HHS Public Access

Author manuscript

Pac Symp Biocomput. Author manuscript; available in PMC 2019 March 14.

Published in final edited form as:

Pac Symp Biocomput. 2019 ; 24: 30–41.

ODAL: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites

Rui Duan[#], **Mary Regina Boland**[#], **Jason H. Moore**, and **Yong Chen**[†]

Department of Biostatistics, Epidemiology & Informatics, University of Pennsylvania, 423 Guardian Drive, PA, 19104, USA

Abstract

Electronic Health Records (EHR) contain extensive information on various health outcomes and risk factors, and therefore have been broadly used in healthcare research. Integrating EHR data from multiple clinical sites can accelerate knowledge discovery and risk prediction by providing a larger sample size in a more general population which potentially reduces clinical bias and improves estimation and prediction accuracy. To overcome the barrier of patient-level data sharing, distributed algorithms are developed to conduct statistical analyses across multiple sites through sharing only aggregated information. The current distributed algorithm often requires iterative information evaluation and transferring across sites, which can potentially lead to a high communication cost in practical settings. In this study, we propose a privacy-preserving and communication-efficient distributed algorithm for logistic regression without requiring iterative communications across sites. Our simulation study showed our algorithm reached comparative accuracy comparing to the oracle estimator where data are pooled together. We applied our algorithm to an EHR data from the University of Pennsylvania health system to evaluate the risks of fetal loss due to various medication exposures.

Keywords

birth outcomes; distributed computing; meta-analysis; multi-site analysis; pregnancy; prenatal; surrogate likelihood

1. Introduction

1.1. Integrate evidence from multiple clinical sites

Electronic Health Records (EHR) contain information collected routinely as a part of clinical care. These data include diagnoses, medications, procedures, imaging and clinical notes. Since 2009, the use of EHR has grown tremendously across the nation. This allows for meaningful use of data recorded there [1, 2]. Institutional data integration is a major

[†]Corresponding: ychen123@upenn.edu.

[#]Co-first author

trend in EHR-based research [3, 4]. Integrating data from different institutions or clinical sites allows us to obtain more meaningful sample size and potentially accelerates knowledge discoveries in a more general population. In particular, for studying relatively rare events or conditions, such as complications from invasive procedures, adverse events associated with new medications, association of disease with a rare gene variant, and many others, integrating EHR data from different clinical sites is critical for obtaining more accurate, generalizable and reproducible results [5]. Moreover, because of the healthcare process biases endemic in EHR, it is necessary to validate findings across multiple sites. This allows for assessment of clinical practice bias (e.g., one drug is prescribed more frequently at a particular hospital), race/ethnic disparities in populations that results in differences in the exposure and/or the outcome at a given site and other types of biases that may be due to the specific research database housed at a given institution [6].

To address these issues, the Observational Health Data Sciences and Informatics (OHDSI) consortium was formed (https://ohdsi.org/) for the primary purpose of developing open source tools that would be shareable across multiple sites. They also developed a Common Data Model [7] to enable each site to map their local data to a common shareable framework. This allows for a single script to be run across multiple sites without alteration. This simultaneously minimizes the probability of a database translation error (when a script is translated from one database structure to another to extract the same type of result) while speeding up the time to results.

Many studies have been conducted that have successfully utilized the OHDSI consortium, including a treatment pathways study [8], a birth season – disease risk study [9, 10] and several pharmacovigilance studies [11]. Using multiple sites allows researchers to study geographic variation [8, 10], which can be caused by regional changes in pollution and other exposures [10].

1.2. Distributed Computing

One barrier of institutional data sharing is regularity and government challenges on privacy protection [12]. In general, patient-level information with regards to important outcomes such as presence/absence of a medical condition or important confounders such as comorbidities, race/ethnicity, and age are not possible to share across institutions. As a consequence, current multi-site studies that rely on consortia, such as the OHDSI consortium [8, 10] or the eMERGE network (Electronic Medical Records and Genomics), can only utilize summary statistics that are shared across institutions. This necessitates the use of meta-analysis methods to aggregate signals from across the network [10].

As of 2018, the OHDSI consortium runs each script locally at a given institution and returns results, typically summary statistics (p-values, effect estimates) to the primary investigator for a given protocol. The Shared Health Research Information Network (SHRINE) has constructed a federated query network whereby analyses are run through the network and results are returned to the investigator [13]. If patient-level information were shareable in a privacy-preserving manner, it would enable more sophisticated patient-level statistical modeling and analyses [14].

Distributed Computing is a strategy where a computational goal is achieved by distributively computing its components from multiple sites. With data from multiple clinical sites, statistical analyses can be performed distributively without sharing patient-level information. For example, motivated by the pSCANNER project (patient-centered Scalable National Network for Effectiveness Research), a distributed algorithm for conducting logistic regression, termed as GLORE (Grid Binary LOgistic Regression), was developed and deployed to pSCANNER consortium [12, 15]. Another example was the WebDISCO (a web service for distributed Cox model learning) method for fitting the Cox proportional hazard model [16] on EHR data from multiple clinical sites without sharing individual patient-level data [17]. These methods proved the utility and plausibility of a distributed privacy-preserving computing approach for obtaining results from multiple sites while still adjusting for patient-level covariates [15].

Despite their usefulness and promise, as acknowledged by the investigators, the aforementioned methods [12, 17] require iteratively transferring information across sites, which is time consuming and labor intensive in practice. Such practical limitation could be one of the barriers to adapt distributed algorithms in research consortia. This limitation motivated researchers to develop non-iterative distributed algorithms [18, 19]. A recently published paper by Jordan et al proposed an innovative one-shot distributive computing framework, where the main idea is to construct a surrogate likelihood function through the use of patient-level data from a local site and aggregated information from other sites [20]. This idea was also proposed in distributed analysis for high-dimensional regression with sparsity [21]. In this study, we exercise the surrogate likelihood idea in logistic regression and develop a One-shot Distributed Algorithm to perform Logistic regressions (termed as ODAL). A major advantage of the proposed method, inherited from the merits of the surrogate likelihood [20], is that it only requires synthesizing summary statistics from multiple clinical sites *once*. Compared to algorithms that require iterative communication across sites, it is more practical to be deployed in research consortia.

2. Material and Method

In this section, we first present our motivating problem, then introduce our proposed method, and describe the design of simulation studies for evaluating the performance of our method.

2.1. Clinical Cohort and Motivating Problem

We extract females treated at one of the hospitals and/or clinics that comprise the University of Pennsylvania health system (abbreviated as UPenn). UPenn clinics are located in the entire Philadelphia Metropolitan area, which includes Delaware and Southern New Jersey. A pregnancy is defined as ‘normal’ if the woman was coded with any of the Z34 ICD-10 codes or a V22 ICD-9 code. A pregnancy is labeled as ending in fetal loss if any ICD-9 code is used within 630 through 639 or O00-O08 in the ICD-10 system. A similar fetal loss definition was used previously [22]. We only include patients who were prescribed or listed as taking at least 1 of the top 100 prescribed medications within 1 year prior to the first diagnosis of either a fetal loss or a normal pregnancy. The demographics of our cohorts are

given in Table 1. P-values for differences between the fetal loss cohort and the normal pregnancy cohort are determined using a t-test. The variable race is dichotomized as white versus non-white in our models. The weight and BMI variables are averages across an individual's entire medical record. The statistics reported in Table 1 are excluding those with 0 weight or 0 BMI (i.e., indicating that no entries are available for those parameters). However, because the average weight and BMI is computed across the individual's entire record the value is smaller for those with longer records containing null entries.

Our proof-of-concept study involves predicting pregnancy outcome: fetal loss versus normal pregnancy. We include 4 relevant demographic covariates: age, race, Body Mass Index (BMI), and weight. We include our 'exposure' term of interest – namely the medication exposure. We ran our algorithm for each of the top 100 medications (ranked by drug prevalence) prescribed within 1 year prior to the pregnancy outcome while adjusting for the 4 demographic confounders. For purposes of this study, we randomly assign each pregnancy id to one of ten clinic IDs to ensure that an equal proportion of data is assigned to each of the ten clinics (approximately 3,557 pregnancies per clinic).

2.2. Algorithm

In this subsection, we introduce the distributed algorithm ODAL. First, we introduce the needed notations. We denote Y to be a binary outcome and z to be a $(p-1)$-dimensional vector, which contains the exposure of interest and potential confounders to be adjusted in a regression model. Let $x = (1, z)$. Suppose we have N observations from K different sites. Without loss of generality, we assume that each site contains n observations, noting that the algorithm also applies to sites with unequal sample sizes. Let (x_{ij}, Y_{ij}) denotes the i-th observation in the j-th site. Under the assumption of a logistic regression model, the log likelihood function for the combined data can be written as

$$L(\beta) = \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{n}\left[Y_{ij}x_{ij}^{T}\beta - log\left\{\left(1 + exp\left(x_{ij}^{T}\beta\right)\right)\right\}\right],$$

where β is a p-diemsional vector including the regression intercept and coefficients. Since the individual patient-level information is not allowed to be transferred across sites, we cannot obtain $L(\beta)$ directly. To tackle this challenge, we apply Taylor expansion on the log likelihood function (1) around an initial value $\bar{\beta}$, and obtain

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

Suppose we have the full access to the data stored in a local site (without loss of generality, assume it is the site 1). The log-likelihood at the local site can be written as

$$L_1(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_{i1}x_{i1}^{T}\beta - log\left\{(1 + exp\left(x_{i1}^{T}\beta\right)\right\}\right]. \quad (2)$$

Similarly, we can expand the local log likelihood function $L_1(\beta)$ around an initial value $\bar{\beta}$,

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

Using the idea from Jordan et al. 2018 [20], the higher order terms of the local likelihood $L_1(\beta)$ in (2) can be used to approximate the higher order terms of the combined likelihood $L(\beta)$ in (1), resulting in the following *surrogate likelihood* function after dropping the constant terms,

$$\tilde{L}(\beta) = L_1(\beta) + \left\{ \frac{1}{K} \Sigma_{k=1}^{K} \nabla L_k(\bar{\beta}) - \nabla L_1(\bar{\beta}) \right\} \beta, \quad (3)$$

where $\nabla L_k(\beta) = \frac{1}{n} \sum_{i=1}^{n} \left[Y_{ik} - exp\left(x_{ik}^T \beta\right) / \left\{ 1 - exp\left(x_{ik}^T \beta\right) \right\} \right] x_{ik}$.

There are several notable features of the above surrogate likelihood. First, the terms $L_1(\beta)$ and $\nabla L_1(\bar{\beta})$ can be calculated using data from the local site. Secondly, the term $\nabla L_k(\bar{\beta})$ can be computed from the site k and transferred to the local site. Note that each $\nabla L_k(\bar{\beta})$ is with dimension p and contains only aggregated information. Therefore, the information transferring maintains low communication cost and is privacy preserving. The ODAL estimator is then obtained locally by minimizing the surrogate likelihood function in equation (3), i.e.

$$\tilde{\beta} = \arg\max_{\beta} \tilde{L}(\beta).$$

Regarding the initial value $\bar{\beta}$, a nature choice of $\bar{\beta}$ is the maximum likelihood estimator of the local likelihood $L_1(\beta)$. A detailed algorithm is outlined below.

Algorithm: ODAL

1. Initial value: obtain $\bar{\beta} = \arg\max_{\beta} L_1(\beta)$. using data in the local site (i.e., site 1), where $L_1(\beta)$ is the log likelihood of logistic regression defined in equation (2)
2. Initial communication: transfer $\bar{\beta}$ to the other sites (i.e., sites 2, 3, …, K)
3. For j = 2 to K
4. **do** compute $\nabla L_j(\bar{\beta})$, where $L_j(\beta)$ is defined similarly as in equation (2)
5. transfer $\nabla L_j(\bar{\beta})$ to the local site
6. **end**
7. Compute the surrogate likelihood $\tilde{L}(\beta)$ defined in equation (3)
8. obtain $\tilde{\beta} = \arg\max_{\beta} \tilde{L}(\beta)$
9. **return** $\tilde{\beta}$

2.3. Simulation Design

To evaluate the empirical performance of the ODAL method, we consider a setting where a binary outcome is associated with two continuous risk factors and two binary risk factors. We generate the two continuous variables from a standard normal distribution $N(0,1)$ and a uniform distribution $U(0,1)$ respectively. The two binary variables are generated from Bernoulli distributions with probability 0.1 and 0.5 respectively. Slightly different from the previous notation, we let x denote the vector of all the risk factors. The outcome Y is generated from Bernoulli distribution, with the conditional probability satisfying the logistic regression model,

$$logit(Pr(Y = 1|x)) = \alpha + x^T\beta,$$

where $logit(p) = log\{p/(1 - p)\}$, β is the vector of coefficients and α is the regression intercept.

To mimic a distributed research network, we generate total number of N subjects and randomly divide them into K sub-datasets. The local dataset is set to be the first sub-dataset and the number of subjects of the local dataset is n. We design the simulation study to investigate the relative accuracy of the ODAL compared to the two methods:

- **i.** the pooled estimator: the individual patient-level data pooled from all clinical sites are used, which can serve as a gold standard for the best possible accuracy; i.e., the estimate that maximizes the log likelihood in equation (1);
- **ii.** (ii) the local estimator: only individual patient-level data from the local site are used; i.e., the estimate that maximizes the log likelihood in equation (2).

We use mean square error (MSE) to summarize the performance of the three estimators and consider the following four scenarios:

- **A.** We randomly generate data for N patients, and evenly divide them into 10 sites. We increase N from 1000 to 10000. This reflects a setting where a network, such as PEDSnet (the National Pediatric Learning Health System), contains a fixed number of pediatric hospitals, but the number of patients increases over time and is updated quarterly [23],
- **B.** We randomly generate data from K sites each has 1000 patients, and increase K from 2 to 100. This is a setting where a consortium involves a growing number of clinical sites, and the number of patients per site is relatively stable. For example, the Hospital Compare dataset (https://www.medicare.gov/hospitalcompare/search.html) contains results from Meaningful Use measures (e.g., 30-day réadmissions) for increasing number of hospitals reporting those measures, however the average hospital size remained relatively constant.
- **C.** We randomly generate data for 10000 patients, and evenly divide them into K sites. We increase K from 2 to 100. This setting is included to investigate the relative performance of the ODAL for a small versus large number of clinical sites, while holding the total number of patients fixed. Depending on how the

data are stored in each hospital, the investigators may choose to perform a distributive analysis on the hospital-level or on the clinic-level.

D. We randomly generate data for 10000 patients divide them into 10 sites. The local site has sample size n, and other 9 sites evenly split the rest of data. We increase n from 100 to 9100. This setting is to investigate the performance of the ODAL when the relative size of the local site, compared to the total number of patients, increases from a small percentage to a large proportion. For example, OHDSI contains many sites of varying sizes from 0.5 million patients to hundreds of millions of patients. Depending on where an investigator is located, the 'local' dataset will vary with regards to the proportion of the dataset as a whole.

3. Results

3.1. Simulation Results

Figure 2 presents the mean square errors of the ODAL, the pooled and the local estimators under four different scenarios. Overall, it shows that in all considered scenarios, the ODAL provides estimates with comparable accuracy as the best possible pooled estimates. In Setting A, where number of sites is fixed and each site has relatively the same number of subjects, ODAL can reach almost the same accuracy as the pooled estimator when total sample size is relatively large. When total sample size is limited, ODAL can still provide much more accurate estimation than the local estimator (MSE of the local estimator is 15 times higher than MSE of ODAL when $N = 1000$). *This suggests that by borrowing simple gradient information* $\nabla L_k(\bar{\beta})$ *from other sites, the ODAL estimate gained substantial statistical efficiency compared to the estimate using the data at the local site alone*. Setting B shows that by borrowing information from more sites, the accuracy of estimation increases. In addition, the ODAL and the pooled estimators provide estimates with negligible difference in accuracy.

Setting C shows that by dividing a fixed number of subjects into increasing number of sites, as expected, the performance of the pooled estimator stays the same. ODAL performs as good as the pooled estimator when the number of sites is relatively small. With increasing number of sites, ODAL has slightly increased amount of error, but is much more accurate compared to the local estimator (MSE of the local estimator is 13 times of the MSE of the ODAL estimate when $K = 100$). The results from Setting C suggest that ODAL can guarantee reasonable accuracy even when the number of sites are moderately large. Such investigation also provides quantitative guidance on choosing between performing the distributed analysis at the clinic level (relatively large number of sites) or the hospital level (relatively small number of sites). Setting D considers the influence of number of subjects contained in the local sites on the accuracy of each methods. As expected, the local estimator performs worse with smaller number of subjects in the local site. The change of local sample size does not influence the performance of the pooled estimator since the total sample size is fixed. Compared to the pooled estimator, the ODAL performs almost the same where the ratio of MSE decreases from 1.22 to 1.00 with the increase of local sample size.

The distributed algorithm GLORE, which leads to exactly the same estimate as the pooled estimator, requires cross-site iterations until a convergence is reached. In our simulation, the number of iterations to obtain the pooled estimates ranges from 6 to 10 using the glm() function in R 3.4.1. In the case with more covariates involved, it may require larger number of iterations to achieve convergence, which creates a substantial burden in communication across clinical sites.

3.2. Fetal Loss Prediction via ODAL

We apply ODAL to the EHR data described in Section 2.1 to evaluate the risks of fetal loss due to various medication exposures. We include the top 100 medications prescribed within 1 year prior to a normal pregnancy or fetal loss outcome. We randomly assign each of our pregnancies to 1 of 10 clinics to test the performance of ODAL. We include one medication at a time adjusting for maternal age, race/ethnicity (collapsed to a binary variable of White versus non-White), weight and BMI. Figure 3 compares the estimates from ODAL to the pooled estimator. The average relative difference in the odds ratios between ODAL and the pooled estimator is 0.0046 across all 100 medications. This indicates that the result from ODAL is very close to the result that would be achieved if all individual-level data are pooled together for the analysis.

Figure 4 presents the top 10 medications that are positively associated (left panel) and bottom 10 medications that are negatively associated (right panel) with fetal loss. To compare our findings with existing knowledge in the literature, we use information on the pregnancy safety of the drug using the Food and Drug Administration (FDA)'s A-X category system. This information is readily obtainable from drugs.com, a freely available online resource, for drugs and their various therapeutic uses and effects (https://www.drugs.com/). Each drug's FDA category is shown above in Figure 4. Drugs in category A are drugs where no fetal risk has been observed in controlled human studies, category B drugs are drugs with no evidence of fetal risk in animal models but well-controlled human studies are lacking, category C drugs are drugs where fetal risk has been shown in animal models but the effects are unknown in humans while category D and X are drugs with known evidence of some fetal risk in humans and animals [24]. Of the top 10 drugs associated with fetal loss Figure 4, six are either category D or X with known evidence of fetal risk in the literature. Three drugs are category C pain relievers, two are drug combos of Tylenol (acetaminophen) with an opioid (codeine or oxycodone) while ibuprofen is an over-the-counter pain reliever. The only category A or B drug in the top 10 is hydrochlorothiazide (a diuretic that treats hypertension), a category B drug. However, hydrochlorothiazide is considered a category D drug, and contra-indicated in pregnancy, is used to treat pregnancy-related hypertension. Therefore, there is likely a dosage that is fetal toxic. In the ten medications that are negatively associated with fetal loss, we identify 8 types of prenatal vitamins with folic acid, docosahexaenoic acid (DHA) and metoclopramide hcl. These findings are consistent with the literature on the importance of prenatal vitamins to prevent early term miscarriages and fetal loss. For example, it has been suggested by many studies that folic acid has positive impacts on preventing early pregnancy loss [25]. In summary, the ODAL method leads to estimates that are highly consistent with the pooled estimates, and the identified associations are also consistent with our current understanding of these medications.

4. Discussion

The integration of EHR data from multiple healthcare databases increases statistical sample size and heterogeneity of exposure, as well as reduces clinical bias and improves the power of statistical analyses. The rise of large healthcare networks, such as ODHSI, pSCANNER, SHRINE and PEDSnet provide platforms for data integration and evidence synthesis [23]. To avoid sharing individual-level information, distributed algorithms have been developed which can conduct population-level analyses in a privacy-preserving manner. In this paper, we propose a novel privacy-preserving and communication-efficient distributed algorithm to study binary outcomes with a set of risk factors using logistic regression. As demonstrated by our simulation study and the application to fetal loss data analysis, our algorithm provides a close approximation to the pooled estimator where all patient-level information is pooled together.

The communication efficiency of our algorithm comes from two aspects. First, in contrast to the existing iterative algorithms such as GLORE and WebDISCO, our algorithm does not require iterative communication across sites. This is crucial especially in the healthcare field where data and information exchange often require large amount of administrative work and technical support. On the other hand, the intermediate result that need to be transferred in the ODAL method is only the first gradient of the likelihood function evaluated at an initial value, which is a vector with dimension equal to the number of parameters p. In contrast, for algorithms such as GLORE [12], in each iteration, the value of the second gradient of the likelihood function need to be transferred, which is a $p \times p$ matrix. When studying a large amount of risk factors, for example large number of potential confounders or genetic variations, the dimension of the matrices can be big which might cause high communication cost for transferring the data.

On the other hand, ODAL requires access of individual patient-level data for one clinical site, in order to construct the surrogate likelihood function. In situations where individual patient-level data are inaccessible in any site, GLORE is preferred.

The OHDSI consortium consists of many partner institutions where patient-level data sharing is not permissible as this often conflicts with regional legislation. In this instance an individual researcher may have patient-level data available at their given site, but then would deploy their algorithms at other sites without having access to the patient-level data. For these situations ODAL is ideal because aggregated information from other sites is only borrowed once without having access to the patient-level data in those countries and regions where that is impermissible.

Deploying ODAL within OHDSI and other large consortia would enable us to further validate our findings with regards to medications taken within 1-year prior to normal pregnancy or fetal loss diagnoses. Validation of these results and also larger scale assessment of medications that potentially increase the risk of fetal loss is still much needed. Algorithms have been developed to assess the fetal effect of category C medications [22], but these can often be limited by confounding and other local institution-specific biases. Use of ODAL across a large international consortium such as OHDSI would propel adequate assessment of

each drug's fetal toxicity even for those where the effects remain unknown (i.e., category C medications).

In the future, we are planning to extend our method to other types of outcomes, such as continuous, categorical, and time-to-event data. Furthermore, we are developing open-source software packages for directly implementing ODAL on distributed networks. We believe that our algorithm can be a good complement to the existing distributed algorithms.

Acknowledgments

This work is supported in part by the University of Pennsylvania, and National Institutes of Health grants AI116794, DK112217, ES013508, HL134015, LM010098, LM011360, LM012601, TR001263, 1R01LM012607, 1R01AI130460, and the Commonwealth Universal Research Enhancement Program grant from the Pennsylvania Department of Health.

References

1. Blumenthal D and Tavenner M, The "meaningful use" regulation for electronic health records. N Engl J Med, 2010 363(6): p. 501–4. [PubMed: 20647183]
2. Andreu-Perez J, et al., Big data for health. IEEE J Biomed Health Inform, 2015 19(4):1193–1208. [PubMed: 26173222]
3. Smith B, et al., The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol, 2007.
4. Goble C and Stevens R, State of the nation in data integration for bioinformatics. J Biomed Inform, 2008 41(5): p. 687–93. [PubMed: 18358788]
5. Sutton AJ, et al., Meta-analysis of rare and adverse event data. Expert Rev Pharmacoecon Outcomes Res, 2002 2(4): p. 367–79. [PubMed: 19807443]
6. Katzan IL and Rudick R.A.J.S.t.m., Time to integrate clinical and research informatics. 2012 4(162): p. 162fs41–162fs41.
7. Overhage JM, et al., Validation of a common data model for active safety surveillance research. Journal of the American Medical Informatics Association, 2011 19(1): p. 54–60. [PubMed: 22037893]
8. Hripcsak G, et al., Characterizing treatment pathways at scale using the OHDSI network. Proceedings of the National Academy of Sciences, 2016 113(27): p. 7329–7336.
9. Boland MR, et al., Birth month affects lifetime disease risk: a phenome-wide method. Journal of the American Medical Informatics Association, 2015 22(5): p. 1042–1053. [PubMed: 26041386]
10. Boland MR, et al., Uncovering exposures responsible for birth season-disease effects: a global study. Journal of the American Medical Informatics Association, 2017 25(3): p. 275–288.
11. Large-scale adverse effects related to treatment evidence standardization (LAERTES): an open scalable system for linking pharmacovigilance evidence sources with clinical data. Journal of biomedical semantics, 2017 8: p. 1–15. [PubMed: 28049518]
12. Wu Y, et al., Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. J Am Med Inform Assoc, 2012 19(5): p. 758–64. [PubMed: 22511014]
13. Weber GM, et al., The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. Journal of the American Medical Informatics Association, 2009 16(5): p. 624–630. [PubMed: 19567788]
14. Boyle D and Rafael N, BioGrid Australia and GRHANITE™™: privacy-protecting subject matching. Studies in health technology and informatics, 2011 168: p. 24–34. [PubMed: 21893908]
15. Ohno-Machado L, et al., pSCANNER: patient-centered Scalable National Network for Effectiveness Research. Journal of the American Medical Informatics Association, 2014 21(4): p. 621–626. [PubMed: 24780722]
16. Cox DR, Regression models and life-tables, in Breakthroughs in statistics. 1992, Springer p. 527–541.

17. Lu C-L, et al., WebDISCO: a web service for distributed cox model learning without patient-level data sharing. 2015 22(6): p. 1212–1219.
18. Zhang Y, Wainwright MJ, and Duchi JC. Communication-efficient algorithms for statistical optimization in Advances in Neural Information Processing Systems. 2012.
19. Battey H, et al., Distributed testing and estimation under sparse high dimensional models. 2018 46(3): p. 1352–1382.
20. Jordan MI, Lee JD, and Yang Y.J.J.o.t.A.S.A., Communication-efficient distributed statistical inference. 2018(just-accepted).
21. Wang J, et al., Efficient distributed learning with sparsity. 2016.
22. Boland MR, Polubriaginof F, and Tatonetti NP, Development of A Machine Learning Algorithm to Classify Drugs Of Unknown Fetal Effect. Scientific reports, 2017 7(1): p. 12839.
23. Forrest CB, et al., PEDSnet: a national pediatric learning health system. 2014 21(4): p. 602–606.
24. Boothby LA and Doering P.L.J.A.o.P., FDA labeling system for drugs in pregnancy. 2001 35(11): p. 1485–1489.
25. Nelen WL, et al., Homocysteine and folate levels as risk factors for recurrent early pregnancy loss. 2000 95(4): p. 519–524.

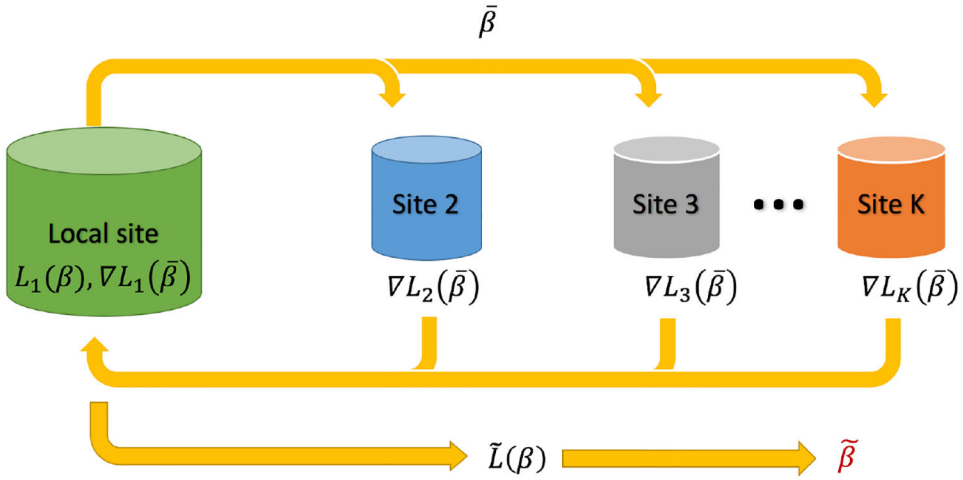

Figure 1.

Schematic illustration of ODAL. Using data from the local site (i.e., site 1), the local estimator $\bar{\beta}$ is calculated and transferred to other sites. The intermediate term $\nabla L_j(\bar{\beta})$ is then evaluated at each site *j (j=2, …, K)* and transferred back to the local site. Combined with $\nabla L_1(\bar{\beta})$ and L_1 (β) we construct the surrogate function $\tilde{L}(\beta)$ in the local site and obtain the ODAL estimator $\tilde{\beta}$ by maximizing $\tilde{L}(\beta)$.

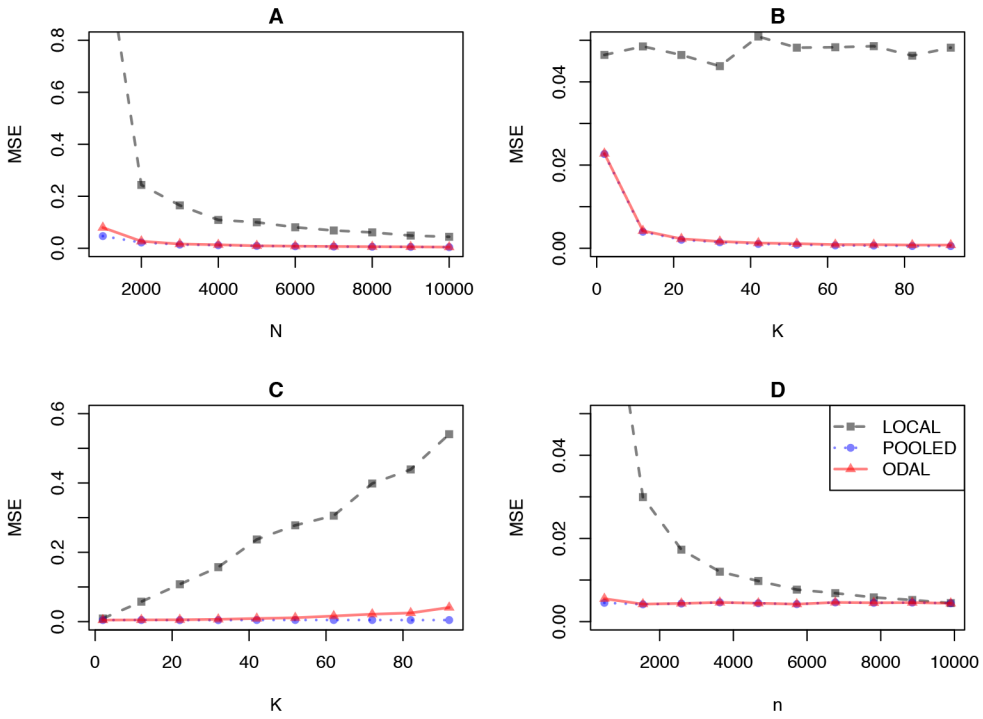

Figure 2.

Mean square errors (MSE) of ODAL, the pooled and the local estimators under settings A, B, C and D. In setting A (upper left panel), we evenly divide N subjects in to 10 sub-datasets, and increase N from 1000 to 10000. In setting B (upper right panel), each site contains 1000 subjects and the number of sites K is then increased from 2 to 100. In setting C (lower left panel), we generate 10000 subjects, and evenly divide them into K sub-datasets, where K increases from 2 to 100. In setting D (lower right panel), we generate 10000 subjects, and divide them into 10 sub-datasets, where the local dataset has n subjects and the other 9 sub-datasets has the equal number of subjects. We increase n from 100 to 9900.

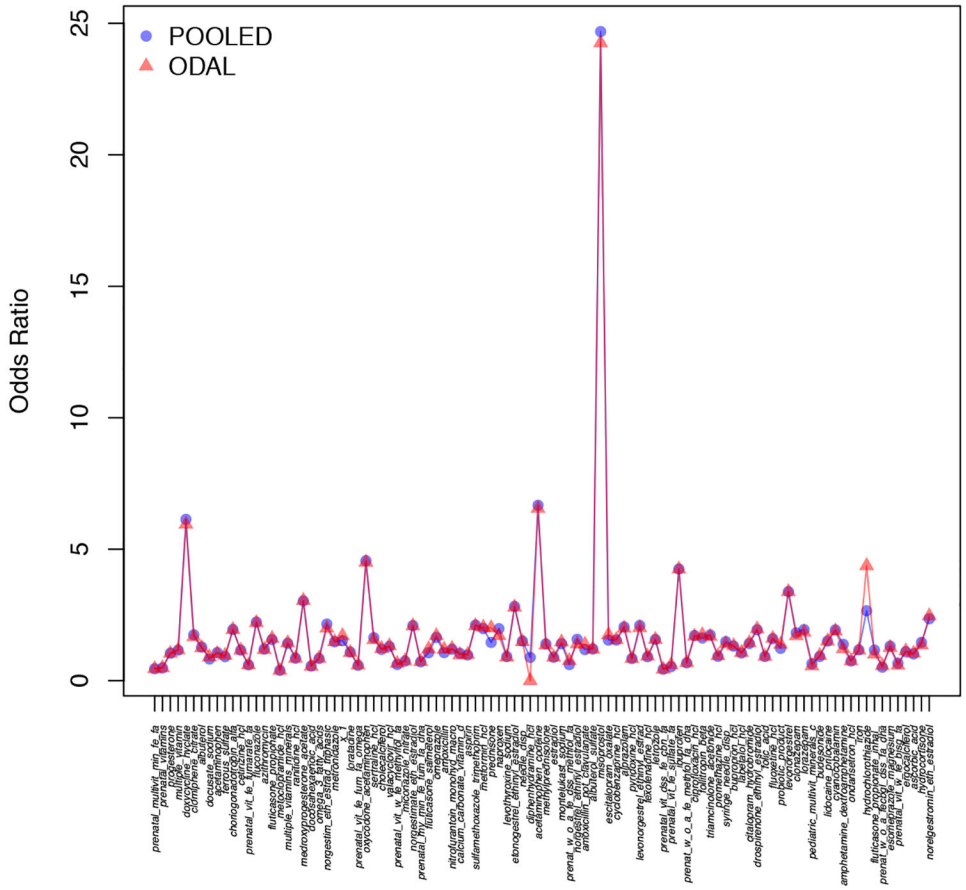

Figure 3:

Odds ratio estimates from the ODAL method (red triangles) and the pooled data (blue circles) for 100 medications and their associations with fetal loss. The 100 medications from left to right are sorted by their prevalence in the population.

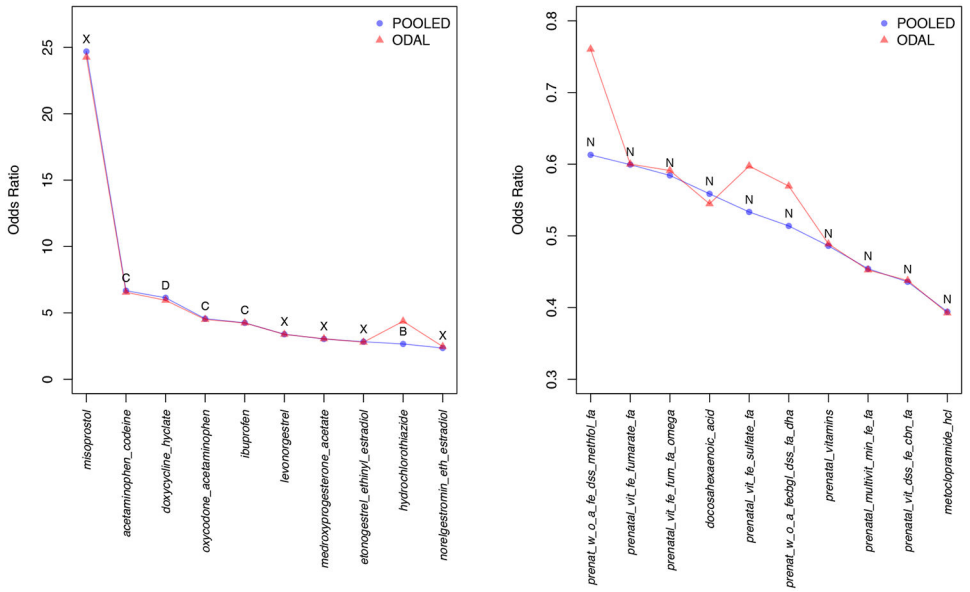

Figure 4:

Odds ratio estimates from ODAL and the pooled estimator for the top 10 medications positively associated (left panel) and negatively associated (right panel) with fetal loss. On the left panel, the ten medications are misoprostol, acetaminophen codeine, doxycycline hyclate, oxycodone acetaminophen, ibuprofen, levonorgestrel, medroxyprogesterone acetate, etonogestrel ethinyl estradiol, hydrochlorothiazide and norelgestromin eth estradiol. On the right panel, the ten acronyms are referring to prenatal vitamins (without vit. A) with DHA, iron, folic acid and docusate sodium; Prenatal vitamins with Iron fumarate, Folic Acid; Prenatal vitamin with Folic Acid and DHA; DHA; Prenatal vitamins with Iron, Sulfate, and Folic Acid; Prenatal vitamin (without vit. A) with DHA, Folic Acid, Extra Iron and Docusate sodium; Prenatal vitamins; Prenatal multi-vitamin with Folic Acid and minimum Iron; Prenatal vitamins with Iron, Docusate sodium, and Folic Acid; Metoclopramide hcl. The letter on each medication shows the FDA assigned pregnancy category, where A, B and C means of no or unknown risk, D and X means of risk. N means the medication is not assigned a category. Detailed interpretations can be found at https://chemm.nlm.nih.gov/pregnancycategories.htm.

Table 1.

Demographics of Pregnancies Treated at UPenn Health System

Demographics	Normal Pregnancy (N=30,810)	Fetal Loss (M=4,763)	P-value
Race			
White *	13911 (45.2%)	2291 (48.1%)	
African American	12918 (41.9%)	1871 (39.3%)	
Other	1916 (6.2%)	274 (5.8%)	
Asian	2065 (6.7%)	327 (6.9%)	
Age	29.40	32.15	<0.001
Weight (pounds)	126.26	115.43	<0.001
Body Mass Index	19.06	16.61	<0.001

* For race, we only used a binary variable for white versus non-white