Title

Ogburn et al. Respond to "Estimation and Bounds Under Data Fusion".

Journal

American Journal of Epidemiology, 191(4)

ISSN

0002-9262

Authors

Ogburn, Elizabeth L

Rudolph, Kara E

Morello-Frosch, Rachel

et al.

Publication Date

2022-03-24

DOI

10.1093/aje/kwab195

Peer reviewed

American Journal of Epidemiology

Vol. 191, No. 4
https://doi.org/10.1093/aje/kwab195
Advance Access publication:
July 7, 2021

Response to Invited Commentary

Ogburn et al. Respond to "Estimation and Bounds Under Data Fusion"

Elizabeth L. Ogburn*, Kara E. Rudolph, Rachel Morello-Frosch, Amber Khan, and Joan A. Casey

* Correspondence to Dr. Elizabeth L. Ogburn, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Baltimore, MD 21205 (e-mail: eogburn@jhsph.edu).

Initially submitted June 5, 2021; accepted for publication June 30, 2021.

We thank Miao et al. (1) for an excellent, enlightening, and forward-looking commentary on our paper (2). We are grateful to them for pointing out that all need not be lost when a variable required for an analysis is not directly measured in the primary data. Methods for data fusion, which is an active area of statistical research, attempt to solve precisely this problem. Miao et al. propose 3 ingenious approaches to data fusion (1).

The first method, estimation with a nonlinear imputation model, describes a setting in which the common current practice that we critiqued in our paper (2) is in fact valid. In other words, the approach they describe is another answer to the question we posed in our paper: When is it acceptable to use estimates from auxiliary regression models? Recall that V is a variable that is required for the primary analysis but not available in the primary data, and Z is the collection of auxiliary data covariates used as independent variables in the model for V. Miao et al. demonstrated (1) that it is acceptable to use estimates from auxiliary regression models when 1) V is an exposure or covariate; 2) the true relationship between the outcome, V, and Z is linear with no V-Z interaction; 3) all covariates in the primary analysis are available in the auxiliary data and included in Z; 4) the joint distribution of V and Z is the same in the primary and auxiliary data; and 5) the true relationship between V and Z and the model $g(Z)$ used to estimate V in the auxiliary data are both nonlinear. This method hinges on the linearity of the outcome in V and Z contrasted with the nonlinearity of the relationship between V and Z, and in the applications that we reference in our paper (2), we are skeptical that this contrast could be strongly justified on scientific bases. In particular, we are skeptical that the outcome would be linear in V and Z in most epidemiologic settings of interest. (Note that it does not suffice to be interested in *estimating* a linear model in the primary analysis; the linear model must capture the *true* relationship between the outcome and V and Z.) Miao et al. express a similar skepticism in their conclusion (1).

The second method, estimation with validation data, relies on a second auxiliary data set. Typically, the auxiliary data set includes V and Z but *not* other variables W required for the primary analysis. (If it did, researchers could simply use the auxiliary data to run the primary analysis without worrying about estimating V.) If another data set is available with data W and V, then it may be possible to correct the biases that we describe in our paper. Miao et al. show that this is indeed the case when 1) the relationship between V, W, and Z is linear and 2) the mean of the product of W and V, $E[WV]$, is the same in the primary data and the second auxiliary data set (1). We remain skeptical that the assumption of linearity will hold in most epidemiologic settings, and it could also be difficult for researchers to find a second auxiliary data set meeting the requirements for this method.

Finally, the third method derives bounds for the association or effect of interest. This strikes us as an extremely fruitful and promising direction for further research. When reasonable assumptions do not suffice to identify an estimand of interest, it may still be the case that the set of assumptions implies a *feasible region* for the estimand: a set of values guaranteed to contain its true value. In this case, the feasible region is given by upper and lower bounds on the truth; we can be confident that under our assumptions, the truth lies between these 2 bounds. Miao et al. derive bounds under the assumption of linearity of the relationship between V, W, and Z, but we echo their optimism that similar bounds could be found even if the assumption of linearity were relaxed or replaced. We agree with Miao et al. that the bounding approach is very promising and deserving of more attention (1). We look forward to following their future work in this direction.

As Miao et al. readily admit, none of their 3 approaches is a panacea. These methods require more care, more work, and more data than the existing (flawed) practice of simply replacing V with predictions from the auxiliary model. They probably would require researchers to have access to the auxiliary data, or the ability to request specific analyses using the auxiliary data, which currently is frequently infeasible. They rely on the true relationship between W, V,

and Z being linear and additive. This is a major limitation; even when researchers are interested in the parameters in such a linear model, it may not be the correct structural model. However, our primary takeaway from Miao et al.'s commentary is optimism that more general methods for data fusion may be available soon, and we look forward very much to following the work of these insightful researchers.

ACKNOWLEDGMENTS

Author affiliations: Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, United States (Elizabeth L. Ogburn); Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, New York, United States (Kara E. Rudolph); Department of Environmental Science, Policy and Management, School of Public Health, University of California, Berkeley, Berkeley, California, United States (Rachel Morello-Frosch); Department of Environmental and Occupational Health Sciences, School of Public Health, University of Washington, Seattle, Washington, United States (Amber Khan); and Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, New York, United States (Joan A. Casey).

E.L.O. was supported by National Institutes of Health grant U24OD023382 and Office of Naval Research grant N00014-18-1-2760.

Conflict of interest: none declared.

REFERENCES

1. Ogburn EL, Rudolph KE, Morello-Frosch R, et al. A warning about using predicted values from regression models for epidemiologic inquiry. *Am J Epidemiol*. 2021;190(6): 1142–1147.
2. Miao W, Li W, Hu W, et al. Invited commentary: estimation and bounds under data fusion. *Am J Epidemiol*. 2022; 191(4):674–678.