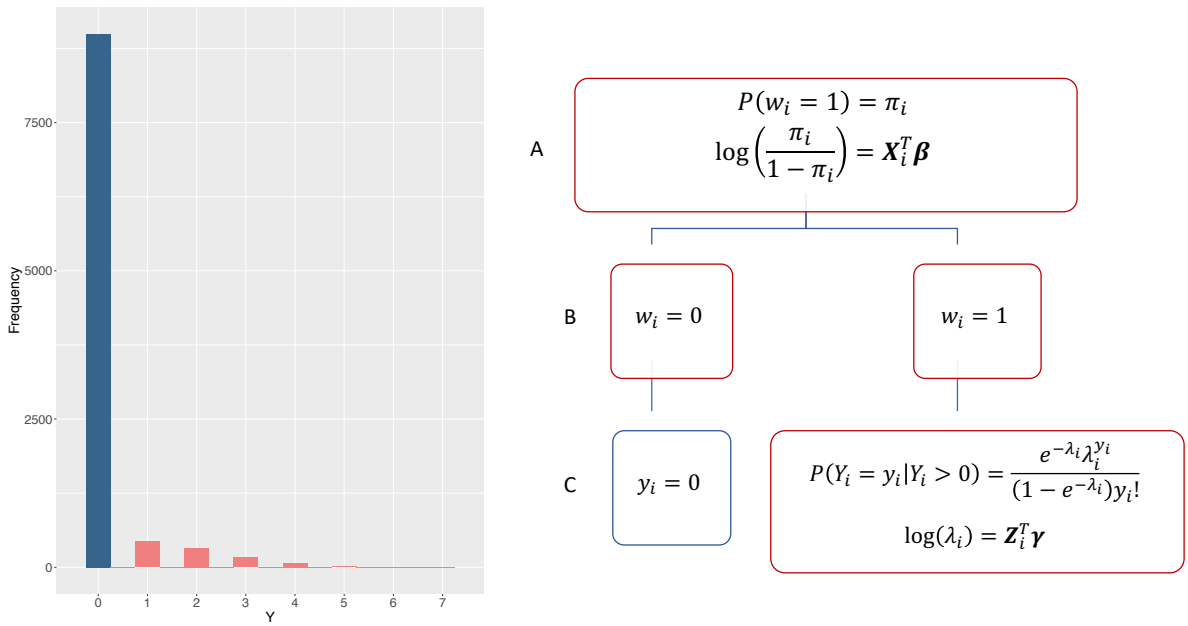

Figure 1. On the left, a histogram displaying counts generated with a Poisson-Logit hurdle distribution with 10% prevalence and an event rate of $\lambda = 1.5$. On the right, a hierarchical diagram visualizing the data generation process in a Poisson-Logit hurdle framework. Independent realizations $w_i \in \{0,1\}$ are generated from a Bernoulli process, with underlying probability π_i modeled using a logit link. Realizations where $w_i = 0$ are zero counts ($y_i = 0$), while realizations where $w_i = 1$ are positive counts ($y_i \in \{1,2,\ldots\}$). The positive counts are generated by a zero-truncated Poisson distribution, with underlying event rate λ_i modeled using a log link.

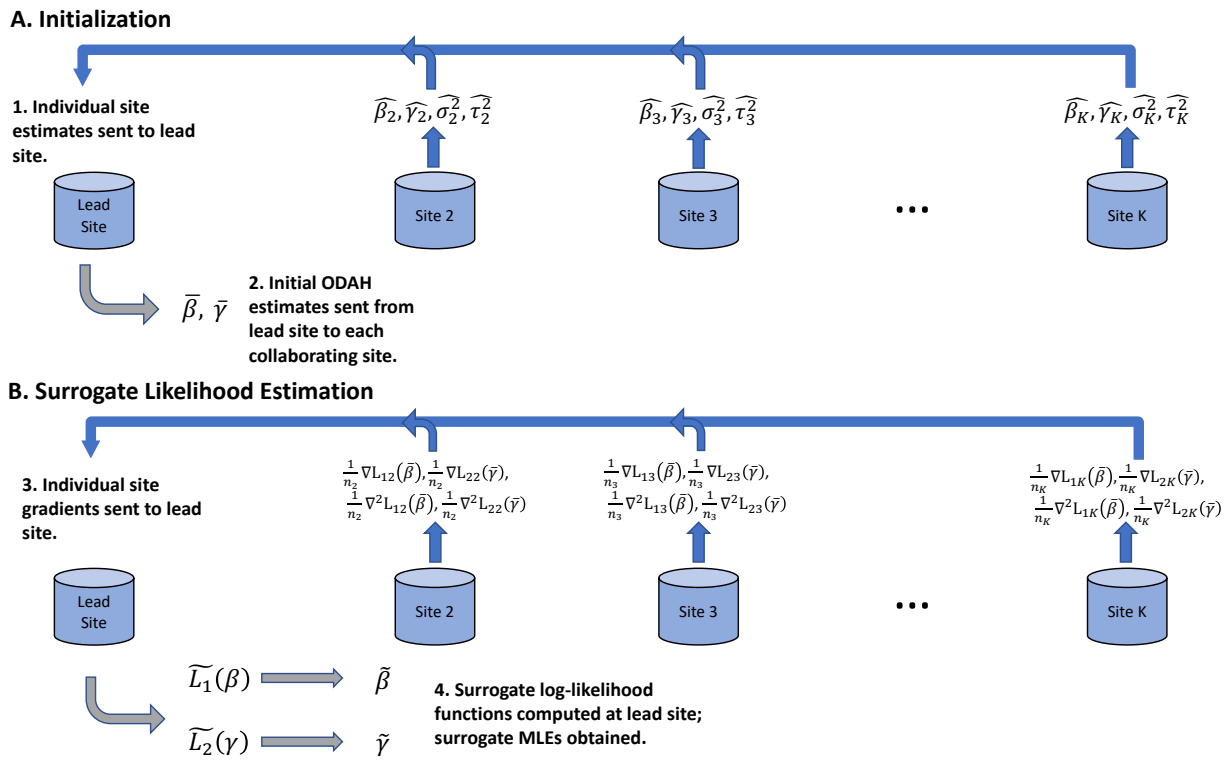

Figure 2. Visual representation of one-shot algorithm for performing hurdle regression (ODAH).

A.) Upper panel: Initialization. Effect size and variance estimates are obtained from fitting separate hurdle models at each site and are sent to the lead site (1); these estimates are then used in a meta-analysis to produce initial estimates for ODAH (2).

B.) Lower panel: Surrogate likelihood estimation. First- and second-order gradients are computed at each site, evaluated at the initial estimates obtained in step (2) and sent to the lead site (3). These gradients are used in conjunction with data from the lead site to construct surrogate likelihood functions $\tilde{L}_1(\boldsymbol{\beta})$ and $\tilde{L}_2(\boldsymbol{\gamma})$, which are then maximized to produce surrogate maximum likelihood estimates $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\gamma}}$ (4).

Simulation Setting			True Parameter Values		
Prevalence	**Event Rate (λ)**	$\boldsymbol{n_{lead}/N}$	$\boldsymbol{\beta_0}$	$\boldsymbol{\gamma_0}$	$\boldsymbol{n_{lead}}$
5%	0.03	0.10	-3.0	-3.6	20,000
2.5%	0.03	0.10	-3.7	-3.6	20,000
1%	0.03	0.10	-4.5	-3.6	20,000
0.5%	0.03	0.10	-5.3	-3.6	20,000
2.5%	0.25	0.10	-3.7	-1.4	20,000
2.5%	0.01	0.10	-3.7	-4.5	20,000
2.5%	0.005	0.10	-3.7	-5.3	20,000
2.5%	0.03	0.19	-3.7	-3.6	38,000
2.5%	0.03	0.28	-3.7	-3.6	56,000
2.5%	0.03	0.37	-3.7	-3.6	74,000

Table 1. Simulation settings varying baseline outcome prevalence β_0, baseline event rate γ_0, and size of lead site n_{lead}.

	Site 1 (n=5456)	Site 2 (n=9111)	Site 3 (n=7893)	Site 4 (n=27288)	Site 5 (n=7996)	Site 6 (n=13074)	Total (n=70818)
Gender							
Female	2589 (47.5%)	4427 (48.6%)	3862 (48.9%)	13458 (49.3%)	4013 (50.2%)	6494 (49.7%)	34843 (49.2%)
Male	2867 (52.5%)	4684 (51.4%)	4031 (51.1%)	13830 (50.7%)	3983 (49.8%)	6580 (50.3%)	35975 (50.8%)
Caucasian Race							
Caucasian	3476 (63.7%)	5508 (60.5%)	4783 (60.6%)	15747 (57.7%)	4649 (58.1%)	9158 (70.0%)	43321 (61.2%)
Other	1980 (36.3%)	3603 (39.5%)	3110 (39.4%)	11541 (42.3%)	3347 (41.9%)	3916 (30.0%)	27497 (38.8%)
Mean Age (across visits) (years)							
Mean (SD)	8.02 (5.48)	7.95 (5.58)	7.77 (5.50)	7.54 (5.60)	7.60 (5.57)	7.04 (5.37)	7.57 (5.54)
Median [Min, Max]	7.87 [0.0216, 18.0]	7.67 [0.0181, 18.0]	7.44 [0.0376, 17.9]	6.79 [0.0158, 17.9]	7.02 [0.0170, 17.9]	6.10 [0.0202, 17.9]	6.97 [0.0158, 18.0]
Insurance Provider							
Public	1997 (36.6%)	3410 (37.4%)	2339 (29.6%)	9545 (35.0%)	2477 (31.0%)	3438 (26.3%)	23206 (32.8%)
Private/Self-Pay	3459 (63.4%)	5701 (62.6%)	5554 (70.4%)	17743 (65.0%)	5519 (69.0%)	9636 (73.7%)	47612 (67.2%)
PC Visits per Year							
Mean (SD)	5.19 (5.14)	5.00 (4.75)	4.84 (4.35)	4.51 (4.59)	5.34 (4.86)	5.17 (4.85)	4.88 (4.72)
Median [Min, Max]	3.52 [0.243, 65.3]	3.68 [0.276, 85.3]	3.50 [0.276, 70.8]	3.16 [0.238, 97.5]	3.95 [0.233, 73.2]	3.83 [0.253, 85.3]	3.50 [0.233, 97.5]
Hospitalization Status							
At least one avoidable hospitalization (AH)	71 (1.3%)	70 (0.8%)	33 (0.4%)	878 (3.2%)	76 (1.0%)	396 (3.0%)	1524 (2.2%)
No AHs	5385 (98.7%)	9041 (99.2%)	7860 (99.6%)	26410 (96.8%)	7920 (99.0%)	12678 (97.0%)	69294 (97.8%)
Total AHs (for those with at least one AH)							
Mean (SD)	1.38 (1.19)	1.51 (1.82)	1.48 (1.64)	1.46 (1.16)	1.46 (0.901)	1.47 (1.58)	1.47 (1.31)
Median [Min, Max]	1.00 [1.00, 10.0]	1.00 [1.00, 15.0]	1.00 [1.00, 10.0]	1.00 [1.00, 10.0]	1.00 [1.00, 5.00]	1.00 [1.00, 16.0]	1.00 [1.00, 16.0]
Follow-up Time							
Mean (SD)	3.43 (2.05)	4.67 (2.76)	4.74 (2.75)	4.72 (2.72)	4.92 (2.77)	4.75 (2.74)	4.64 (2.72)
Median [Min, Max]	3.25 [0.0766, 8.74]	4.58 [0.0766, 8.74]	4.83 [0.0766, 8.74]	4.74 [0.0766, 8.74]	5.08 [0.0766, 8.74]	4.74 [0.0766, 8.74]	4.58 [0.0766, 8.74]

Table 2. Summary statistics describing patient population across six CHOP primary care sites.

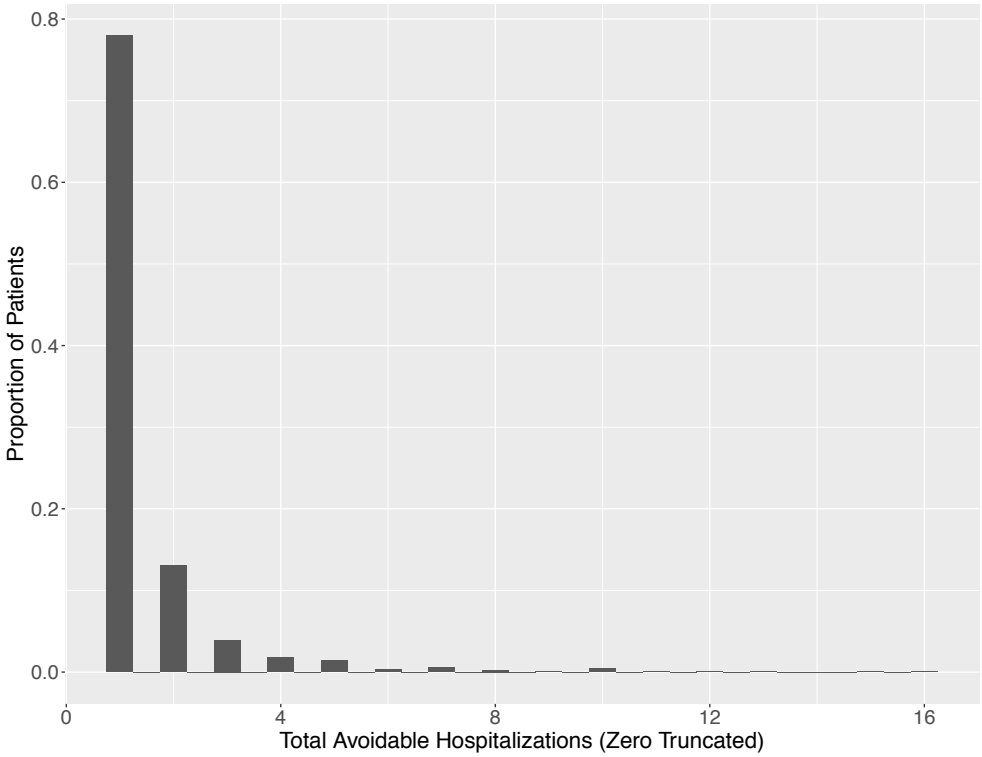

Figure 3. Distribution of total number of avoidable hospitalizations (AHs) for patients with at least one AH in CHOP data sample.

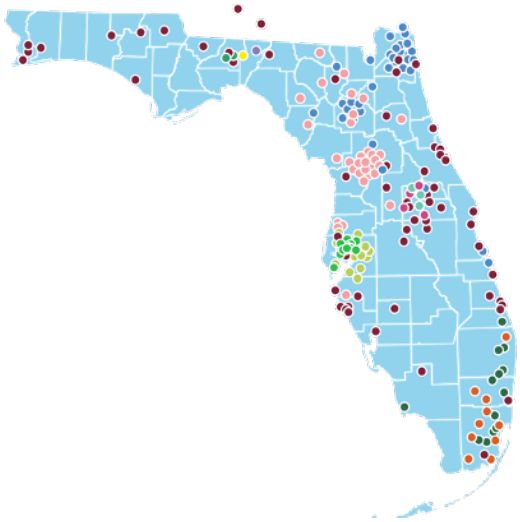

Figure 4. Map detailing locations of OneFlorida clinical partners.

	Site 1 (n=48)	Site 2 (n=226)	Site 3 (n=386)	Total (n=660)
Gender				
Female	22 (45.8%)	90 (39.8%)	178 (46.1%)	290 (43.9%)
Male	26 (54.2%)	136 (60.2%)	208 (53.9%)	370 (56.1%)
Caucasian Race				
Caucasian	25 (52.1%)	165 (73.0%)	302 (78.2%)	492 (74.5%)
Other	23 (47.9%)	61 (27.0%)	84 (21.8%)	168 (25.5%)
Age (years)				
Mean (SD)	51.8 (9.55)	56.2 (11.9)	57.2 (11.9)	56.5 (11.8)
Hispanic				
Yes	12 (25.0%)	9 (4.0%)	226 (58.5%)	247 (37.4%)
No	36 (75.0%)	217 (96.0%)	160 (41.5%)	413 (62.6%)
Charlson Comorbidity Index (CCI)				
Mean (SD)	5.23 (0.52)	5.27 (0.75)	5.24 (0.87)	5.25 (0.81)
Serious Adverse Events (SAEs)				
Mean (SD)	1.81 (1.71)	2.11 (2.19)	1.47 (1.72)	1.72 (1.91)

Table 3. Summary statistics describing patient population across three OneFlorida clinical sites.

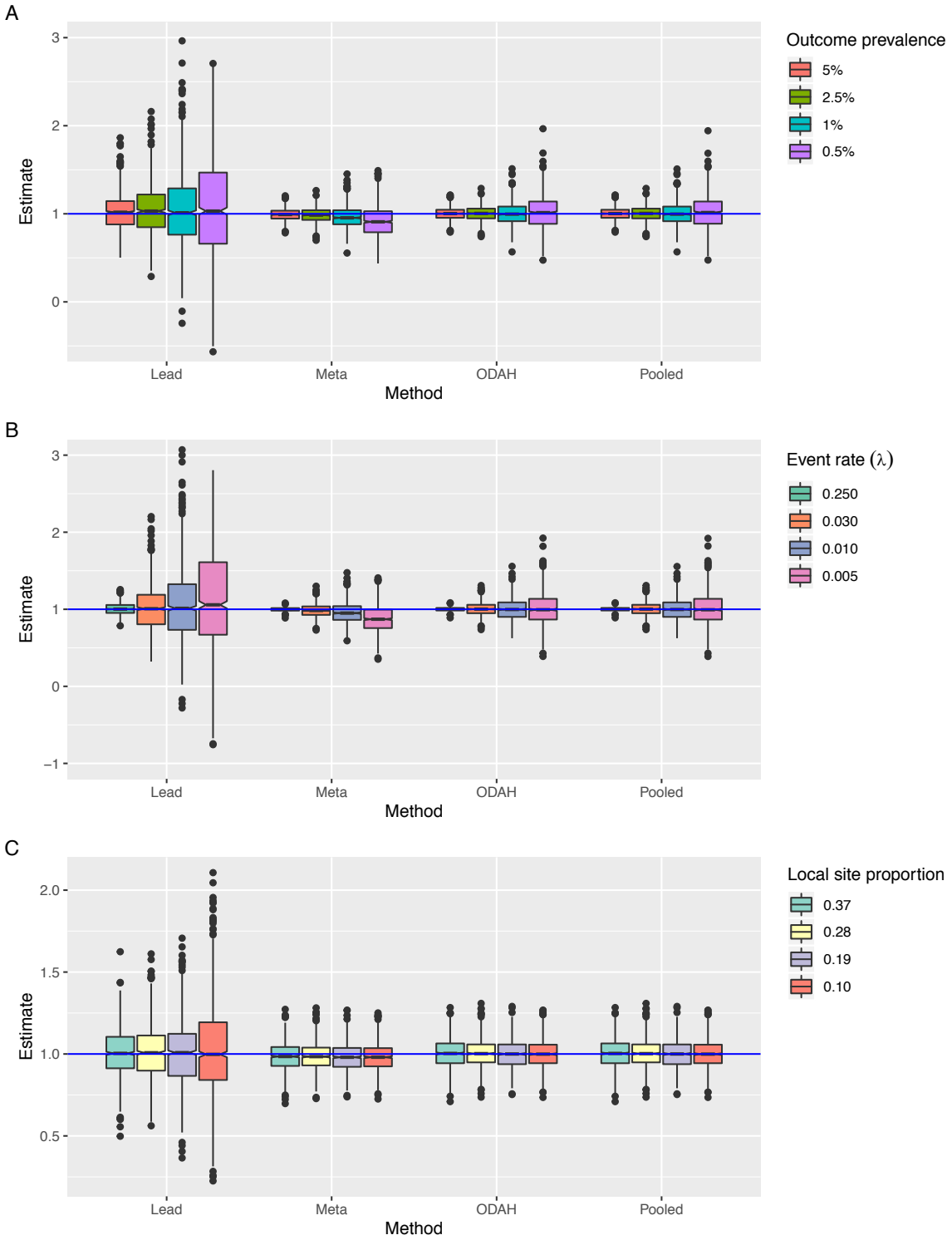

Figure 5. Simulation results for estimating Poisson component covariate γ_2. A) Results for Setting A, fixing n_{lead} = 20,000 and γ_0 = -3.6 ($\lambda = 0.03$) while varying outcome prevalence. B) Results for Setting B, fixing n_{lead} = 20,000 and β_0 = -3.7 (2.5% prevalence) while varying event rate (λ). C) Results for Setting C, fixing β_0 = -3.7 (2.5% prevalence) and $\gamma_0 = -3.6$ ($\lambda = 0.03$) while varying proportion of observations in lead site. Horizontal blue line represents true value of $\gamma_2 = 1$.

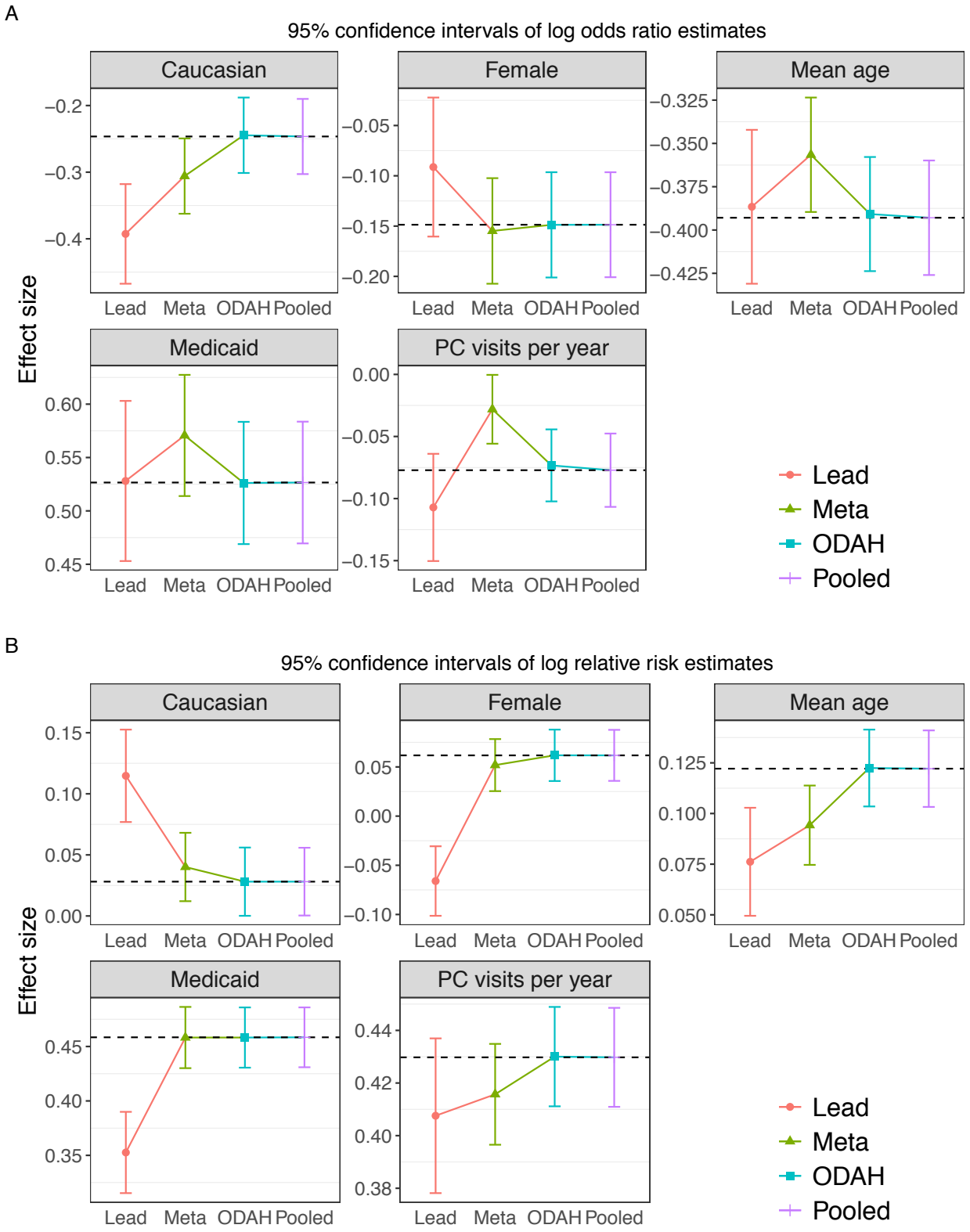

Figure 6. Plots depicting results from CHOP avoidable hospitalization application. Log odds ratio (A) and log relative risk (B) estimates (along with corresponding 95% confidence intervals) for each covariate in the fitted hurdle model. Dashed horizontal line represents pooled estimate, our gold standard for comparing methods.

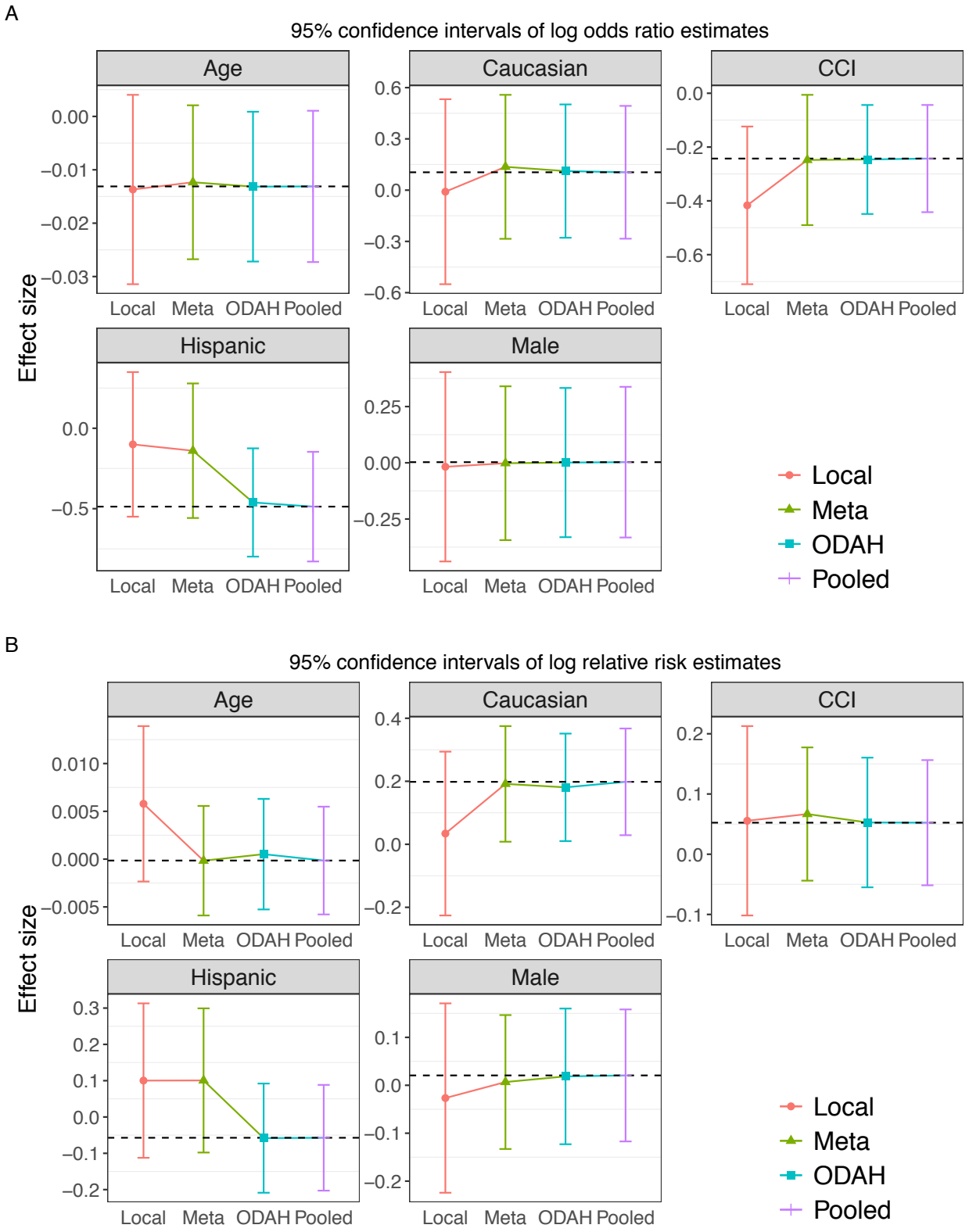

Figure 7. Plots depicting results from OneFlorida serious adverse event application. Log odds ratio (A) and log relative risk (B) estimates (along with corresponding 95% confidence intervals) for each covariate in the fitted hurdle model. Dashed horizontal line represents pooled estimate, our gold standard for comparing methods.

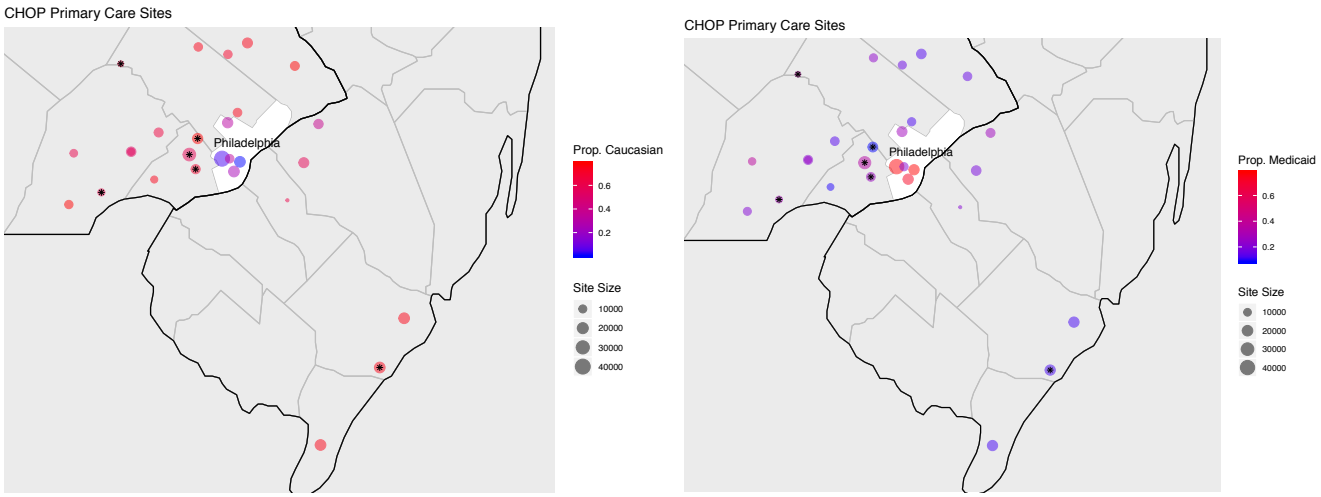

Figure 8. Geographical map of 27 CHOP primary care sites across greater Philadelphia region. In the left map, the proportion of patients of Caucasian race are depicted for each site. In the right map, the proportion of patients using public insurance (Medicaid) at each site is depicted. The size of each site on each map is proportional to the number of patients at the given site. Stars indicate sites used in our data analysis.