Evaluating Public Health Interventions: 8. Causal Inference for Time-Invariant Interventions

We provide an overview of classical and newer methods for the control of confounding of time-invariant interventions to permit causal inference in public health evaluations.

We estimated the causal effect of gender on all-cause mortality in a large HIV care and treatment program supported by the President's Emergency Program for AIDS Relief in Dar es Salaam, Tanzania, between 2004 and 2012. We compared results from multivariable modeling, three propensity score methods, inverse-probability weighting, doubly robust methods, and targeted maximum likelihood estimation. Considerable confounding was evident, and, as expected by theory, all methods considered gave the same result, a statistically significant approximately 20% increased mortality rate in men.

In general, there is no clear advantage of any of these methods for causal inference over classical multivariable modeling, from the point of view of either bias reduction or efficiency. Rather, given sufficient data to adequately fit the multivariable model to the data, multivariable modeling will yield causal estimates with the greatest statistical efficiency. All methods can adjust only for well-measured confounders—if there are unmeasured or poorly measured confounders, none of these methods will yield causal estimates. (*Am J Public Health*. 2018;108:1187–1190. doi:10.2105/AJPH.2018.304530)

Donna Spiegelman, ScD, and Xin Zhou, PhD

In this commentary, the eighth in the series, "Evaluating Public Health Interventions," we provide an overview of classical and newer methods for confounding control to permit causal inference in public health evaluations when either measured confounders or the intervention are time-invariant. We will emphasize that, contrary to widespread belief, classical methods for confounding control for causal inference are as good as, and often better than, newer "causal inference" methods when the confounders or the intervention are time-invariant. By contrast, when the intervention and confounders are both time-varying, only the newer methods for confounding can fully control for causal inference being valid, although typically their use has little impact on the findings, as will be discussed in the next column in this series. Here, we provide an overview of these methods, illustrated by an evaluation of gender disparities in mortality in a large HIV/AIDS treatment program in Tanzania.

ILLUSTRATIVE EXAMPLE

Among adults attending one of 29 HIV/AIDS treatment and care clinics in Dar es Salaam, Tanzania, Hawkins et al. reported that men were at higher risk for earlier mortality, disease progression, and loss to follow-up, although the multivariable relative risks for these outcomes were substantially smaller than those adjusted only for time since enrollment into care.[1] In an analysis updated to 2012, gender disparities in all-cause mortality among 101 214 adult patients attending these 29 clinics were reassessed; 13 674 patients died, and 40% of the patients were men. We estimated the causal parameter of interest, the incidence rate ratio, by using classical and newer methods for confounder control: multivariable modeling, three types of propensity score (PS) methods, inverse-probability weighting (IPW), double robust estimation (DR) and targeted maximum likelihood estimation (TMLE; Table 1). Each method required selecting the confounders to be adjusted for in the health outcome model, in the PS model, or both. We considered two selection strategies: (1) the main effects of all known and suspected risk factors for the outcome and (2) stepwise selection of main effects of all known and suspected risk factors for the outcome and their significant two-way interactions ($P < .05$).

Now, we will discuss methods for causal inference in the presence of a time-invariant intervention or confounders and illustrate key points through the evaluation of gender-related disparities in mortality.

WHAT IS A CONFOUNDER?

A confounder is a variable, which, if omitted from the analysis, would result in a biased estimate of the causal parameter of interest. Generally, known and suspected risk factors for the outcome of interest are potential confounders. Sometimes forgotten by PS methods users, correlates of the intervention of interest that are not independent risk factors for the outcome are not confounders,[2] nor are variables caused by the exposure (mediators) or by the outcome (colliders). Recently, some special cases have been discussed of variables that meet the definition for a confounder as given here, for which adjustment does not necessarily lead to bias reduction in the causal intervention effect estimate.[3] It is not clear how frequently these situations must be considered, and we will not consider them any further in this column.

WHEN IS CONFOUNDING CONTROL NEEDED?

When the data are observational, confounding control is

ABOUT THE AUTHORS

Donna Spiegelman is with the departments of Epidemiology, Biostatistics, Nutrition, and Global Health, Harvard T. H. Chan School of Public Health, Boston, MA. Xin Zhou is with the departments of Epidemiology and Biostatistics, Harvard T. H. Chan School of Public Health.

Correspondence should be sent to Donna Spiegelman, Department of Epidemiology, Harvard T. H. Chan School of Public Health, 677 Huntington Ave, Kresge Building, Room 802, Boston, MA 02115 (e-mail: stdls@hsph.harvard.edu). Reprints can be ordered at http://www.ajph.org by clicking the "Reprints" link.

This article was accepted May 6, 2018.

doi: 10.2105/AJPH.2018.304530

TABLE 1—A Comparison of Classical and New Methods for Confounder Control Using Male Gender in Relation to All-Cause Mortality in Dar es Salaam, Tanzania, 2004–2012

Causal Effect Estimation Strategy	Model Selection Method	HR (95% CI)[a]
Univariable Cox model		1.76 (1.70, 1.82)
Multivariable Cox model	All known and suspected risk factors for outcome	1.24 (1.20, 1.29)
PS stratification	All known and suspected risk factors for outcome	1.23 (1.19, 1.28)
	All known and suspected risk factors for outcome, main effects only, stepwise $P < .05$	1.22 (1.18, 1.27)
PS adjustment	All known and suspected risk factors for outcome, main effects only	1.23 (1.18, 1.27)
	All known and suspected risk factors for outcome, main effects and two-way interactions, stepwise $P < .05$	1.23 (1.18, 1.28)
PS matching		
9390 deaths in 54 530 patients	All known and suspected risk factors for outcome, main effects only	1.20 (1.15, 1.24)
9277 deaths in 53 986 patients	All known and suspected risk factors for outcome, main effects and two-way interactions, stepwise $P < .05$	1.20 (1.16, 1.25)
IPW	All known and suspected risk factors for outcome, main effects only	1.28 (1.20, 1.37)
Doubly robust	All known and suspected risk factors for outcome, main effects only	1.24 (1.19, 1.29)
TMLE	All known and suspected risk factors for outcome, main effects only	1.29 (1.23, 1.34)

Note. CI = confidence interval; HR = hazard ratio; IPW = inverse probability weighting; PS = propensity score; TMLE = targeted maximum likelihood estimation. The sample size was n = 13 674 deaths in 101 214 patients. Confounders included age (< 30, 30 to < 40, 40 to < 50, ≥ 50 y), season of clinic visit (long dry, short rainy, short dry, long rainy), World Health Organization HIV disease stage (I, II, III, IV), year of enrollment (2004–2005, 2006, 2007, 2008, 2009, 2010, 2011–2012), district of health facility (Ilala, Kinondoni, Temeke), facility level (hospital, health center, dispensary), oral candidiasis (yes or no), tuberculosis treatment (yes or no), history of tuberculosis (yes or no), alanine aminotransferase levels greater than 40 IU/L (yes or no), anemia (yes or no), antiretroviral therapy use (yes or no), diarrhea (yes or no), body mass index (kg/m^2, continuous with spline terms), CD4 count (cells/μL, continuous with spline terms), and, where indicated, any two-way interactions significant at $P < .05$.

[a]All P values < .001.

almost always needed for causal effect estimation and inference to eliminate bias. Confounding control is almost never needed in large individually randomized trials, as balance between intervention arms is virtually assured. Confounding control is often needed in cluster randomized trials, including stepped-wedge designs, particularly when the number of clusters is relatively limited, regardless of the size of the within-cluster samples.[4] Because many large-scale public health interventions are cluster-randomized, and because we believe that observational data are grossly underutilized for learning about the effectiveness and cost-effectiveness of public health interventions, largely because of concerns about confounding bias, mastery of these methods is important for accelerating the production and dissemination of knowledge from new and existing resources to advance public health.

CLASSICAL METHODS PROVIDE CAUSAL ESTIMATES

Open any epidemiology textbook, and you will likely see described the three classical methods for confounder control: restriction, matching, and stratification or modeling (e.g., Hennekens,[5] chapter 12). Restriction eliminates confounding at the expense of external generalizability by including study participants only at one level of the confounder. If the study is restricted to one level of all of the risk factors for the outcome under study, the crude analysis will provide a causal inference. Restriction is infeasible in interventions studying outcomes with multiple risk factors, but it may be possible to restrict to a few of the strongest risk factors and then use other available methods to control for bias from the confounders remaining.

In another classical option, at the design stage, participants can be matched on potential confounders, followed by matched analysis using Mantel–Haenszel or McNemar methods or conditional logistic regression, jointly stratified on the matching factors.[6,7] If matching on all potential confounders is undertaken at the design stage of a case–control study and the proper analysis follows, the resulting inference will be causal.

Finally, there are the classical methods of analysis alone to control for confounding: stratification and multivariable modeling. Stratified Mantel–Haenszel methods require that continuous confounders be discretized, potentially inducing residual confounding within strata. Stratified methods have the additional disadvantage that each stratum is composed of a single unique combination of levels of all the potential confounders together. This can lead to a great loss of power, as many uninformative strata are often produced. The advantage of the approach is that it is nonparametric and thus robust to residual confounding attributable to model misspecification, in the sense that no model for confounder–outcome or confounder–exposure association is needed. In addition, it adjusts not only for confounding by the main effects of each covariate but also for all possible higher-order interactions. If all confounders are stratified upon, and there is no residual confounding within strata, the resulting analysis will again provide a causal effect estimate.

The most popular of all of these classical approaches for confounding control is multivariable modeling: linear regression for continuous outcomes; logistic, Poisson, and log-binomial regression for binary outcomes, with log-binomial preferred for interpretability[8]; and Cox regression for comparing incidence rates. Multivariable models can efficiently adjust for bias attributable to a large number of potential confounders that are either continuous or categorical in nature, but have the disadvantage, which we believe to be overstated, that residual confounding

will result to the extent that the model is misspecified. Correct model specification means that on the scale of the model (linear, logistic, log, Cox proportional hazards), all necessary terms are included. As always, be sure to assess effect modification. As always, be sure to investigate nonlinearity of continuous variables[9] or use finely grouped categorical variables. If all potential confounders are adjusted for and the multivariable model is correctly specified, the analysis will provide a causal effect estimate.

As seen in Table 1, substantial confounding was apparent, and the univariate relative risk overestimated the causal relative risk by 42%. Men were found to be at 24% statistically significant, greater risk for earlier mortality, even after extensive adjustment for confounding.

NEWER METHODS GIVE CAUSAL EFFECT ESTIMATES TOO

If the classical methods provide causal inferences under the standard, well-understood assumptions when the intervention is time-varying and the confounders fixed, and when the intervention is fixed and the confounders are time-varying, why the need for new causal methods? For time-invariant confounding, there is no need whatsoever![10] It is quite unfortunate that the classical methods for confounder control are widely misunderstood as "associational" while the newer methods—PSs, DRs, and TMLEs—are considered "causal" methods. Care must be taken with time-varying confounders, which as determinants of the outcome that are, in turn, determined by the intervention, might often be more suitably regarded as mediators and analyzed as such.

Multivariable regression methods eliminate confounding by adjusting for the confounder–outcome relationship; PS methods eliminate confounding by adjusting for the confounder–intervention relationship. Propensity score methods require the building of a model for the intervention on the measured, time-invariant potential confounders. With the possible exceptions in pharmaco- and clinical epidemiology, that model is not generally one about which investigators have a great deal of intuition and, thus, this approach has a greater potential for errors, as investigators may have little idea about the expected magnitude or direction of associations between confounders and the intervention. As with multivariable modeling, the PS model must be correctly specified, or residual bias will result. In addition, it should not be overfit, or finite sample instability will result.

Once the PS model is fit to the data, several approaches are available for obtaining causal estimates. The first is to pair-match on the PS, leading to the discarding of data without close-enough matches and the consequent loss in efficiency at little gain in validity when the chosen matching criterion is unnecessarily tight, as we suspect it often is. The second PS method is to stratify by PS score groups, potentially strengthening efficiency by preserving more of the data for the analysis at the possible cost of increased bias, attributable to residual confounding within strata. The third is to adjust for the PS directly in the multivariable outcome model, ensuring that the relationship between the PS and the outcome is correctly modeled.

Asymptotically, there is no advantage to any of these methods either from the point of view of bias reduction for causal inference or efficiency. In fact, in large samples in which the investigator is likely to model both the outcome and the intervention propensity correctly, classical methods will be uniformly more efficient[10] and, thus, preferred.

When the outcome is rare and the exposure not, and when there is a large number of potential confounders, better confounding control may be obtained through PS methods, which will provide more power to block the bias attributable to confounding on the confounder–exposure association side.[11] As always, one should not adjust for correlates of the intervention that are not risk factors of the outcome (i.e., overmatching[12[p247–249]]) as considerable efficiency can be lost; one should not adjust for mediators of the causal effect between the intervention and the outcome; and one should not adjust for consequences of the outcome (colliders).[2,3] In automated PS modeling in the big-data setting, algorithms are likely to select covariates of these types.

All three of these PS methods gave similar results, a 20% to 25% higher mortality in men. Importantly, these results were similar to the standard multivariable regression results as well. The addition of higher-order interaction terms did not materially change the estimates.[13] The power loss theoretically expected from discarding nearly half of the study's data was not evident in the PS matching methods, perhaps because the sample size was so very large to begin with.

A method closely related to the PS approach is inverse probability weighting (IPW) for causal effect estimation. The PS model is created as described previously. For each study participant, PSs are estimated and the outcome model is fit weighted to the inverse of these PSs. Only the intervention needs to be included in the multivariable outcome model. The robust variance estimator is used for confidence interval construction and testing, overestimating the true variance, generally believed to be by little.[14] As with the other PS methods discussed, this method is not maximum-likelihood and, assuming both the PS and the multivariable model can be validly fit, it will be inefficient relative to direct multivariable modeling. Another challenging practical issue that arises with IPW methods is weight instability. As the probability of the intervention as a function of the PS model approaches 0 or 1, the weights become increasingly unstable and, in finite samples, biased results can be obtained.

The PS and IPW methods allow for an explicit examination of deviations from overlap between intervention and control participants with similar confounder values, analogous to detecting noninformative strata in classic stratified 2 × 2 tables. By examining differences in the distribution of PS between the intervention and control groups, extrapolation to regions of the data where PS scores residing at the tails of the distribution have no match can be avoided, more recently denoted as a violation of the positivity assumption. In the Appendix (available as a supplement to the online version of this article at http://www.ajph.org), it is shown that, with the exception of providing an explicit means for evaluating positivity, in sufficiently large studies,

IPW offers no benefit for the validity of causal inference, and a possible disadvantageous loss of efficiency.

As expected, the results from the IPW analysis of our study were similar to the others, showing a significant 28% increased mortality rate in men in the program.

CAUSAL METHODS FOR CONFOUNDER CONTROL

Doubly robust methods aim to alleviate bias that occurs from model misspecification by constructing estimators that have a causal interpretation when either the outcome model or the PS model is correctly specified.[15,16] Yet another approach for confounder control in causal inference is TMLE,[17] which is also DR and uses the superlearner,[18] a machine-learning approach that combines an optimized mixture of multiple machine-learning algorithms to select confounders for the models from a rich set of main effects, higher-order interactions, and nonlinearities with cross-validation to avoid both overfitting and bias in the causal effect estimate attributable to model misspecification. Neither DR nor TMLE methods have yet been developed with user-friendly software to permit causal estimation of rate ratios from Cox models, but pooled logistic regression[19] can reasonably be substituted in, as shown in our illustrative example.

Again, the results from the DR and TMLE analysis produced similar point and interval estimates of effect, significant 24% and 28% increased mortality in men, respectively. The utility of TMLE was particularly questionable, taking 17 days on our high-speed computer facility to produce these same results.

The Appendix, available as a supplement to the online version of this article at http://www.ajph.org, provides annotated SAS version 9.4 (SAS Institute, Cary NC) and R version 3.4.0 (R Foundation, Vienna, Austria) code used for Table 1.

CONCLUSIONS

All methods considered in this column can only adjust for bias in causal effect estimates attributable to well-measured confounders. Only randomization and quasi-experimental methods, when their own empirically unverifiable assumptions are satisfied, can adjust for unmeasured confounding, while measurement error and misclassification methods can be used to adjust causal effect estimates for residual confounding attributable to imperfectly measured confounders.[20]

As we previously discussed,[13] we saw again that the addition of higher-order interaction terms did not materially change the estimates, suggesting that newer methods, some of which make fewer assumptions about the confounding structure, did not provide any better adjustment for confounding than did the "main effects–only" model. It is our view that model misspecification bias will typically be negligible in large-scale public health evaluations. The illustrative study of gender-related disparities in mortality among HIV/AIDS patients exemplified this point: there was no difference in the results between any of these methods; the classical multivariable approach was completely adequate. When the intervention or the confounders are time-invariant, classical methods are as valid as and more efficient than newer "causal" methods for observational data. AJPH

CONTRIBUTORS

D. Spiegelman conceptualized and wrote the article. X. Zhou analyzed the data and wrote the supplementary material.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grant DP1ES025459.

The authors thank Jamie Robins for his advice about this column and acknowledge that where his disagreements with this column exist, they are philosophical and not technical.

HUMAN PARTICIPANT PROTECTION

Patients were recruited for participation and enrolled in Management and Development for Health supported care and treatment clinics following written informed consent, which was subject to ethical reviews by the Muhimbili University of Health and Allied Sciences and the Harvard School of Public Health institutional review board.

REFERENCES

1. Hawkins C, Chalamilla G, Okuma J, et al. Sex differences in antiretroviral treatment outcomes among HIV-infected adults in an urban Tanzanian setting. *AIDS*. 2011;25(9):1189–1197.

2. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163(12):1149–1156.

3. Ding P, VanderWeele TJ, Robins JM. Instrumental variables as bias amplifiers with general outcome and confounding. *Biometrika*. 2017;104(2):291–302.

4. Crippa A, Khudyakov P, Wang M, et al. A new measure of between-studies heterogeneity in meta-analysis. *Stat Med*. 2016;35(21):3661–3675.

5. Hennekens CH. *Epidemiology in Medicine*. Boston, MA; Toronto, ON: Little, Brown and Company; 1987.

6. Breslow N, Day N. *Statistical Methods in Cancer Research: The Design and Analysis of Cohort Studies*. Vol 2. Lyon, France: International Agency for Research on Cancer; 1987.

7. Breslow N, Day N. *Statistical Methods in Cancer Research: The Analysis of Case–Control Studies*. Vol 1. Lyon, France: International Agency for Research Cancer; 1980.

8. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol*. 1987;125(5):761–768.

9. Govindarajulu US, Malloy EJ, Ganguli B, et al. The comparison of alternative smoothing methods for fitting non-linear exposure-response relationships with Cox models in a simulation study. *Int J Biostat*. 2009;5(1):2.

10. Shah BR, Laupacis A, Hux JE, et al. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol*. 2005;58(6):550–559.

11. Cepeda MS, Boston R, Farrar JT, et al. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol*. 2003;158(3):280–287.

12. Rothman K, Greenland S. *Modern Epidemiology*. Boston, MA: Lippincott, Williams and Wilkins; 1998.

13. Spiegelman D, VanderWeele TJ. Evaluating public health interventions: 6. Modeling ratios or differences? Let the data tell us. *Am J Public Health*. 2017;107(7):1087–1091.

14. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*. 2004;23(19):2937–2960.

15. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005;61(4):962–973.

16. Funk MJ, Westreich D, Wiesen C, et al. Doubly robust estimation of causal effects. *Am J Epidemiol*. 2011;173(7):761–767.

17. Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. *Am J Epidemiol*. 2017;185(1):65–73.

18. Rose S. Mortality risk score prediction in an elderly population using machine learning. *Am J Epidemiol*. 2013;177(5):443–452.

19. D'Agostino RB, Lee M-L, Belanger AJ, et al. Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. *Stat Med*. 1990;9(12):1501–1515.

20. Spiegelman D. Evaluating public health interventions: 4. The Nurses' Health Study and methods for eliminating bias attributable to measurement error and misclassification. *Am J Public Health*. 2016;106(9):1563–1566.