scientific reports

Check for updates

OPEN

An efficient and accurate distributed learning algorithm for modeling multi-site zero-inflated count outcomes

Mackenzie J. Edmondson[1], Chongliang Luo[1], Rui Duan[2], Mitchell Maltenfort[3], Zhaoyi Chen[4,5], Kenneth Locke Jr.[1], Justine Shults[1], Jiang Bian[4,5], Patrick B. Ryan[6], Christopher B. Forrest[3] & Yong Chen[1]✉

Clinical research networks (CRNs), made up of multiple healthcare systems each with patient data from several care sites, are beneficial for studying rare outcomes and increasing generalizability of results. While CRNs encourage sharing aggregate data across healthcare systems, individual systems within CRNs often cannot share patient-level data due to privacy regulations, prohibiting multi-site regression which requires an analyst to access all individual patient data pooled together. Meta-analysis is commonly used to model data stored at multiple institutions within a CRN but can result in biased estimation, most notably in rare-event contexts. We present a communication-efficient, privacy-preserving algorithm for modeling multi-site zero-inflated count outcomes within a CRN. Our method, a one-shot distributed algorithm for performing hurdle regression (ODAH), models zero-inflated count data stored in multiple sites without sharing patient-level data across sites, resulting in estimates closely approximating those that would be obtained in a pooled patient-level data analysis. We evaluate our method through extensive simulations and two real-world data applications using electronic health records: examining risk factors associated with pediatric avoidable hospitalization and modeling serious adverse event frequency associated with a colorectal cancer therapy. In simulations, ODAH produced bias less than 0.1% across all settings explored while meta-analysis estimates exhibited bias up to 12.7%, with meta-analysis performing worst in settings with high zero-inflation or low event rates. Across both applied analyses, ODAH estimates had less than 10% bias for 18 of 20 coefficients estimated, while meta-analysis estimates exhibited substantially higher bias. Relative to existing methods for distributed data analysis, ODAH offers a highly accurate, computationally efficient method for modeling multi-site zero-inflated count data.

The recent advent of "big data" has had significant implications for health care, spawning several advancements in management and analysis of large-scale patient data[1]. Much of this is a result of the widespread adoption of electronic health records (EHRs), patient data collected during routine and emergency clinical visits. Though EHRs are primarily used as a written record of health care delivery, substantial effort has been made in using these data secondarily to generate real-world evidence (RWE), evidence produced as a result of analyzing observational health data outside of clinical trials. RWE quality can be substantially improved from analyzing pooled data, patient records aggregated across health systems. This is especially true in the context of studying rare outcomes, where outcome prevalence at any single institution may not be large enough to result in an analysis with meaningful conclusions. Pooled patient data from several healthcare systems also allows for study of a sample likely to be more representative of the population of interest.

While pooling patient data from several institutions is ideal, doing so is not always possible. Regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General

[1]Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [3]Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA, USA. [4]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, USA. [5]Cancer Informatics Shared Resource, University of Florida Health Cancer Center, Gainesville, FL, USA. [6]Janssen Research and Development, Titusville, NJ, USA. ✉email: Ychen123@upenn.edu

Data Protection Regulation (GDPR) in the European Union often prevent inter-site sharing of patient-level data that is not de-identified[2,3]. Sharing de-identified individual patient data (IPD) can also be dangerous according to several studies demonstrating the susceptibility of these data to re-identification, causing concern among patients[4–6]. Further, significant computational burden associated with storing and analyzing massive datasets makes data centralization less appealing in some settings. As a result of these restrictions and concerns, much interest has been demonstrated recently in distributed clinical research networks (CRNs), multi-site distributed data networks which allow for analyses across institutions without the need for data centralization[7,8]. In a CRN, each individual institution or health system maintains control over its own data, drastically reducing risk of violating patient privacy through avoiding IPD exchange. Examples of CRNs include the National Patient-Centered Clinical Research Network (PCORnet)[9], a CRN with patient data from 348 health systems in the United States, and the Sentinel System, a national CRN for monitoring performance of FDA-regulated medical products[10,11].

To protect patient privacy, CRNs largely invoke methods for synthesizing and analyzing aggregate data, summary measures obtained from individual sites without any information that could reveal patient identity. In comparative effectiveness research performed in CRNs, meta-analysis is frequently used. Meta-analysis, which only requires effect size and variance estimates from each individual site, is easy to implement and widely accepted in medical literature; it is the primary analysis method used in comparative effectiveness studies conducted by the Observational Health Data Sciences and Informatics (OHDSI) collaborative, an international CRN made up of over 100 different databases[12–16]. Beyond statistical approaches to privacy preservation, several other methods are available for making multi-site analyses of patient data more secure. Differential privacy methods[17] allow researchers to add random noise to patient-level data and obtain results close to those using raw data, while methods incorporating homomorphic encryption[18] produce results identical to those using unencrypted data. Similarly, blockchain technology can be used to implement a secure, decentralized distributed network which eliminates reliance on a coordinating center, useful for Health Information Exchange applications[19]. Swarm Learning and ModelChain are examples using blockchain technology for building and sharing privacy-preserving predictive models across institutions[20,21]. While useful in certain settings, the computation time required for blockchain-based analyses can be a limitation in healthcare settings where both accurate and efficient solutions are desirable.

While suitable for many applications, meta-analysis has been shown to result in biased or imprecise effect estimates in the context of rare events and limited sample sizes[22]. In only sharing site-level point and variance estimates, meta-analysis does not utilize any additional aggregate information that could be obtained from ongoing studies with access to their own patient-level data. Distributed regression methods are an alternative to meta-analysis which leverage access to patient-level data within individual sites, allowing for fitting a regression model distributively across institutions without sharing IPD. Rather than sharing only site-specific regression estimates, distributed regression methods reconstruct or approximate pooled regression estimates (estimates calculated using all pooled patient-level data) using aggregate, summary-level data supplied by each participating database. While several distributed regression algorithms have been developed, many require several rounds of communication among sites until convergence, resulting in analysis that is both time consuming and computationally expensive[23,24]. More recently, a class of non-iterative distributed algorithms has been proposed by Duan et al.[22,25]; these methods use a surrogate likelihood approach to generate estimates comparable to those from pooled analysis using IPD only at the lead site, incorporating aggregate information from collaborating sites to better approximate the complete data likelihood[26]. Methods based on the surrogate likelihood approach are one-shot algorithms, requiring only one or two rounds of non-iterative communication among institutions to offer a communication-efficient alternative for performing distributed regression.

To our knowledge, despite the growing collection of methods for analyzing data in CRNs, no distributed regression method for modeling count outcomes currently exists. Count data are abundant in EHRs, administrative claims, and other sources of electronic health data, with examples including length of stay, number of primary care or emergency department visits, and number of laboratory tests administered. To explore associations between count outcomes and a set of clinical covariates, Poisson or Negative Binomial regression is typically used. In practice, medical count data can be zero-inflated, where zero counts are in excess; zeros often make up the majority of observed counts for rare outcomes, far exceeding the number expected in Poisson or Negative Binomial distributions. In several applications, empirical distributions of zero-inflated counts can also feature a small number of observations with relatively large counts. This is a common occurrence in distributions of health care expenditure, for example, which feature a large proportion of patients with no expenses at one end and a smaller proportion of patients with large expenses at the other[27]. In these settings, one can use hurdle regression, which uses two separate processes for modeling zero and non-zero (positive) counts. The first part models whether an observation will have a zero or positive count, commonly through logistic regression, while the second estimates a count for an observation given that the count is positive, typically using zero-truncated Poisson or Negative Binomial regression. Hurdle regression allows one to separately investigate the effect of covariates on the probability of experiencing an outcome and on the expected frequency of an outcome given that it occurs at least once, improving interpretation in settings where these two processes are driven by different parameters.

We propose a novel method, a one-shot distributed algorithm for hurdle regression (ODAH), to distributively model zero-inflated count outcomes stored in multiple institutions. Using the surrogate likelihood approach, our method for modeling count outcomes is an efficient, non-iterative algorithm which requires two rounds of privacy-preserving communication among sites to generate accurate and precise population-level estimates closely approximating those from pooled analysis. 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 and modeling serious adverse event frequency for colorectal cancer patients. The results from these analyses demonstrate that coefficients produced from ODAH are generally less biased than those from meta-analysis when compared to the gold standard pooled estimate. Our non-iterative ODAH method is both

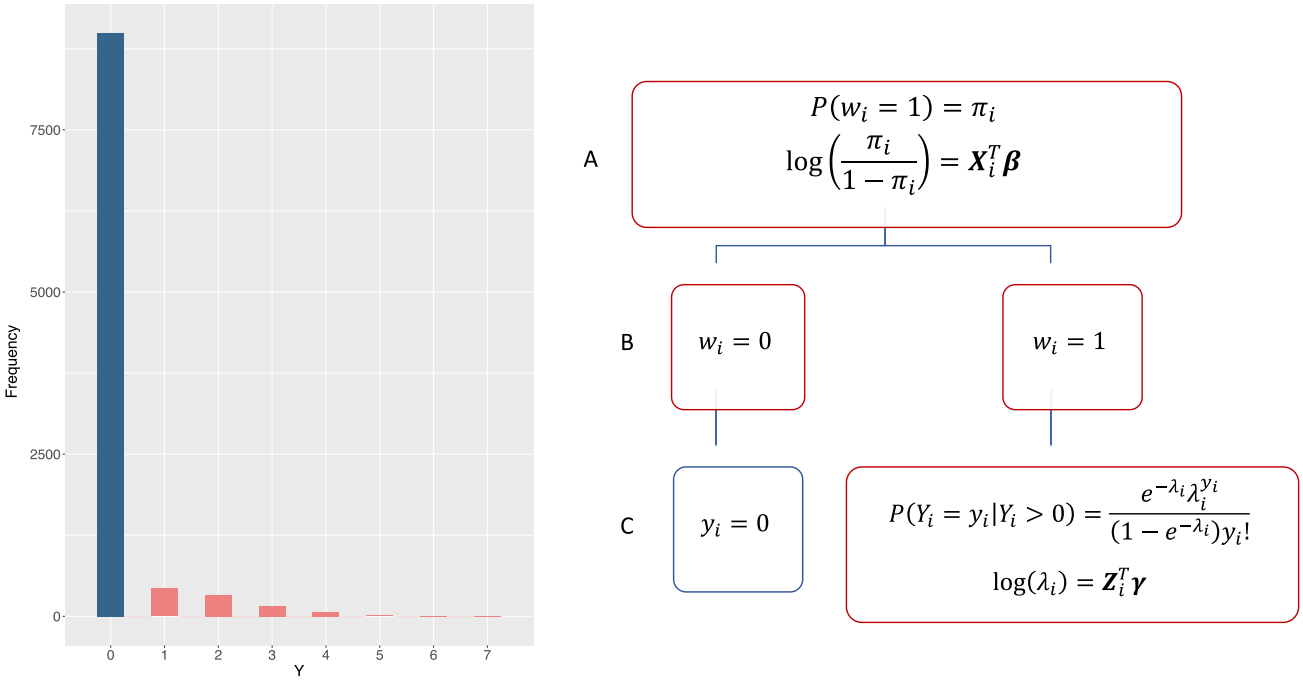

Figure 1. On the left, a histogram displaying counts generated with a Poisson–Logit hurdle distribution with 10% prevalence and a zero-truncated event rate of $\lambda = 1.5$. On the right, a hierarchical diagram visualizing the data generation process in a Poisson–Logit hurdle framework. Independent 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.

communication-efficient and highly accurate serving as a worthwhile method for analyzing zero-inflated count outcomes in CRN and distributed regression settings.

Methods

Poisson–Logit hurdle model. A hurdle model is a two-part model which specifies two separate processes, one for generating zero values and another for generating values given that they are non-zero[28]. The hurdle model is useful for modeling a count outcome with excess zeros, modeling the zero and positive counts independently. Figure 1 depicts a typical zero-inflated distribution of counts, as well as a schematic overview detailing the sequential nature of the Poisson–Logit hurdle model.

In this paper, we invoke the hurdle model to model zero-inflated count outcomes common in healthcare data. To derive our 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 $w_1, w, \ldots, w_n \in \{0, 1\}$ 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 $\boldsymbol{\beta}$:

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

Next, positive counts are modeled using a zero-truncated Poisson model. Let $y_1, y_2, \ldots, y_n \in \{0, 1, 2, \ldots\}$ 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 $P(Y_i = y_i|Y_i\rangle 0) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{(1-e^{-\lambda_i})y_i!}$. Thus, we can write the mixture probability mass function of the Poisson hurdle model as

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

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 $\boldsymbol{\gamma}$:

$$\log(\lambda_i) = \boldsymbol{Z}_i^T \gamma. \tag{3}$$

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

$$L_1(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i X_i^T \boldsymbol{\beta} - \log\left(1 + e^{X_i^T \boldsymbol{\beta}}\right). \tag{4}$$

and

$$L_2(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \left(-e^{Z_i^T \boldsymbol{\gamma}} + Y_i Z_i^T \boldsymbol{\gamma} - \log\left(1 - e^{-e^{Z_i^T \boldsymbol{\gamma}}}\right) - \log(Y_i!)\right). \tag{5}$$

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[29,30]. 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[31,32]. 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 the outcome. This offers an interpretation of estimated model coefficients that is more appropriate in many clinical settings. Further, in having two sets of parameters which can be estimated independently, one avoids the complexity of zero counts coming from a mixture distribution as is the case when using zero-inflated distributions.

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

$$L_{1N}(\boldsymbol{\beta}) = \frac{1}{N} \sum_{j=1}^{K} \sum_{i=1}^{n_j} X_{ij}^T \boldsymbol{\beta} - \log\left(1 + e^{X_{ij}^T \boldsymbol{\beta}}\right) \tag{6}$$

and

$$L_{2N}(\boldsymbol{\gamma}) = \frac{1}{N} \sum_{j=1}^{K} \sum_{Y_{ij}>0} \left(-e^{Z_{ij}^T \boldsymbol{\gamma}} + Y_{ij} Z_{ij}^T \boldsymbol{\gamma} - \log\left(1 - e^{-e^{Z_{ij}^T \boldsymbol{\gamma}}}\right) - \log\left(Y_{ij}!\right)\right) \tag{7}$$

In the CRN context, 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.[26] and later adapted to the clinical data setting by Duan et al.[22,25], 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. The goal with surrogate likelihood estimation is to closely approximate the log likelihood functions for the combined data that we do not have access to, constructing a proxy for the combined-data log likelihoods near a neighborhood of some true parameter value. 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. Since our method is based on approximating the combined-data log likelihoods, an assumption for the algorithm is that data from different sites are homogeneously distributed. Additionally, the outcome being modeled given the covariates should be approximately Poisson-distributed and zero-inflated.

The surrogate log likelihood function for each component of the hurdle model can be expressed as

$$\widetilde{L_1}(\boldsymbol{\beta}) = L_{11}(\boldsymbol{\beta}) + \left\{\nabla L_{1N}(\overline{\boldsymbol{\beta}}) - \nabla L_{11}(\overline{\boldsymbol{\beta}})\right\}\beta + \frac{1}{2}(\boldsymbol{\beta} - \overline{\boldsymbol{\beta}})^2 \left\{\nabla^2 L_{1N}(\overline{\boldsymbol{\beta}}) - \nabla^2 L_{11}(\overline{\boldsymbol{\beta}})\right\} \tag{8}$$

and

$$\widetilde{L_2}(\boldsymbol{\gamma}) = L_{21}(\boldsymbol{\gamma}) + \{\nabla L_{2N}(\overline{\boldsymbol{\gamma}}) - \nabla L_{21}(\overline{\boldsymbol{\gamma}})\}\gamma + \frac{1}{2}(\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma}})^2 \left\{\nabla^2 L_{2N}(\overline{\boldsymbol{\gamma}}) - \nabla^2 L_{21}(\overline{\boldsymbol{\gamma}})\right\}, \tag{9}$$

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

$$\nabla^g L_{1N}(\overline{\boldsymbol{\beta}}) = \frac{1}{N} \sum_{j=1}^{K} n_j \nabla^g L_{1j}(\overline{\boldsymbol{\beta}}) \tag{10}$$

and

$$\nabla^g L_{2N}(\overline{\boldsymbol{\gamma}}) = \frac{1}{N}\sum_{j=1}^{K} n_j \nabla^g L_{2j}(\overline{\boldsymbol{\gamma}}) \tag{11}$$

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 zero-truncated 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

$$\tilde{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \tilde{L}_1(\boldsymbol{\beta}) \tag{12}$$

and

$$\tilde{\boldsymbol{\gamma}} = \operatorname*{argmax}_{\boldsymbol{\gamma}} \tilde{L}_2(\boldsymbol{\gamma}). \tag{13}$$

Well-chosen $\overline{\boldsymbol{\beta}}$ and $\overline{\boldsymbol{\gamma}}$ will increase the accuracy of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\gamma}}$, 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),

$$\overline{\boldsymbol{\beta}} = \frac{\sum_{j=1}^{K} \hat{\boldsymbol{\beta}}_j \boldsymbol{\omega}_j}{\sum_{j=1}^{K} \boldsymbol{\omega}_j}, \text{ with } \boldsymbol{\omega}_j = \frac{1}{\widehat{\boldsymbol{\sigma}}_j^2}. \tag{14}$$

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 $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\gamma}}_1$ obtained via fitting the hurdle model of interest at the lead site. This has been shown to perform well when the lead site is largely representative of the entire multi-site sample and eliminates one round of communication among sites relative to using the meta-analytic initial estimate[22].

ODAH builds upon currently available methods using the surrogate likelihood approach for distributed regression. While the logistic component uses the same model featured in the ODAL algorithm developed by Duan et al.[25] to model the probability of a binary outcome, our method incorporates an additional zero-truncated Poisson component to model the frequency of a given outcome. This extra component is especially useful in settings where significant proportions of patients experience either zero or several instances of an outcome, avoiding a loss of potentially valuable information that would occur if the outcome were dichotomized and analyzed using logistic regression alone.

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 regression. 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 lead site can then initiate ODAH by, following its own hurdle model fitting, computing initial estimates via 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. Figure 2 depicts a schematic diagram for the algorithm.

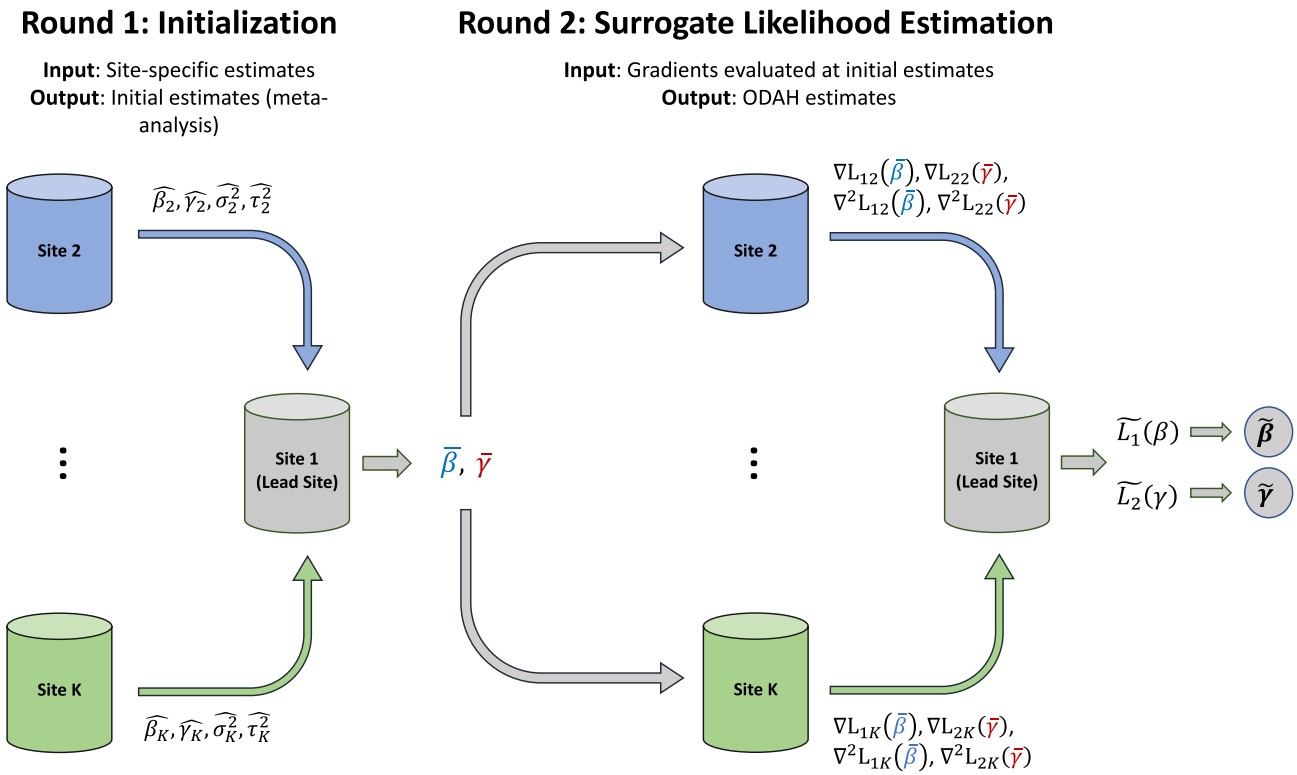

Figure 2. Visual representation of one-shot distributed algorithm for hurdle regression (ODAH). In the initialization round, coefficient ($\hat{\beta}_i, \hat{\gamma}_i$) and variance ($\hat{\sigma}_i^2, \hat{\tau}_i^2$) estimates from fitting separate hurdle models at each collaborating site are sent to the lead site; these estimates are then used together with lead site estimates in a meta-analysis to produce initial estimates ($\bar{\beta}, \bar{\gamma}$) for ODAH, which are sent to each collaborating site. In the surrogate likelihood estimation round, first-order ($\nabla L_{1i}, \nabla L_{2i}$) and second-order ($\nabla^2 L_{1i}, \nabla^2 L_{2i}$) gradients are computed at each site, evaluated at the received initial estimates and sent to the lead site. These gradients are used in conjunction with data from the lead site to construct surrogate likelihood functions $\tilde{L}_1(\beta)$ and $\tilde{L}_2(\gamma)$, which are then maximized to produce surrogate maximum likelihood estimates $\tilde{\beta}$ and $\tilde{\gamma}$.

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 $\widetilde{L_1}(\boldsymbol{\beta})$ (12) and $\widetilde{L_2}(\boldsymbol{\gamma})$ (13).
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 $\widetilde{\boldsymbol{\gamma}}$.

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 patient-level data. Performance is evaluated in terms of bias relative to pooled estimates, coefficients estimated in an analysis where all patient-level data is used; we treat pooled estimation as our gold standard, an ideal scenario featuring centralized data that is typically unattainable in practice. We additionally examine performance of hurdle regression using data only from the lead site, emulating a single-site analysis.

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 our avoidable hospitalization analysis ($X_1 \sim N(3, 2), X_1 \in (0, 18)$), while X_2 was generated using a Bernoulli distribution with the probability of success representing that of public insurance use among patients in the same analysis ($X_2 \sim Bern(0.33)$). Our covariate of interest was X_2, with X_1 assumed to be a confounder. The outcome Y given covariates X_1 and X_2 was generated from the Poisson–Logit hurdle model described in the "Methods" section, using logistic regression to model the process generating zero or positive counts and zero-truncated Poisson regression to estimate counts given that they are positive. Note that while the hurdle model has two components, each of which can use its own unique set of covariates, the sets of covariates making up each component of the model are identical in our simulations. We seek to estimate $\boldsymbol{\beta} = \{\beta_0, \beta_1, \beta_2\}$ and $\boldsymbol{\gamma} = \{\gamma_0, \gamma_1, \gamma_2\}$, each 3×1 vectors of regression coefficients quantifying associations between our simulated count outcome and risk factors.

Motivated by our rare-event 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 (mean event rate for patients in our avoidable hospitalization analysis, denoting number of hospitalizations per year): 5%, 2.5%, 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%. Note that these event rates include zero counts, with smaller event rates corresponding to more severe zero-inflation.

In all settings, we fixed the number of sites $K = 10$ and total population size $N = 200{,}000$. In settings where we vary outcome prevalence or event rate, we set $n_1 = n_2 = \cdots = n_{10}$ so all sites had the same number of

Simulation setting			True parameter values		
Prevalence (%)	Event rate (λ)	n_{lead}/N	β_0	γ_0	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}.

observations. We also explored the effect of the lead site being larger than 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 in terms of bias relative to pooled estimates across 1000 simulations to examine the variability in method performance. In all settings, we assume true coefficient values $\{\beta_1, \gamma_1\} = -1$ and $\{\beta_2, \gamma_2\} = 1$.

Application 1: Pediatric avoidable hospitalization. About one-third of pediatric healthcare costs are associated with hospital admissions, the majority of which are unplanned[33]. Unplanned hospitalizations associated with a diagnosis treatable at the primary care level are considered avoidable[34]. By studying which risk factors are most strongly associated with avoidable hospitalizations (AHs), hospital systems can identify patient subpopulations for which primary care should be improved, ideally leading to an overall reduction in hospital costs or admissions[35]. Because pediatric avoidable 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. Further, the rarity of pediatric AH makes analyses studying this outcome susceptible to zero-inflation, making this application a suitable use case for hurdle regression.

In this analysis, we applied ODAH to study risk factors associated with pediatric AH using data from the Children's Hospital of Philadelphia (CHOP) health system. 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.

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 to explicitly model the counts (Fig. 3). This, combined with substantial zero-inflation, makes Poisson–Logit hurdle regression appropriate for modeling these data. The logistic component of the hurdle model will model the probability of a patient experiencing at least one AH, while the zero-truncated Poisson component will model the total number of hospitalizations for a patient given that they experience at least one.

As in our simulations, we used an identical set of covariates for both hurdle model components and evaluated method performance by calculating relative bias to the pooled estimate for lead site analysis, meta-analysis, and ODAH. To estimate the 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.

Application 2: Serious adverse events. Our second analysis studied a population of patients with colorectal cancer (CRC) 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 frequency of serious adverse events (SAEs). The data analyzed are from the OneFlorida Clinical Research Consortium, containing 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 and Medicare claims, cancer registry data, vital statistics, and EHRs from its clinical partners. These data are centralized in a HIPAA limited dataset that contains detailed patient and clinical variables, including demographics, encoun-

	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Total
	(n = 5456)	(n = 9111)	(n = 7893)	(n = 27,288)	(n = 7996)	(n = 13,074)	(n = 70,818)
Gender							
Female	2589 (47.5%)	4427 (48.6%)	3862 (48.9%)	13,458 (49.3%)	4013 (50.2%)	6494 (49.7%)	34,843 (49.2%)
Male	2867 (52.5%)	4684 (51.4%)	4031 (51.1%)	13,830 (50.7%)	3983 (49.8%)	6580 (50.3%)	35,975 (50.8%)
Caucasian race							
Caucasian	3476 (63.7%)	5508 (60.5%)	4783 (60.6%)	15,747 (57.7%)	4649 (58.1%)	9158 (70.0%)	43,321 (61.2%)
Other	1980 (36.3%)	3603 (39.5%)	3110 (39.4%)	11,541 (42.3%)	3347 (41.9%)	3916 (30.0%)	27,497 (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%)	23,206 (32.8%)
Private/self-pay	3459 (63.4%)	5701 (62.6%)	5554 (70.4%)	17,743 (65.0%)	5519 (69.0%)	9636 (73.7%)	47,612 (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%)	26,410 (96.8%)	7920 (99.0%)	12,678 (97.0%)	69,294 (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.

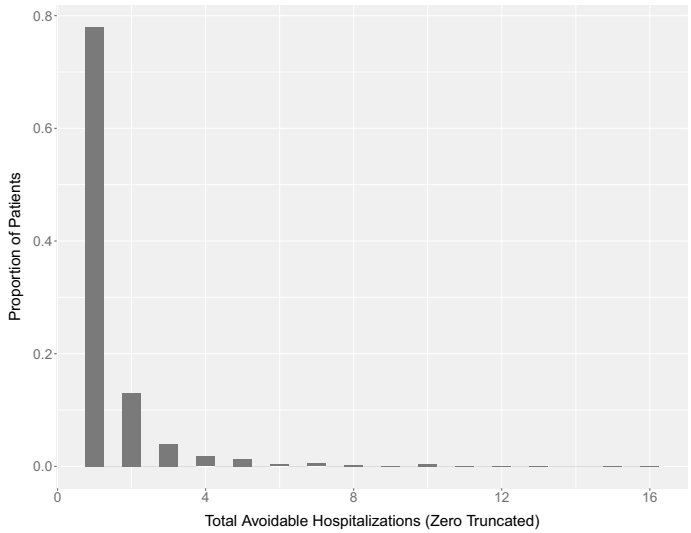

Figure 3. Distribution of total number of avoidable hospitalizations (AHs) for patients with at least one AH in CHOP data sample.

ters, diagnoses, procedures, vitals, medications, and labs, following the PCORnet Common Data Model. 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[36]. Figure 4 shows the geographic locations of OneFlorida partners.

To define an SAE in this analysis, 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

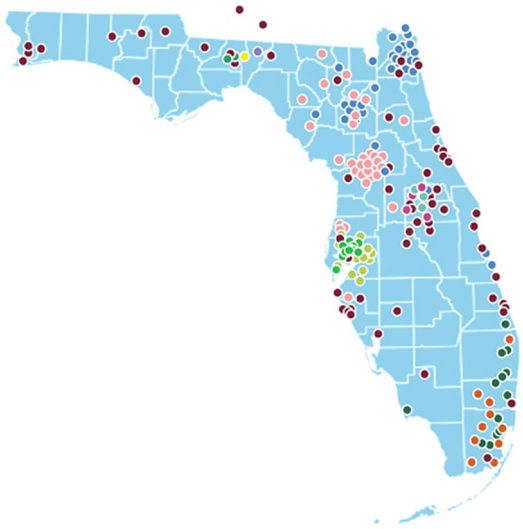

Figure 4. Map detailing locations of OneFlorida clinical partners.

	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)
Patients with 0 SAEs	12 (25%)	53 (23.5%)	151 (39.1%)	216 (32.7%)
Zero-truncated mean (SD)	2.42 (1.56)	2.75 (2.11)	2.42 (1.61)	2.55 (1.82)

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

180 days after first FOLFIRI prescription[37]. 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 AH 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 lead site (Table 3). To evaluate ODAH using these data, we modeled SAE frequency given the extracted clinical information noted above for each patient. We evaluated method performance as we did for our simulations and AH analysis, again using the same set of covariates in each component of the hurdle model.

Results

Simulation study results. Figure 5 depicts simulation results from evaluating method performance across all scenarios described in Table 1. 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

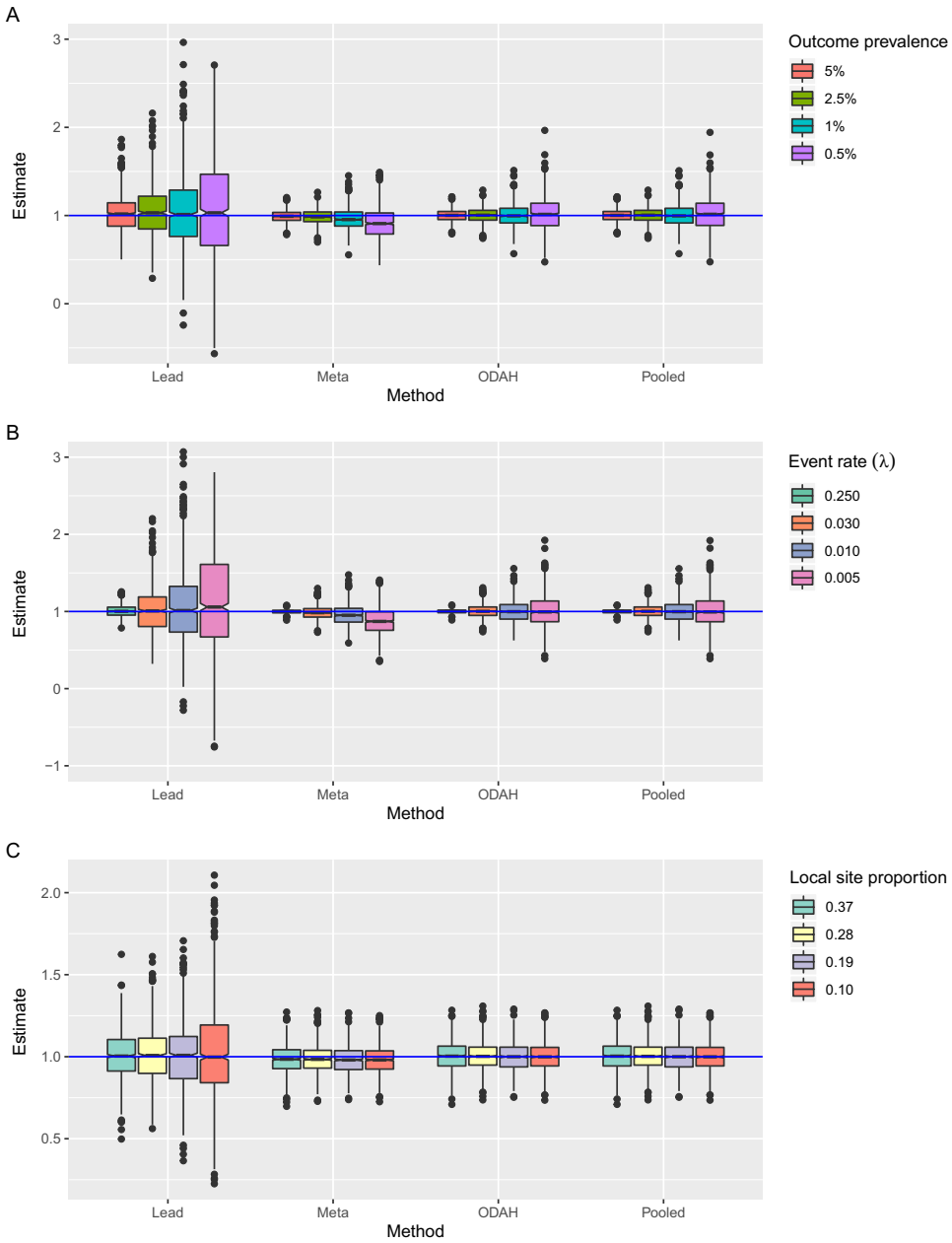

Figure 5. Simulation results for estimating zero-truncated Poisson component covariate γ_2. (**A**) Results for Setting A, fixing $n_{lead}=20{,}000$ and $\gamma_0=-3.6$ ($\lambda = 0.03$) while varying outcome prevalence. (**B**) Results for Setting B, fixing $n_{lead}=20{,}000$ and $\beta_0=-3.7$ (2.5% prevalence) while varying event rate (λ). (**C**) Results for Setting C, fixing $\beta_0=-3.7$ (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 $\gamma_2 = 1$.

therefore present the simulation results for estimating γ_2, leaving β_2 estimation results for the Supplement. Due to select iterations of lead site analysis resulting in outlying estimates, the median bias for the lead site estimate across iterations is reported rather than the mean.

When lead site size and event rate were fixed at 20,000 and $\lambda = 0.03$, respectively, we varied outcome prevalence to study how each method performed relative to pooled analysis, the gold standard (Fig. 5A). In all prevalence levels examined, ODAH performed nearly as well as pooled analysis, with negligible difference in terms of bias and variance of its estimate; 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 (Fig. 5B). 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 $\lambda = 0.005$ (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 (Fig. 5C). Variance of lead site analysis estimates decreased with increasing lead site size.

Application 1: results—pediatric avoidable hospitalization. Figure 6 depicts our avoidable hospitalization (AH) analysis results. Regression coefficient estimates for each covariate in the fitted hurdle model are shown along with their corresponding 95% confidence interval.

Log odds ratio estimates (estimated by the logistic component of the hurdle model) when using ODAH were close to the pooled estimates, with relative bias 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 (estimated by the zero-truncated 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).

Application 2 results: serious adverse events. Results from using ODAH to model serious adverse event (SAE) frequency in colorectal cancer patients using data from OneFlorida are shown in Fig. 7, displayed similarly to the CHOP AH results. In this application, we again see our method exhibiting low bias relative 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 relative 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 zero-truncated Poisson component (compared to 5.3% and 1.8% for ODAH, respectively).

Discussion

We introduced a non-iterative, privacy-preserving algorithm for performing distributed hurdle regression with zero-inflated count outcomes. As demonstrated by simulations and two real-world EHR applications, our method consistently produced parameter estimates comparable to and sometimes more accurate than those produced by meta-analysis. Our method's utility is especially evident in settings featuring a count outcome with marked zero-inflation and very low event rate, as we demonstrated the tendency of only meta-analysis to produce biased estimates under these circumstances in both simulations and real-data analysis. We also showed the benefit of using ODAH over meta-analysis in settings where use of individual site estimates alone may not result in accurate population-level estimation. In the analysis modeling SAE frequency, bias relative to pooled estimation for meta-analysis was greatest for the Hispanic ethnicity coefficient. The proportion of Hispanic patients at each site in this analysis was 4%, 25%, and 58.5%, potentially resulting in highly varied individual-site estimates for this coefficient and resulting in biased meta-analysis estimates. In addition to site-level estimates, our method incorporates aggregate information in the form of gradients to better approximate the complete data likelihood and result in lower bias relative to pooled estimates.

There are several advantages to using our method for performing privacy-preserving data analysis. By using distributed regression, our approach is well-suited for multi-site studies which are ongoing. 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 best suited for studies 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. Further, 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

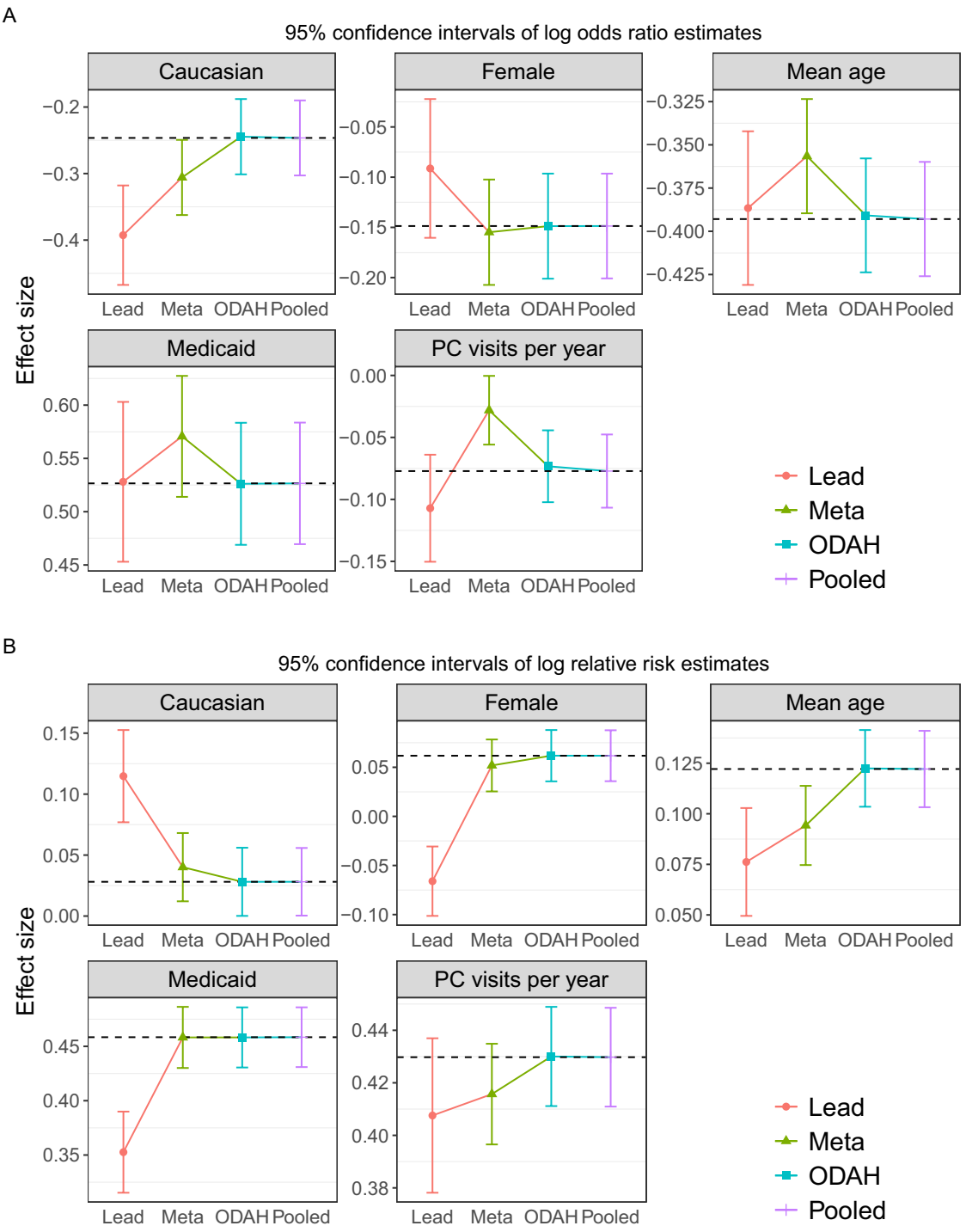

Figure 6. Plots depicting results from CHOP avoidable hospitalization analysis. Log odds ratio (**A**) and log relative risk (**B**) estimates (along with corresponding 95% confidence intervals) for each covariate in the fitted hurdle model. Dashed horizontal line represents pooled estimate, our gold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

site is as large as possible; this helps to ensure the surrogate likelihood is a close approximation to the complete data likelihood.

In terms of limitations of our method, one is that it assumes 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 from each collaborating site. This implicitly assumes that study data are independent and identically distributed across all sites, which may not hold in some real-world settings. As evidenced by Fig. 8, geographical heterogeneity among the patient population

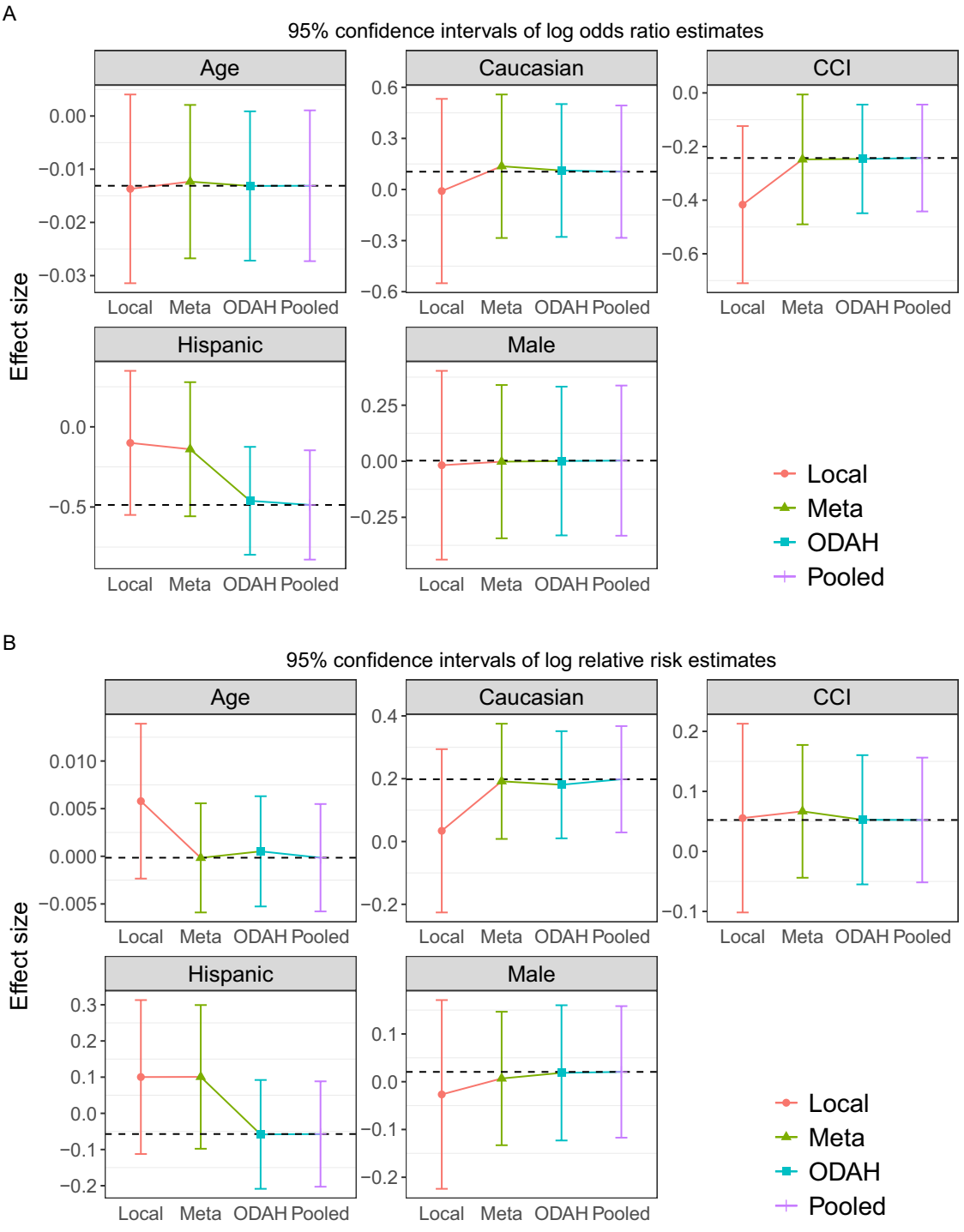

Figure 7. Plots depicting results from OneFlorida serious adverse event application. Log odds ratio (**A**) and log relative risk (**B**) estimates (along with corresponding 95% confidence intervals) for each covariate in the fitted hurdle model. Dashed horizontal line represents pooled estimate, our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

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 relatively homogeneous subsets of institutions. Our group is currently working to develop distributed regression methods which can explicitly model site-specific effects. Additionally, the zero-truncated Poisson component of our method does not currently account for overdispersion in the outcome. Overdispersion in count data is common and should be accounted for when necessary to ensure calculation of robust standard errors. A distributed regression method for modeling count

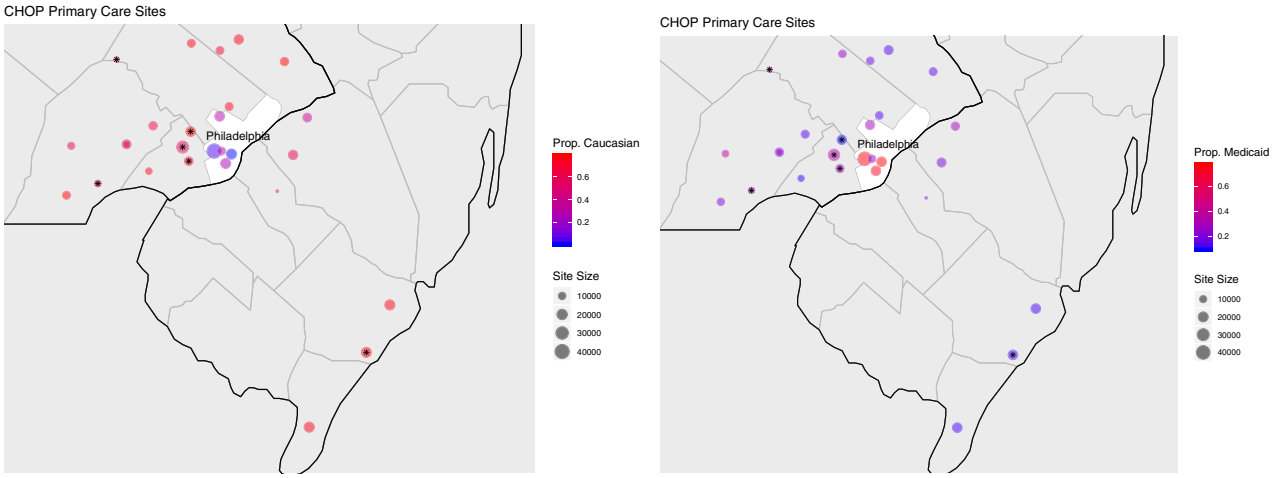

Figure 8. Geographical map of 27 CHOP primary care sites across greater Philadelphia region. In the left map, the proportion of patients of Caucasian race are depicted for each site. In the right map, the proportion of patients using public insurance (Medicaid) at each site is depicted. The size of each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

outcomes which accounts for overdispersion is currently being developed by our group. In both of our real data applications, we did not find strong evidence of overdispersion. Another limitation of this study is that we did not explore method performance in the context of a large number of covariates. While our simulations and data applications featured relatively small collections of risk factors, analyzing a larger collection of covariates may be of interest in big data settings, so evaluating our method in this context would be useful. 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 real data, particularly in terms of covariate imbalance. For example, 52% of patients in the CHOP data 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.

Our group continues to construct methods for performing non-iterative, privacy-preserving distributed inference, ideal for use within CRNs which seek to collaborate on analyses without sharing patient-level data. We seek to create distributed methods for the types of outcomes most common in healthcare, so far producing methods for modeling binary[25], time-to-event[22], and now zero-inflated count outcomes. Further, we look to develop distributed methods with a greater emphasis on data security, incorporating techniques such as homomorphic encryption as was done in works concerning distributed linear and logistic regression[17,38]. While additional methods are being developed and implemented, we believe ODAH is worthy of consideration when one seeks to perform distributed regression on zero-inflated count outcome data.

The code for ODAH is available within the "pda" package in R. Instructions for package installation can be found at https://github.com/Penncil/pda. Details concerning all methods our group has developed for performing distributed regression can be found at https://PDAmethods.org; an overview and sample code for ODAH are available at https://PDAmethods.org/portfolio/odah/. To implement ODAH, one institution will serve as the lead site and coordinate the analysis with other collaborating sites. In order for institutions to collaborate with one another, all data being analyzed must adhere to a common data model to ensure that the same data definitions are used across institutions. Additionally, all institutions must analyze the exact same set of variables. Once these requirements have been met, the procedure outlined in the Methods section can be followed to conduct a privacy-preserving distributed analysis using ODAH.

Conclusion

We introduced an accurate, communication-efficient, privacy-preserving algorithm (ODAH) for performing distributed hurdle regression for settings featuring a zero-inflated count outcome. By only requiring patient-level data from one site, we limit between-site communication to sharing only aggregate statistics in at most two rounds, preserving patient privacy and keeping necessary data exchange to a minimum. In an extensive simulation study and two real-world data analyses, ODAH exhibited higher estimation accuracy than meta-analysis, most notably in the context of rare events. We believe ODAH can be a useful method for analyzing zero-inflated count outcomes in a clinical research network where patient-level data cannot be shared.

Received: 11 April 2021; Accepted: 13 September 2021
Published online: 04 October 2021

References

1. Murdoch, T. B. & Detsky, A. S. The inevitable application of big data to health care. *JAMA* **309**(13), 1351–1352 (2013).

2. Arellano, A. M., Dai, W., Wang, S., Jiang, X. & Ohno-Machado, L. Privacy policy and technology in biomedical data science. *Annu. Rev. Biomed. Data Sci.* **1**, 115–129 (2018).
3. Phillips, M. International data-sharing norms: from the OECD to the General Data Protection Regulation (GDPR). *Hum. Genet.* **137**, 575–582 (2018).
4. Benitez, K. & Malin, B. Evaluating re-identification risks with respect to the HIPAA privacy rule. *J. Am. Med. Inform. Assoc.* **17**(2), 169–177. https://doi.org/10.1136/jamia.2009.000026 (2010).
5. Jiang, X., Sarwate, A. D. & Ohno-Machado, L. Privacy technology to support data sharing for comparative effectiveness research: a systematic review. *Med. Care* **51**, S58 (2013).
6. McGraw, D. Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data. *J. Am. Med. Inform. Assoc.* https://doi.org/10.1136/amiajnl-2012-000936 (2012).
7. Brown, J. S. *et al.* Distributed health data networks: A practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Med. Care* **48**, S45–S51 (2010).
8. Maro, J. C. *et al.* Design of a national distributed health data network. *Ann. Intern. Med.* **151**(5), 341–344 (2009).
9. Fleurence, R. L. *et al.* Launching PCORnet, a national patient-centered clinical research network. *J. Am. Med. Inform. Assoc.* **21**(4), 578–582 (2014).
10. Brown, J. S., Maro, J. C., Nguyen, M. & Ball, R. Using and improving distributed data networks to generate actionable evidence: The case of real-world outcomes in the Food and Drug Administration's Sentinel system. *J. Am. Med. Inform. Assoc.* **27**(5), 793–797 (2020).
11. Robb, M. A. *et al.* The US Food and Drug Administration's Sentinel Initiative: expanding the horizons of medical product safety. *Pharmacoepidemiol. Drug Saf.* **21**(1), 9 (2012).
12. Voss, E. A. *et al.* Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J. Am. Med. Inform. Assoc.* **22**(3), 553–564 (2015).
13. Hripcsak, G. *et al.* Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud. Health Technol. Inform.* **216**, 574 (2015).
14. Hripcsak, G. *et al.* Characterizing treatment pathways at scale using theOHDSI network. *Proc. Natl. Acad. Sci. USA* **113**(27), 7329–7336 (2016).
15. Vashisht, R. *et al.* 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 Netw. Open* **1**(4), E181755-e181755 (2018).
16. Boland, M. R. *et al.* Uncovering exposures responsible for birth season–disease effects: A global study. *J. Am. Med. Inform. Assoc.* **25**(3), 275–288 (2017).
17. Dwork, C. *et al.* The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9**(3–4), 211–407 (2014).
18. Hall, R., Fienberg, S. E. & Nardi, Y. Secure multiple linear regression based on homomorphic encryption. *J. Offic. Stat.* **27**(4), 669 (2011).
19. Kuo, T. T., Kim, H. E. & Ohno-Machado, L. Blockchain distributed ledger technologies for biomedical and health care applications. *J. Am. Med. Inform. Assoc.* **24**(6), 1211–1220 (2017).
20. Warnat-Herresthal, S. *et al.* Swarm learning for decentralized and confidential clinical machine learning. *Nature* **594**(7862), 265–270 (2021).
21. Kuo, T. T. & Ohno-Machado, L. (2018). Modelchain: Decentralized privacy-preserving healthcare predictive modeling framework on private blockchain networks. arXiv preprint arXiv:1802.01746.
22. Duan, R. *et al.* Learning from local to global: An efficient distributed algorithm for modeling time-to-event data. *J. Am. Med. Inform. Assoc. JAMIA* **27**(7), 1028–1036 (2020).
23. 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.* **19**, 758–764 (2012).
24. Lu, C. L. *et al.* WebDISCO: A web service for distributed cox model learning without patient-level data sharing. *J. Am. Med. Inform. Assoc.* **22**(6), 1212–1219 (2015).
25. Duan, R. *et al.* Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *J. Am. Med. Inform. Assoc.* **27**(3), 376–385 (2020).
26. Jordan, M. I., Lee, J. D. & Yang, Y. Communication-efficient distributed statistical inference. *J. Am. Stat. Assoc.* **114**(526), 668–681 (2019).
27. Deb, P. & Norton, E. C. Modeling health care expenditures and use. *Annu. Rev. Public Health* **39**, 489–505 (2018).
28. Cameron, A. C. & Trivedi, P. K. *Regression analysis of count data* (Cambridge University Press, 1998).
29. Rose, C. E., Martin, S. W., Wannemuehler, K. A. & Plikaytis, B. D. On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *J. Biopharm. Stat.* **16**(4), 463–481 (2006).
30. Pittman, B., Buta, E., Krishnan-Sarin, S., O'Malley, S. S., 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; https://doi.org/10.1093/ntr/nty072.
31. Neelon, B., Chang, H. H., Ling, Q. & Hastings, N. S. Spatiotemporal hurdle models for zero-inflated count data: Exploring trends in emergency department visits. *Stat. Methods Med. Res.* **25**(6), 2558–2576 (2016).
32. Hofstetter, H., Dusseldorp, E., Zeileis, A. & Schuller, A. A. Modeling caries experience: advantages of the use of the hurdle model. *Caries Res.* **50**(6), 517–526 (2016).
33. Bui, A. L. *et al.* Spending on children's personal health care in the United States, 1996–2013. *JAMA Pediatr.* **171**(2), 181–189 (2017).
34. Lu, S. & Kuo, D. Z. Hospital charges of potentially preventable pediatric hospitalizations. *Acad. Pediatr.* **12**(5), 436–444 (2012).
35. Maltenfort, M. G., Chen, Y. & Forrest, C. B. Prediction of 30-day pediatric unplanned hospitalizations using the Johns Hopkins Adjusted Clinical Groups risk adjustment system. *PloS One.* **14**(8), 0221233 (2019).
36. Bian, J. *et al.* Implementing a hash-based privacy-preserving record linkage tool in the OneFlorida clinical research network. *Jamia Open* **2**, 562–569 (2019).
37. 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
38. Kim, M., Lee, J., Ohno-Machado, L. & Jiang, X. Secure and differentially private logistic regression for horizontally distributed data. *IEEE Trans. Inf. Forensics Secur.* **15**, 695–710 (2019).

Acknowledgements

We sincerely thank our three reviewers for their insightful and constructive feedback which significantly improved this work. All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee.

Author contributions

M.J.E. and Y.C. designed methods and experiments; C.B.F. and M.M. provided data from the Children's Hospital of Philadelphia; J.B. provided data from the University of Florida; M.J.E. and C.L. designed and conducted simulation experiments; M.J.E. and Z.C. conducted data analysis; all authors interpreted the results and provided instructive comments; and M.J.E. drafted the main manuscript. All authors have approved the manuscript.

Funding

Funding was provided by National Institutes of Health (1R01LM012607, R01CA246418, 1R01AI130460, 1R01AG073435) and Pennsylvania Department of Health (4100072543). This work was supported partially through Patient-Centered Outcomes Research Institute (PCORI) Project Program Awards (ME-2019C3-18315 and ME-2018C3-14899).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-99078-2.

Correspondence and requests for materials should be addressed to Y.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.