


RESEARCH ARTICLE

Open Access



Privacy-protecting estimation of adjusted risk ratios using modified Poisson regression in multi-center studies

Di Shu , Jessica G. Young and Sengwee Toh

Abstract

Background: Multi-center studies can generate robust and generalizable evidence, but privacy considerations and legal restrictions often make it challenging or impossible to pool individual-level data across data-contributing sites. With binary outcomes, privacy-protecting distributed algorithms to conduct logistic regression analyses have been developed. However, the risk ratio often provides a more transparent interpretation of the exposure-outcome association than the odds ratio. Modified Poisson regression has been proposed to directly estimate adjusted risk ratios and produce confidence intervals with the correct nominal coverage when individual-level data are available. There are currently no distributed regression algorithms to estimate adjusted risk ratios while avoiding pooling of individual-level data in multi-center studies.

Methods: By leveraging the Newton-Raphson procedure, we adapted the modified Poisson regression method to estimate multivariable-adjusted risk ratios using only summary-level information in multi-center studies. We developed and tested the proposed method using both simulated and real-world data examples. We compared its results with the results from the corresponding pooled individual-level data analysis.

Results: Our proposed method produced the same adjusted risk ratio estimates and standard errors as the corresponding pooled individual-level data analysis without pooling individual-level data across data-contributing sites.

Conclusions: We developed and validated a distributed modified Poisson regression algorithm for valid and privacy-protecting estimation of adjusted risk ratios and confidence intervals in multi-center studies. This method allows computation of a more interpretable measure of association for binary outcomes, along with valid construction of confidence intervals, without sharing of individual-level data.

Keywords: Distributed analysis, Modified Poisson regression, Multi-center studies, Odds ratio, Privacy protection, Risk ratio

Background

In studies where the outcome variable is binary, a logistic regression model is commonly used for convenient estimation of the adjusted (i.e., conditional on measured covariates) odds ratio comparing exposed to unexposed individuals [1, 2]. However, the odds ratio is not easily interpretable because, unlike the risk ratio, it is not a direct measure of a ratio of probabilities, often of primary interest to patients and clinicians [3]. Although the odds

ratio approximates the risk ratio under the rare disease assumption (e.g., odds of the outcome < 10% in all exposure and confounder categories) [4], it can be quite different from the risk ratio and produce misleading results when this assumption is not met [3, 5–8].

Log-binomial regression can directly estimate adjusted risk ratios without requiring the rare disease assumption, but it is susceptible to non-convergence issues when the maximum likelihood estimators lie near the boundary of the parameter space [9, 10]. Poisson regression is another approach to estimating adjusted risk ratios and does not have any known convergence problems in its parameter

* Correspondence: Di_Shu@harvardpilgrim.org
Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA



space. This approach provides consistent estimates of adjusted risk ratios but incorrect estimates of the variance because it relies on a Poisson distributed, rather than binomially distributed, outcome. In practice, standard implementation of Poisson regression tends to produce conservative confidence intervals [11]. As a solution to these challenges, Zou [12] proposed a modified Poisson regression approach that allows direct estimation of adjusted risk ratios even when the rare disease assumption is not met. This approach avoids the convergence issues typically observed in log-binomial regression and, unlike conventional Poisson regression, provides consistent variance estimates and confidence intervals with the correct nominal coverage.

A growing number of studies are now conducted within multi-center distributed data networks [13]. Such collaborations combine data from multiple sources to generate more reliable evidence using larger and more representative samples. Within these networks, each data-contributing site (i.e., data partner) maintains physical control of their data and may not always be able or willing to share individual-level data for analysis. For example, the Sentinel System is a national program funded by the U.S. Food and Drug Administration to proactively monitor the safety of regulated medical products using electronic healthcare data from multiple data partners [14]. In multi-center studies like those conducted within the Sentinel System, it is often crucial to minimize sharing of sensitive individual-level data to protect patient privacy. The development and applications of analytic methods that enable valid statistical analysis without pooling individual-level data are therefore increasingly important.

Privacy-protecting distributed algorithms to conduct logistic regression analyses have been previously developed [15–18]. To our knowledge, there are currently no distributed algorithms to estimate adjusted risk ratios via modified Poisson regression while avoiding pooling of individual-level data across data partners. In this paper, we propose such an algorithm and provide example R [19] code to implement the algorithm. We also illustrate in simulated and real-world data examples that our algorithm produces adjusted risk ratio estimates and standard errors equivalent to those obtained from the corresponding pooled individual-level data analysis.

Methods

Theory of modified Poisson regression for pooled individual-level data

We begin by describing the general theory of modified Poisson regression in single-database studies. Let X be a vector of covariates, E a binary exposure indicator ($E = 1$ if exposed and $E = 0$ if unexposed), and Y the binary

outcome variable ($Y = 1$ if the outcome occurs and $Y = 0$ otherwise).

Let Z be a vector of information on the exposure and covariates. Specifically, $Z = (1, g(E, X^T))^T$ where $g(E, X^T)$ is a vector containing a specified function of E and X . Assume the risk of the outcome conditional on E and X can be written as

$$P(Y = 1|E, X) = \exp(\beta^T Z) \tag{1}$$

where β is an unknown vector of parameters. An example of model 1 is $P(Y = 1|E, X) = \exp(\beta_0 + \beta_E E + \beta_X^T X)$. In this special case, the risk ratio of the outcome comparing the exposed to unexposed and adjusting for covariates X is given by $P(Y = 1|E = 1, X)/P(Y = 1|E = 0, X) = \exp(\beta_E)$.

Alternatively, we might assume a more flexible model that allows interactions between E and X : $P(Y = 1|X, E) = \exp(\beta_0 + \beta_E E + \beta_X^T X + \beta_{EX}^T EX)$. Under this model, the risk ratio of the outcome comparing the exposed to unexposed and adjusting for covariates X is given by $P(Y = 1|E = 1, X)/P(Y = 1|E = 0, X) = \exp(\beta_E + \beta_{EX}^T X)$, which depends on the value of X .

Suppose we have an independent and identically distributed sample of size n . For each individual i , the following variables are measured: Let X_i be a vector of covariates, E_i a binary exposure indicator ($E_i = 1$ if exposed and $E_i = 0$ if unexposed), Y_i the binary outcome variable ($Y_i = 1$ if the outcome occurs and $Y_i = 0$ otherwise), and Z_i a vector of information on the exposure and covariates, i.e., $Z_i = (1, g(E_i, X_i^T))^T$.

Zou [12] provided the theoretical justification for his proposed approach in the setting of a 2 by 2 table (a binary exposure and no covariates). The justification for this approach can be established more generally using the theory of unbiased estimating equations [20]. Provided model 1 is correctly specified, we have $E\{Y - \exp(Z^T \beta) | Z\} = 0$, which leads to the unbiased estimating equation

$$\sum_{i=1}^n \{Y_i - \exp(Z_i^T \beta)\} Z_i = 0 \tag{2}$$

Solving (2) for β gives $\hat{\beta}$, a consistent and asymptotically normal estimator for the true β .

A consistent estimator of the variance of $\hat{\beta}$ is then given by the sandwich variance estimator [20]

$$\widehat{\text{var}}(\hat{\beta}) = \left\{ H(\hat{\beta})^{-1} \right\} B(\hat{\beta}) \left\{ H(\hat{\beta})^{-1} \right\} \tag{3}$$

where $H(\hat{\beta}) = -\sum_{i=1}^n \exp(Z_i^T \hat{\beta}) Z_i Z_i^T$ and $B(\hat{\beta}) = \sum_{i=1}^n \{Y_i - \exp(Z_i^T \hat{\beta})\}^2 Z_i Z_i^T$. Zou [12] referred to this procedure as *modified Poisson regression* because (2) is equivalent to the score equation for the Poisson

likelihood but the variance estimator does not rely on the Poisson distribution assumption (clearly unreasonable for binary outcomes).

Distributed algorithm for conducting modified Poisson regression in multi-center studies

Suppose the n individuals' data are physically stored in K data partners that are unable to share their individual-level data with the analysis center. For $k = 1, \dots, K$, let Ω_k denote the set of indexes of individuals who are members of the k^{th} data partner.

When the individual-level data are available to the analysis center, the estimator $\hat{\beta}$ and its corresponding variance estimator can be obtained with off-the-shelf statistical software. However, when the individual-level data are not available, (2) cannot be directly solved for β to obtain $\hat{\beta}$, and $\widehat{var}(\hat{\beta})$ cannot be directly calculated using (3). Here we describe a distributed algorithm that produces identical $\hat{\beta}$ and $\widehat{var}(\hat{\beta})$ in multi-center studies where individual-level data are not pooled.

We leverage the Newton-Raphson method such that $\hat{\beta}$ can be obtained using an iteration-based procedure with only summary-level information being shared between the data partners and the analysis center in each iteration. The r^{th} iterated estimate of β using the Newton-Raphson method is

$$\beta^{(r)} = \beta^{(r-1)} - \{H^{(r)}\}^{-1} S^{(r)} \tag{4}$$

where $\beta^{(r-1)}$ is the $(r-1)^{\text{th}}$ iterated estimate of β , $S^{(r)} = \sum_{i=1}^n \{Y_i - \exp(Z_i^T \beta^{(r-1)})\} Z_i$, and $H^{(r)} = -\sum_{i=1}^n \exp(Z_i^T \beta^{(r-1)}) Z_i Z_i^T$.

We observe that $S^{(r)}$ and $H^{(r)}$ can be re-written as summation of site-specific quantities:

$$S^{(r)} = \sum_{k=1}^K S_k^{(r)} \tag{5}$$

where

$$S_k^{(r)} = \sum_{i \in \Omega_k} \{Y_i - \exp(Z_i^T \beta^{(r-1)})\} Z_i \tag{6}$$

and

$$H^{(r)} = \sum_{k=1}^K H_k^{(r)} \tag{7}$$

where

$$H_k^{(r)} = -\sum_{i \in \Omega_k} \exp(Z_i^T \beta^{(r-1)}) Z_i Z_i^T \tag{8}$$

Therefore, to calculate $S^{(r)}$ and $H^{(r)}$, each data partner $k = 1, \dots, K$ only needs to calculate and share with the

analysis center the summary-level information $S_k^{(r)}$ and $H_k^{(r)}$.

Next, consider estimation of the variance of $\hat{\beta}$. We observe that

$$H(\hat{\beta}) = \sum_{k=1}^K H_k(\hat{\beta}) \tag{9}$$

where

$$H_k(\hat{\beta}) = -\sum_{i \in \Omega_k} \exp(Z_i^T \hat{\beta}) Z_i Z_i^T \tag{10}$$

and

$$B(\hat{\beta}) = \sum_{k=1}^K B_k(\hat{\beta}) \tag{11}$$

where

$$B_k(\hat{\beta}) = \sum_{i \in \Omega_k} \{Y_i - \exp(Z_i^T \hat{\beta})\}^2 Z_i Z_i^T \tag{12}$$

To calculate $H(\hat{\beta})$ and $B(\hat{\beta})$, each data partner $k = 1, \dots, K$ only needs to calculate and share the summary-level information $H_k(\hat{\beta})$ and $B_k(\hat{\beta})$ after receiving the value of $\hat{\beta}$ from the analysis center. Unlike in the estimation of β , no iterations are needed in the sandwich variance estimation.

We summarize our distributed algorithm for conducting modified Poisson regression in multi-center studies below.

Point estimation

Step 0 (Determination of starting values) The analysis center specifies the starting values for the components of $\beta^{(0)}$ and sends these values to all data partners. Then, for each iteration r until the convergence criteria are met, the following two steps are repeated:

Step 1 (r^{th} iteration of data partners) Each data partner $k = 1, \dots, K$ calculates $S_k^{(r)}$ and $H_k^{(r)}$ using (6) and (8), respectively, based on $\beta^{(r-1)}$ received from the analysis center. All data partners then share the values of $S_k^{(r)}$ and $H_k^{(r)}$ with the analysis center.

Step 2 (r^{th} iteration of the analysis center) The analysis center calculates $S^{(r)}$ and $H^{(r)}$ using (5) and (7), respectively. The analysis center then calculates $\beta^{(r)}$ using (4) and shares the value of $\beta^{(r)}$ with all data partners.

The iteration procedure is considered to have converged when the change in the estimates between iterations is within a user-specified tolerance value. In numerical studies to be presented later, we considered a

convergence criterion to be met at the $(R + 1)^{th}$ iteration if $\max_l |\delta_l^{(R+1)}| < 10^{-8}$, where $\delta_l^{(R+1)} = \beta_l^{(R+1)} - \beta_l^{(R)}$ if $|\beta_l^{(R)}| < 0.01$ and $\delta_l^{(R+1)} = (\beta_l^{(R+1)} - \beta_l^{(R)})/\beta_l^{(R)}$ otherwise, and $\beta_l^{(R)}$ is the l^{th} element of $\beta^{(R)}$. Once achieving convergence, the analysis center shares the final estimate $\hat{\beta}$ with all data partners.

Variance estimation

Step 1 (Calculation of summary-level information by data partners) Each data partner $k = 1, \dots, K$ calculates $H_k(\hat{\beta})$ and $B_k(\hat{\beta})$ using (10) and (12), respectively, and then shares the values of $H_k(\hat{\beta})$ and $B_k(\hat{\beta})$ with the analysis center.

Step 2 (Calculation of the variance estimate by the analysis center) The analysis center calculates $H(\hat{\beta})$ and $B(\hat{\beta})$ using (9) and (11), respectively, and then calculates the estimated variance $\widehat{var}(\hat{\beta})$ using (3).

Due to mathematical equivalence, the above procedure would provide the same point estimates and sandwich variance estimates as the analysis that uses individual-level data pooled across data partners.

Results

Analysis of simulated data

We considered a simulation design that enabled us to assess the performance of the proposed summary-level modified Poisson method in the presence of multiple data partners, multiple covariates (including but not limited to data source indicators), and differences in exposure prevalence and outcome incidence across data partners. Although modified Poisson regression is broadly applicable with rare and common outcomes, here we considered a scenario with common outcomes, where logistic regression would provide biased estimates of adjusted risk ratios.

Specifically, we simulated a distributed network with three (i.e., $K = 3$) data partners and $n = 10000$ individuals

with 5000, 2000, and 3000 individuals contributing data from the first, second, and third data partners, respectively. We considered five covariates X_1, X_2, X_3, X_4 and X_5 . We generated X_1 as a Bernoulli variable with a mean (i.e., $P(X_1 = 1)$) of 0.6, X_2 as a continuous variable following the standard uniform distribution, X_3 as a continuous variable following the unit exponential distribution, X_4 as an indicator that an individual contributed data from the first data partner, and X_5 as an indicator that an individual contributed data from the second data partner.

The exposure E was generated from a Bernoulli distribution with the probability of being exposed ($E = 1$) defined as $1/\{1 + \exp(0.73 - X_1 - X_2 + X_3 - 0.2X_4 + 0.2X_5)\}$, indicating a non-randomized study. This setting led to different exposure prevalences across data partners. The resulting exposure prevalence was approximately 40% overall, 43% for the first data partner, 34% for the second data partner, and 38% for the third data partner. The outcome Y was generated from a Bernoulli distribution with the probability of having the outcome ($Y = 1$) defined as $\exp(\mathbf{Z}^T \beta) = \exp(-0.1 - 0.5E - 0.4X_1 - 0.6X_2 - 0.5X_3 - 0.1X_4 + 0.1X_5)$ such that the true adjusted risk ratio comparing the exposed to unexposed was $\exp(-0.5) = 0.61$. The resulting outcome incidence (i.e., risk) varied across the three data partners. This incidence was about 30% for the entire pooled data, 27% for the first data partner, 35% for the second data partner, and 31% for the third data partner.

As the reference, we first fit a modified Poisson regression model using pooled individual-level data (Table 1). We then implemented our proposed distributed algorithm that did not require sharing of individual-level data to estimate β . Based on the starting value $\beta^{(0)} = \mathbf{0}$, the analysis took seven iterations to converge. The individual-level and summary-level methods produced identical point estimates and sandwich variance-based standard errors (Table 1). The Additional file 1 provides the summary-level information shared between the data partners and the analysis center during each iteration.

Table 1 Point Estimates and Standard Errors Using the Summary-Level Modified Poisson Method and Pooled Individual-Level Data Analysis: Analysis of Simulated Data

Covariates	Summary-Level Modified Poisson Method		Pooled Individual-Level Data Analysis	
	Parameter Estimates	Standard Errors	Parameter Estimates	Standard Errors
Intercept	-0.09702882	0.03751991	-0.09702882	0.03751991
Exposure	-0.48776703	0.03462485	-0.48776703	0.03462485
X_1	-0.38249121	0.02968448	-0.38249121	0.02968448
X_2	-0.62968463	0.05161073	-0.62968463	0.05161073
X_3	-0.50382664	0.02389781	-0.50382664	0.02389781
X_4	-0.09079213	0.03389126	-0.09079213	0.03389126
X_5	0.13274741	0.03814272	0.13274741	0.03814272

Analysis of real-world data

To further illustrate our method, we analyzed a dataset created from the IBM® Health MarketScan® Research Databases, which contain de-identified individual-level healthcare claims information from employers, health plans, hospitals, and Medicare and Medicaid programs fully compliant with U.S. privacy laws and regulations (e.g., Health Insurance Portability and Accountability Act). The study dataset included 9736 patients aged 18–79 years who received sleeve gastrectomy or Roux-en-Y gastric bypass between 1/1/2010 and 9/30/2015. The outcome of interest was any hospitalization during the 2-year follow-up period after surgery. The exposure variable was set to 1 if the patient received sleeve gastrectomy and 0 if the patient received Roux-en-Y gastric bypass. We estimated the risk ratio of hospitalization comparing sleeve gastrectomy with Roux-en-Y gastric bypass using the pooled individual-level data analysis and the summary-level information approach, adjusting for the following covariates identified during the 365-day period prior to the surgery: age; sex; Charlson/Elixhauser combined comorbidity score; diagnosis of asthma, atrial fibrillation, atrial flutter, coronary artery disease, deep vein thrombosis, gastroesophageal reflux disease, hypertension, ischemic stroke, myocardial infarction, pulmonary embolism, and sleep apnea; use of anticoagulants, assistive walking device, and home oxygen; unique drug classes dispensed and unique generic medications dispensed.

Of the 9736 patients in the study dataset, 7877 (81%) patients underwent the sleeve gastrectomy procedure and 1859 (19%) patients had the Roux-en-Y gastric bypass procedure. The outcome event was not rare in the study, with 1485 (19%) sleeve gastrectomy patients and 608 (33%) Roux-en-Y gastric bypass patients having at least one hospitalization during the two-year follow-up period. We randomly partitioned the dataset into three smaller datasets with 2000, 3000 and 4736 patients to create a “simulated” distributed data network. As the reference, the pooled individual-level data analysis produced $\hat{\beta}_E = -0.4632219$ with a standard error 0.0422368, and a 95% confidence interval: $-0.5460061, -0.3804377$. These results corresponded to an adjusted risk ratio of $\exp(\hat{\beta}_E) = 0.63$ with a 95% confidence interval: 0.58, 0.68. Based on the starting value $\beta^{(0)} = \mathbf{0}$, the proposed summary-level modified Poisson method took seven iterations to converge and produced point estimates and sandwich variance-based standard errors identical to those observed in the corresponding pooled individual-level data analysis (Table 2). The adjusted odds ratio from logistic regression was 0.53. As expected, interpreting the estimated adjusted odds ratio as an estimate of the adjusted risk ratio amplified the protective effect of sleeve gastrectomy compared to Roux-en-Y gastric bypass, resulting in an effect estimate that was further from the null (suggesting a 10% greater relative protective effect) than the modified Poisson regression estimate.

As expected, we had difficulty fitting a log-binomial regression model within this bariatric surgery dataset. Under the starting value $\beta^{(0)} = \mathbf{0}$, the first iterated estimate $\beta^{(1)}$ could not be calculated because the formula for $\beta^{(1)}$ under the log-binomial regression model includes $1 - \exp(\mathbf{Z}_i^T \beta^{(0)})$ in the denominator, which takes the value 0 when $\beta^{(0)} = \mathbf{0}$. We also considered three non-zero starting values. We first set the starting values for all parameters to 0.05, but the analysis stopped at the second iteration due to matrix singularity. We then let the starting values be the estimates obtained from a logistic regression model fit using the entire bariatric surgery dataset, and the analysis converged with $\hat{\beta}_E = -0.7369683$. Finally, we specified the starting values as the estimates obtained from the modified Poisson regression fit using the entire bariatric surgery dataset, and the analysis converged with $\hat{\beta}_E = -0.4544135$. These results illustrated the convergence problems of log-binomial regression and the sensitivity of this method to starting values. In comparison, the modified Poisson analysis had no convergence problems and its estimates remained the same as those presented in Table 2 when using these alternative starting values.

Discussion

In this paper, we proposed and demonstrated – in both simulated and real-world data – a method that adapts the modified Poisson approach to directly estimate adjusted risk ratios in multi-center studies where sharing of individual-level data is not always feasible or preferred. Our method produced the same risk ratio estimates and sandwich variance estimates as the corresponding pooled individual-level data analysis without pooling individual-level data across data partners. The required summary-level information does not contain any potentially identifiable individual-level data and therefore offers better privacy protection. Analytic methods like the one we proposed here complement appropriate governance and data use agreements to enable the conduct of multi-center studies, especially when sharing of individual-level data is challenging.

In terms of privacy protection, the proposed summary-level modified Poisson method serves as an intermediate approach between meta-analysis of site-specific effect estimates from modified Poisson analyses and modified Poisson analysis using pooled individual-level data. Compared to meta-analysis of site-specific effect estimates, our method requires more granular information, but the shared information is summary-level without detailed individual-level data. Unlike meta-analysis that generally only produces approximate results, our method produces results identical to those obtained from the corresponding pooled individual-level data analysis.

Table 2 Point Estimates and Standard Errors Using the Summary-Level Modified Poisson Method and Pooled Individual-Level Data Analysis: Analysis of Real-World Data

Covariates	Summary-Level Modified Poisson Method		Pooled Individual-Level Data Analysis	
	Parameter Estimates	Standard Errors	Parameter Estimates	Standard Errors
Intercept	-1.5653147	0.1040146	-1.5653147	0.1040146
Exposure ^a	-0.4632219	0.0422368	-0.4632219	0.0422368
Demographics				
Age (years)	0.0029253	0.0018694	0.0029253	0.0018694
Female sex	0.0860355	0.0478638	0.0860355	0.0478638
Combined comorbidity score	0.0543412	0.0125169	0.0543412	0.0125169
Diagnosis of				
Asthma	-0.0212109	0.0538624	-0.0212109	0.0538624
Atrial fibrillation	0.0835820	0.1249359	0.0835820	0.1249359
Atrial flutter	0.1638268	0.2598721	0.1638268	0.2598721
Coronary artery disease	0.1597157	0.0746721	0.1597157	0.0746721
Deep vein thrombosis	0.1738365	0.1371837	0.1738365	0.1371837
Gastroesophageal reflux disease	-0.0586971	0.0393428	-0.0586971	0.0393428
Hypertension	0.0046885	0.0448018	0.0046885	0.0448018
Ischemic stroke	-0.1592596	0.1986055	-0.1592596	0.1986055
Myocardial infarction	0.2069423	0.1509918	0.2069423	0.1509918
Pulmonary embolism	0.2846648	0.1436404	0.2846648	0.1436404
Sleep apnea	-0.0261550	0.0396516	-0.0261550	0.0396516
Use of				
Anticoagulants	0.1064905	0.1205252	0.1064905	0.1205252
Assistive walking device	0.0908909	0.1394227	0.0908909	0.1394227
Home oxygen	0.0732620	0.1162295	0.0732620	0.1162295
Number of drug dispensing				
Unique drug classes	-0.0579458	0.0166577	-0.0579458	0.0166577
Unique generic medications	0.0673783	0.0136406	0.0673783	0.0136406

^asleeve gastrectomy vs. Roux-en-Y gastric bypass

Compared to the pooled individual-level analysis, however, our method requires multiple file transfers between the data partners and the analysis center. Although this need for information exchange at each iteration means that our proposed method is more labor-intensive to implement in practice, recent advancements in bioinformatics now allow semi-automated or fully-automated file transfers between data partners and the analysis center [17, 21–27]. For general users who may not have access to such technical infrastructure, we have developed R code that allows manual implementation of our proposed method. This R code (available in Additional file 2) illustrates the analysis of the simulated data from this study but can be easily modified to accommodate different numbers of covariates or different numbers of participating data partners.

We assumed the outcome occurrence between individuals to be independent in our analysis. In some real-

world situations, this independence assumption may be violated. In our case, it is possible that individuals who seek care in the same delivery system have correlated outcomes. To account for correlated data, Zou and Donner [28] extended the modified Poisson approach to settings with correlated binary outcomes in single-database studies. Future work will extend the proposed summary-level modified Poisson method to analyze correlated data in multi-center distributed data environments.

Conclusions

In conclusion, we proposed a privacy-protecting approach to directly estimate adjusted risk ratios using modified Poisson regression analysis for multi-center studies. This approach does not require sharing of individual-level data across data partners but produces results that are identical to those obtained from the corresponding pooled individual-level data analysis.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12874-019-0878-6>.

Additional file 1. Reports the shared summary-level information in the simulated data example.

Additional file 2. Provides replication R code for the data generation and analysis of the simulated data example.

Acknowledgements

The authors thank Qoua Her at the Harvard Pilgrim Health Care Institute for his help with the creation of the real-world dataset. The authors also thank Xiaojuan Li and Jenna Wong at the Harvard Pilgrim Health Care Institute for their comments on an earlier draft of this manuscript.

Authors' contributions

The draft was written by DS, JGY and ST. DS conceived the idea, wrote the R program and performed all data analyses. JGY and ST helped supervise the presentation of methodology, and the design and interpretations of data analyses. All authors read and approved the final manuscript.

Funding

Dr. Toh was funded in part by the Patient-Centered Outcomes Research Institute (ME-1403-11305), the National Institute of Biomedical Imaging and Bio-engineering (U01 EB023683), and a Harvard Pilgrim Health Care Institute Robert H. Ebert Career Development Award. The funding bodies had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

Replication R code for generating and analyzing the simulated data example is available online and can be easily modified by interested readers for their own multi-center analyses. The real-world data created from the IBM® MarketScan® Research Databases are currently not available for public sharing.

Ethics approval and consent to participate

This study, including the secondary data analysis using the IBM® MarketScan® Research Databases that contain de-identified individual-level healthcare claims information, was approved as a non-human subject research project by the Institutional Review Board at Harvard Pilgrim Health Care.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 15 February 2019 Accepted: 22 November 2019

Published online: 05 December 2019

References

- Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979;66(3):403–11.
- Hosmer DW, Lemeshow S, Sturdivant RX. *Applied logistic regression*. 3rd ed. Hoboken: Wiley; 2013.
- Norton EC, Dowd BE, Maciejewski ML. Odds ratios - current best practice and use. *JAMA*. 2018;320(1):84–5.
- Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol*. 1987;125(5):761–8.
- Altman DG, Deeks JJ, Sackett DL. Odds ratios should be avoided when events are common. *BMJ*. 1998;317(7168):1318.
- Holcomb WL Jr, Chaiworapongsa T, Luke DA, Burgdorf KD. An odd measure of risk: use and misuse of the odds ratio. *Obstet Gynecol*. 2001;98(4):685–8.
- Knol MJ, Le Cessie S, Algra A, Vandenbroucke JP, Groenwold RH. Overestimation of risk ratios by odds ratios in trials and cohort studies: alternatives to logistic regression. *CMAJ*. 2012;184(8):895–9.
- Tajeu GS, Sen B, Allison DB, Menachemi N. Misuse of odds ratios in obesity literature: an empirical analysis of published studies. *Obesity*. 2012;20(8):1726–31.

- Wacholder S. Binomial regression in glim: estimating risk ratios and risk differences. *Am J Epidemiol*. 1986;123(1):174–84.
- Skove T, Deddens J, Petersen MR, Endahl L. Prevalence proportion ratios: estimation and hypothesis testing. *Int J Epidemiol*. 1998;27(1):91–5.
- McNutt LA, Wu C, Xue X, Hafner JP. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am J Epidemiol*. 2003;157(10):940–3.
- Zou G. A modified Poisson regression approach to prospective studies with binary data. *Am J Epidemiol*. 2004;159(7):702–6.
- Toh S, Platt R, Steiner JF, Brown JS. Comparative-effectiveness research in distributed health data networks. *Clin Pharmacol Ther*. 2011;90(6):883–7.
- Ball R, Robb M, Anderson SA, Dal PG. The FDA's sentinel initiative - a comprehensive approach to medical product surveillance. *Clin Pharmacol Ther*. 2016;99(3):265–8.
- Fienberg SE, Fulp WJ, Slavkovic AB, Wrobel TA. "Secure" log-linear and logistic regression analysis of distributed databases. In: Domingo-Ferrer J, Franconi L, editors. *Privacy in Statistical Databases*. PSD 2006. Lecture notes in computer science, vol 4302. Berlin, Heidelberg: Springer; 2006.
- Karr AF, Fulp WJ, Vera F, Young SS, Lin X, Reiter JP. Secure, privacy-preserving analysis of distributed databases. *Technometrics*. 2007;49(3):335–45.
- Jiang W, Li P, Wang S, Wu Y, Xue M, Ohno-Machado L, et al. WebGLORE: a web service for grid LOGistic REGression. *Bioinformatics*. 2013;29(24):3238–40.
- El Emam K, Samet S, Arbuckle L, Tamblyn R, Earle C, Kantarcioglu M. A secure distributed logistic regression protocol for the detection of rare adverse drug events. *J Am Med Inform Assoc*. 2013;20(3):453–61.
- R Core Team. R: A Language and Environment for Statistical Computing. Vienna. URL <https://www.R-project.org/>: R Foundation for Statistical Computing; 2018.
- Stefanski LA, Boos DD. The calculus of M-estimation. *Am Stat*. 2002;56(1):29–38.
- Her QL, Malenfant JM, Malek S, Vilik Y, Young J, Li L, et al. A query workflow design to perform automatable distributed regression analysis in large distributed data networks. *eGEMS*. 2018;6(1):11.
- Jiang X, Wu Y, Marsolo K, Ohno-Machado L. Development of a web service for analysis in a distributed network. *eGEMS*. 2014;2(1):22.
- Wolfson M, Wallace SE, Masca N, Rowe G, Sheehan NA, Ferretti V, et al. DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. *Int J Epidemiol*. 2010;39(5):1372–82.
- Wu Y, Jiang X, Kim J, Ohno-Machado L. Grid binary LOGistic REGression (GLORE): building shared models without sharing data. *J Am Med Inform Assoc*. 2012;19(5):758–64.
- Lu CL, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, et al. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. *J Am Med Inform Assoc*. 2015;22(6):1212–9.
- Narasimhan B, Rubin DL, Gross SM, Bendersky M, Lavori PW. Software for distributed computation on medical databases: a demonstration project. *J Stat Softw*. 2017;77(13):22.
- Meeker D, Jiang X, Matheny ME, Farcas C, D'Arcy M, Pearlman L, et al. A system to build distributed multivariate models and manage disparate data sharing policies: implementation in the scalable national network for effectiveness research. *J Am Med Inform Assoc*. 2015;22(6):1187–95.
- Zou G, Donner A. Extension of the modified Poisson regression model to prospective studies with correlated binary data. *Stat Methods Med Res*. 2013;22(6):661–70.

Publisher's Note

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