Vol. 190, No. 2
DOI: 10.1093/aje/kwaa205
Advance Access publication:
September 25, 2020

Meta-Analysis and Sparse-Data Bias

David B. Richardson*, Stephen R. Cole, Rachael K. Ross, Charles Poole, Haitao Chu, and Alexander P. Keil

* Correspondence to Dr. David B. Richardson, Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 (e-mail: david.richardson@unc.edu).

Initially submitted February 28, 2020; accepted for publication September 22, 2020.

Meta-analyses are undertaken to combine information from a set of studies, often in settings where some of the individual study-specific estimates are based on relatively small study samples. Finite sample bias may occur when maximum likelihood estimates of associations are obtained by fitting logistic regression models to sparse data sets. Here we show that combining information from small studies by undertaking a meta-analytical summary of logistic regression estimates can propagate such sparse-data bias. In simulations, we illustrate 2 challenges encountered in meta-analyses of logistic regression results in settings of sparse data: 1) bias in the summary meta-analytical result and 2) confidence interval coverage that can worsen rather than improve, in terms of being less than nominal, as the number of studies in the meta-analysis increases.

cohort studies; logistic regression; meta-analysis; regression analysis

Abbreviations: CI, confidence interval; OR, odds ratio.

A meta-analysis of epidemiologic study results is undertaken to combine information from a set of studies. Often a motivation for conducting a meta-analysis is that some, or all, of the individual study-specific estimates are based on relatively few data. A meta-analytical approach involves the aggregation of information to yield a summary effect estimate that often has greater statistical precision than is possible to obtain in any individual study. Meta-analytical approaches to deriving a summary effect estimate typically involve calculation of a weighted average of the effect estimates from the individual studies (1).

However, finite sample bias may occur when epidemiologic estimates of associations are obtained by fitting logistic regression models to sparse data (2). In this paper, we show that undertaking a meta-analytical summary of logistic regression estimates derived from sparse data can propagate such bias. Consequently, a meta-analytical summarization of logistic regression estimates may not yield a consistent summary estimate that converges to the true association, and the resultant confidence interval may not have the desired nominal coverage.

Here, we describe the problem in the context of meta-analysis of published results from multivariable logistic regression models and illustrate it using simulations.

METHODS

Meta-analysis of logistic regression estimates

We focus on a setting where epidemiologic results have been obtained from multivariable logistic regression models. Subsequently, an investigator wishes to conduct a meta-analysis based on published estimates of odds ratios and associated confidence intervals derived by these logistic regression model fittings. As is typical, the investigator does not have the original, individual-level data for all of the studies included in the meta-analysis. The parameter of primary interest in the meta-analysis, β, denotes the covariate-adjusted natural log of the odds ratio per unit of exposure (noting that a unit change in exposure may refer to settings in which the exposure variable is a binary, ordinal, or continuous variable). The data structure available for this summarization of epidemiologic findings is a table of estimates of β and associated confidence intervals. Let $i = 1 \ \ldots \ k$ index the k estimates to be summarized in the meta-analysis. Let β_i denote the estimate of the natural log of the odds ratio per unit of exposure for study i; and let L_i and U_i denote the associated lower and upper confidence limits for β_i. In standard meta-analytical techniques, study-specific variance estimates are used to calculate the

inverse-variance–weighted average estimate of association, which is reported as the summary estimate of association (3). We address a summary meta-analytical estimate of the association based on a common-effect model, often referred to as a fixed-effect meta-analysis (1), and a summary meta-analytical estimate of the association based on a random-effects model (see Web Appendix 1, available at https://academic.oup.com/aje).

Logistic regression estimates obtained using maximum likelihood methods are known to be susceptible to finite (or sparse) sample bias away from the null when data are sparse (2–4). This bias tends to increase as the number of covariates included in a multivariable regression model increases, while the number of outcomes remains fixed; bias may similarly arise with matched case-control study designs as the number of strata defining matching factors increases (4, 5). Given that meta-analytical approaches to deriving a summary effect estimate involve calculation of a weighted average of the results of the individual studies, if a study in the meta-analysis suffers sparse-data bias, then this approach to combining information will propagate that bias.

Simulation example

In simulation-based assessments, we allow that a meta-analysis encompasses estimates of association derived from multiple studies of the same exposure, outcome, and set of covariates, obtained under the same logistic model form. We simulated data under scenarios in which the number of individuals per study was small (n = 200–300), moderate (n = 300–500), or large (n = 500–750), the number of studies per meta-analysis was 5, 10, or 15, and the number of covariates was 5, 10, or 15. In each simulation, the number of people in a study was drawn from a uniform distribution over the specified range of study size; for each cohort member, we generated independent standard normal covariates $\boldsymbol{Z}$. We generated a random binary exposure, E, with dependence on covariates by specifying that E took a value of 1 with probability $\frac{1}{(1+\exp(-(\log(0.25)+\boldsymbol{\phi Z})))}$, where $\boldsymbol{\phi}$ was set to 0.20, 0.10, or 0. We generated a binary outcome, Y, with dependence of Y on covariates $\boldsymbol{Z}$ and exposure E encoded by specifying that Y took a value of 1 with probability $\frac{1}{(1+\exp(-(\log(0.11)+\boldsymbol{\varphi Z}+\eta_f)))}$, where $\boldsymbol{\varphi}$ was set to 0.20, 0.10, or 0 and η_f was set to 1, 0.50, or 0.20. We estimated the multivariable adjusted log odds ratio for each of the studies included in each meta-analysis. We used maximum likelihood to fit a logistic regression model for the outcome that included the exposure, E, and covariates $\boldsymbol{Z}$ as independent variables with no product terms for interaction between covariates. For each scenario, 1,000 meta-analyses were simulated; then, we calculated a summary meta-analytical estimate of the association based on a common-effect model approach (Web Appendix 1). To summarize the results, we calculated the average estimate, the average of the squared difference between the estimate and the simulation's specified true value, η_f, and the 95% confidence interval coverage of the true value for each scenario.

We then repeated the set of simulations using a similar simulation approach, but we generated a binary outcome, Y, that took a value of 1 with probability $\frac{1}{(1+\exp(-(\log(0.11)+\boldsymbol{\varphi Z}+\eta_r)))}$, where $\eta_r \sim N(\mu, \sigma)$ with μ = 1, 0.50, or 0.20 and σ = 1, 0.5, or 0.1. For each scenario, 1,000 meta-analyses were simulated; then, we calculated a summary meta-analytical estimate of the association based on a random-effects model (Web Appendix 1).

RESULTS

Table 1 summarizes the simulation results for common-effect meta-analyses in which the exposure-outcome association, as quantified by the log odds per unit of exposure, was 1.0. The expected incidence of the outcome across the simulation scenarios was 14%–16%, and the number of outcome events was approximately 40, 60, and 95 when the number of individuals per study was small (n = 200–300), moderate (n = 300–500), and large (n = 500–750), respectively. The meta-analytical results were approximately unbiased when the number of individuals per study was large (n = 500–750) and the number of covariates per regression model was small (5). The degree of bias and the mean squared error tended to increase as the study size decreased and as the number of covariates included in the regression model increased. The 95% confidence interval coverage tended to improve (i.e., become closer to the nominal 95% value) as the study size increased and as the number of covariates included in the regression model decreased, while confidence interval coverage tended to worsen (i.e., become less than nominal) as the number of studies per meta-analysis increased and as the number of covariates in the models increased. Web Tables 1 and 2 show simulation results for common-effect meta-analyses in which the exposure-outcome association, as quantified by the log odds per unit of exposure, was 0.50 and 0.20, respectively (Web Appendix 2), with similar patterns of bias and confidence interval coverage as those observed in Table 1.

Table 2 summarizes the simulation results for a random-effects meta-analysis in which the exposure-outcome association, as quantified by the log odds per unit of exposure, was normally distributed with a mean of 1.0. When the between-study variance in the exposure effect was small (σ = 0.1), the simulation results of the random-effects meta-analysis (Table 2) were very similar to those for the common-effect meta-analysis (Table 1). When the between-study variance in the exposure effect was larger (σ = 0.5 or σ = 1.0), the degree of bias in the random-effects meta-analysis was greater than in the common-effect meta-analysis. The confidence interval coverage tended to improve (i.e., become closer to the nominal 95% value) as the study size increased and as the between-study variance in the exposure effect decreased. When the between-study variance in the exposure effect was larger (σ = 0.5 or σ = 1.0), the confidence interval coverage did not tend to worsen as the number of studies per meta-analysis increased. Web Tables 3 and 4 show simulation results for random-effects meta-analyses in which the exposure-outcome association, as quantified by the log odds per unit of exposure, was normally distributed with a mean of 0.5 and 0.2, respectively (Web Appendix 2), with similar patterns observed as those illustrated in Table 2.

Table 1. Simulated Results[a] for a Common-Effect Meta-Analysis in Which the Exposure-Outcome Association, as Quantified by the Log Odds per Unit of Exposure, Is 1.0

η_f	Simulation Setup			Common-Effect Meta-Analysis		
	No. of Covariates per Model	No. of Studies per Meta-Analysis	No. of Subjects per Study	Mean Estimate	95% CI Coverage, %	Mean Squared Error
1.0	15	15	500–750	1.04	90	0.007
			300–500	1.07	88	0.012
			200–300	1.11	84	0.025
1.0	15	10	500–750	1.04	91	0.009
			300–500	1.07	91	0.016
			200–300	1.11	89	0.031
1.0	15	5	500–750	1.04	93	0.015
			300–500	1.06	93	0.027
			200–300	1.10	91	0.053
1.0	10	15	500–750	1.03	94	0.005
			300–500	1.04	91	0.010
			200–300	1.08	89	0.019
1.0	10	10	500–750	1.03	94	0.007
			300–500	1.05	94	0.014
			200–300	1.08	91	0.025
1.0	10	5	500–750	1.03	94	0.015
			300–500	1.05	93	0.025
			200–300	1.08	93	0.044
1.0	5	15	500–750	1.02	94	0.005
			300–500	1.03	94	0.008
			200–300	1.04	93	0.015
1.0	5	10	500–750	1.02	96	0.006
			300–500	1.03	95	0.011
			200–300	1.04	95	0.019
1.0	5	5	500–750	1.02	94	0.013
			300–500	1.03	96	0.020
			200–300	1.04	95	0.034

Abbreviation: CI, confidence interval.

[a] Results for scenarios in which the covariate-exposure association, ϕ, was set at 0.20 and the covariate-outcome association, φ, was set at 0.20.

Web Tables 5 and 6 illustrate results for common-effect and random-effects meta-analyses, respectively, under simulation settings in which we varied the magnitudes of associations between the covariates and exposure and between the covariates and the outcome (Web Appendix 2). Holding study size fixed, bias and confidence interval coverage were similar across the range of settings examined for covariate-exposure and covariate-outcome associations.

DISCUSSION

Meta-analyses of regression results are often undertaken in the context of epidemiologic literature that may include small studies that individually lead to relatively imprecise estimates. A common-effect meta-analysis is an approach used to combine information across studies and derive a more precise summary estimate of association than is obtained in any individual study. However, this meta-analytical approach does not reduce the sparse-data bias that may affect maximum likelihood estimates. Sparse-data bias tends to arise when the number of outcome events in an analysis is small; and holding the number of events fixed, bias tends to increase as the number of covariates included in a regression model increases. In our simulations, common-effect meta-analytical results were biased away from the null when sample sizes were small and the number of model covariates was large; and as the number of studies included in meta-analyses increased (holding other

Table 2. Simulated Results[a] for A Random-Effects Meta-Analysis in Which the Exposure-Outcome Association, as Quantified by the Log Odds per Unit of Exposure, Is Normally Distributed With a Mean of 1.0

$\eta_r \sim N(\mu, \sigma)$		Simulation Setup		Random-Effects Meta-Analysis	
μ	σ	No. of Studies per Meta-Analysis	No. of Subjects per Study	Mean Estimate	95% CI Coverage, %
1.0	0.1	15	500–750	1.04	92
	0.5			1.05	92
	1.0			1.07	92
1.0	0.1	15	300–500	1.07	89
	0.5			1.09	89
	1.0			1.11	90
1.0	0.1	10	500–750	1.04	92
	0.5			1.05	91
	1.0			1.07	90
1.0	0.1	10	300–500	1.07	91
	0.5			1.09	90
	1.0			1.11	89
1.0	0.1	5	500–750	1.05	93
	0.5			1.05	89
	1.0			1.06	89
1.0	0.1	5	300–500	1.07	92
	0.5			1.10	87
	1.0			1.13	86

Abbreviation: CI, confidence interval.

[a] Results for scenarios with 15 covariates per regression model in which the covariate-exposure association, ϕ, was set at 0.20 and the covariate-outcome association, φ, was set at 0.20.

simulation parameters unchanged), the confidence interval coverage for the meta-analytical summary worsened. This is probably because confidence intervals for meta-analytical summary estimates become tighter as the number of studies in a meta-analysis increases while the bias remains (6).

Similarly, in our simulations, random-effects meta-analytical summaries were biased away from the null when sample sizes were small; we further noted that bias tended to be larger for random-effects meta-analyses than for common-effect meta-analyses and tended to increase as the between-study variance in the exposure effect increased. In part, this may be because smaller studies have relatively larger weights in the summary effect estimate in a random-effects meta-analysis than in a common-effect meta-analysis. The simulations illustrate simple scenarios of meta-analyses of logistic regression model parameter estimates derived under conditions of correct model specification. Sparse-data bias also could affect conclusions regarding homogeneity of estimates included in a meta-analysis—a problem not addressed in our current analyses but a potentially useful topic for future work.

Sparse-data bias affects logistic regression models (7) as well as Poisson and Cox regression models (2). The potential for sparse-data bias within studies to propagate through meta-analyses adds another limitation to the use of logistic regression methods; other issues that have been raised with odds ratios relate to problems of interpretation, collapsibility, valid estimation, and transportability relative to more substantively meaningful measures such as risk and prevalence differences and ratios (8–10). We focused on the setting in which a meta-analysis of aggregate data is undertaken (summary estimates of odds ratios and associated confidence intervals) derived from multiple studies that are conceptually identical, involving not only the same exposure and outcome variables but also the same covariates, modeled under the same logistic form. In practice, meta-analyses often diverge from these conditions, and some estimates may be derived from models with covariate-adjustment sets that may differ between studies. Given the noncollapsibility of the odds ratio (meaning that the covariate-conditional odds ratio may differ from the crude odds ratio even in the absence of confounding), meta-analysts should be cautious in these settings, as the meta-analysis will combine estimates from studies that estimated different (covariate-conditional) effects. Meta-analyses of summary estimates of risk ratios, a collapsible measure of association, do not suffer from this problem; however, susceptibility to sparse-data bias is a concern for estimates of risk ratios, as it is for odds ratios

(2, 6), and problems with model convergence are commonly encountered when estimating risk ratios in multivariable binomial regression models (although several approaches for addressing such problems have been described (11–14)).

A meta-analytical approach may be desirable for summarizing a result derived from the regression model estimates obtained from each individual study. Our findings suggest that consideration of potential sparse-data bias is warranted. A variety of methods have been proposed for reducing sparse-data bias in regression estimates for individual studies (15, 16); and approaches such as Firth's correction, which is one method for penalization as an approach to dealing with sparse-data bias (2), can make regression estimates more resistant to sparse-data bias (16). Of course, in the setting of primary interest, where access to the individual-level study data is not feasible and the aggregate data are covariate-conditional logistic regression estimates of log odds ratios, such corrections are useful only if they were employed by the investigators who reported the study-specific regression model estimates being meta-analyzed. However, the literature suggests that such bias is often unaccounted for in the published literature (2). A useful guideline, albeit one with limitations (17), for assessment of potential sparse-data bias in the individual studies included in a meta-analysis is that sparse-data bias typically is minimal in regression analyses that include at least 10 events per variable in the model (18). Consistent with that guideline, in our simulations there was little evidence of sparse-data bias propagating in meta-analyses of logistic regression results for scenarios where the expected number of events per study was approximately 10 times the number of variables in the regression model (e.g., simulation scenarios with 5 covariates and 300 or more subjects per study (Table 1)).

When undertaking meta-analyses of epidemiologic findings derived from models susceptible to sparse-data bias, we suggest the need for caution and attention to sparse-data bias and less-than-nominal confidence interval coverage in the resultant meta-analytical summary result.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States (David B. Richardson, Stephen R. Cole, Rachael K. Ross, Charles Poole, Alexander P. Keil); and Division of Biostatistics, School of Public Health, University of Minnesota Twin Cities, Minneapolis-St. Paul, Minnesota, United States (Haitao Chu).

Conflict of interest: none declared.

REFERENCES

1. Bender R, Friede T, Koch A, et al. Methods for evidence synthesis in the case of very few studies. *Res Synth Methods*. 2018;9(3):382–392.
2. Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. *BMJ*. 2016;352:i1981.
3. Sutton AJ, Abrams KR, Jones DR, et al. *Methods for Meta-Analysis in Medical Research*. Chichester, United Kingdom: John Wiley & Sons Ltd.; 2000.
4. Jewell NP. Small-sample bias of point estimators of the odds ratio from matched sets. *Biometrics*. 1984;40(2): 421–435.
5. Greenland S, Schwartzbaum JA, Finkle WD. Problems due to small samples and sparse data in conditional logistic regression analysis. *Am J Epidemiol*. 2000;151(5): 531–539.
6. Lin L. Bias caused by sampling error in meta-analysis with small sample sizes. *PLoS One*. 2018;13(9):e0204056.
7. Nemes S, Jonasson JM, Genell A, et al. Bias in odds ratios by logistic regression modelling and sample size. *BMC Med Res Methodol*. 2009;9:Article 56.
8. Kass PH, Greenland S. Conflicting definitions of confounding and their ramifications for veterinary epidemiologic research: collapsibility vs comparability. *J Am Vet Med Assoc*. 1991;199(11):1569–1573.
9. Sackett DL, Deeks JJ, Altman DG. Down with odds ratios! *BMJ Evid Based Med*. 1996;1(6):164–166.
10. McNutt LA, Wu C, Xue X, et al. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am J Epidemiol*. 2003;157(10):940–943.
11. Deddens JA, Petersen MR. Re: "Estimating the relative risk in cohort studies and clinical trials of common outcomes" [letter]. *Am J Epidemiol*. 2004;159(2):213–214.
12. Spiegelman D, Hertzmark E. Easy SAS calculations for risk or prevalence ratios and differences. *Am J Epidemiol*. 2005; 162(3):199–200.
13. Wacholder S. Binomial regression in GLIM: estimating risk ratios and risk differences. *Am J Epidemiol*. 1986;123(1): 174–184.
14. Zou G. A modified Poisson regression approach to prospective studies with binary data. *Am J Epidemiol*. 2004; 159(7):702–706.
15. Walter SD. Small sample estimation of log odds ratios from logistic regression and fourfold tables. *Stat Med*. 1985;4(4): 437–444.
16. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Stat Med*. 2002;21(16): 2409–2419.
17. van Smeden M, de Groot JA, Moons KG, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Med Res Methodol*. 2016;16(1): Article 163.
18. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49(12):1373–1379.