American Journal of Epidemiology

Vol. 185, No. 10
DOI: 10.1093/aje/kwx012
Advance Access publication:
April 18, 2017

Invited Commentary

Invited Commentary: Bias Attenuation and Identification of Causal Effects With Multiple Negative Controls

Wang Miao and Eric Tchetgen Tchetgen*

* Correspondence to Eric J. Tchetgen Tchetgen, Department of Biostatistics, T.H. Chan School of Public Health, Harvard University, 677 Huntington Ave, Boston, Massachusetts 02115 (e-mail: etchetge@hsph.harvard.edu).

Initially submitted October 14, 2016; accepted for publication December 7, 2016.

In this commentary, we describe several extensions to the interesting and important negative control exposure approach for partial confounding adjustment in time-series analysis proposed by Flanders et al. (*Am J Epidemiol*. 2017;185(10):941–949). Specifically, by leveraging the availability of exposure time series, we show that under certain additional fairly reasonable assumptions, one can incorporate both past and future exposures as multiple negative control exposures to further attenuate confounding bias. We further describe 2 specific settings in which multiple controls can be used to fully account for confounding bias; the first assumes a forward-in-time version of the familiar autoregressive model for the exposure time series, while the second combines a negative control exposure with a negative control outcome for joint indirect adjustment of confounding. We briefly illustrate how one might apply our proposed framework in time-series studies. Both the original method of Flanders et al. and our proposed extensions are particularly well-suited for time-series data such as the air pollution study considered in their paper, and as such should be considered in routine environmental health studies.

air pollution study; bias attenuation; identification; negative control; time-series study

RELATED WORK IN THE MEASUREMENT ERROR LITERATURE

We congratulate Flanders et al. (1) on an interesting paper in which they use a negative control exposure to reduce confounding bias in a time-series study about air pollution effects, and we are grateful to the editor for inviting us to discuss this work. As we further discuss below, in addition to addressing the important challenge of unobserved confounding in time-series studies of environmental epidemiology, their work contributes to the literature on measurement error.

Flanders et al. studied bias attenuation of the causal effect of X_t on Y_t under the model in Figure 1, where, adopting their notation, X_{t+1} is a negative control exposure that does not causally affect the outcome Y_t. To guarantee bias attenuation, the path from U_t to X_{t+1} has to be present; otherwise, Y_t is independent of X_{t+1} conditional on X_t, and the extended model will result in identical bias as that of the final model of Flanders et al. (1). However, the extended model can achieve bias attenuation even if the arrow between X_t and X_{t+1} is absent, as long as the other conditions required by Flanders et al. (1) are met. In this case, X_{t+1} serves as a nondifferential proxy of the unmeasured confounder U_t, which has previously been studied by Greenland and Lash (2, 3) and Ogburn and VanderWeele (4, 5). For a binary confounder, Greenland (2) suggested that adjustment by a nondifferential proxy generally reduces confounding bias; for a polytomous confounder, Ogburn and VanderWeele (4, 5) showed that certain monotonicity assumptions are indispensable to guarantee such bias attenuation. Flanders et al., in fact, describe practical conditions for bias attenuation with a continuous confounder, which contributes to the literature on nondifferential measurement error. Inclusion of X_{t+1} in the extended model is consistent with adjustment by the nondifferential proxy as suggested by Greenland (2) and Ogburn and VanderWeele (4). However, in Figure 1, the nondifferential measurement error assumption does not hold for X_{t+1}, in which case, Flanders et al. describe conditions for bias attenuation and thus extend the previous methods of partial confounding adjustment to the case of differential measurement error.

FURTHER REDUCTION OF BIAS

Both in their theoretical analysis and application, Flanders et al. did not include future covariates $\mathbf{C}_{t+s}$ or past exposures

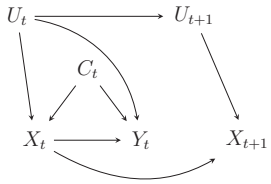

Figure 1. Directed acyclic graph for a negative control exposure X_{t+1}.

X_{t-s}, $s \geq 1$ in the extended model. However, inclusion of $\mathbf{C_{t+1}}$ or X_{t-1} may further reduce confounding bias. From Figure 2, X_{t+1} is a collider on the path $U_t \rightarrow U_{t+1} \rightarrow X_{t+1} \leftarrow \mathbf{C_{t+1}}$. Conditioning on X_{t+1} induces an association between $\mathbf{C_{t+1}}$ and U_t, and thus with X_t and Y_t. As a result, $\mathbf{C_{t+1}}$ performs as an incidental confounder, and also needs to be adjusted for X_{t+1} is included in the extended model. Assuming

$$Y_t(x_t, u_t, c_t) = \beta_0 + \beta_1 x_t + \beta_2 c_t + \beta_3 u_t + \varepsilon_t,$$

we consider the following model

$$E(U_t \mid x_t, x_{t+1}, c_t, c_{t+1}) = \alpha_0 + \alpha_1 x_t + \alpha_2 c_t + \alpha_3 x_{t+1} + \alpha_4 c_{t+1}.$$

Therefore, we have

$$E(Y_t \mid x_t, x_{t+1}, c_t) = (\beta_0 + \alpha_0\beta_3) + (\beta_1 + \alpha_1\beta_3)x_t + (\beta_2 + \alpha_2\beta_3)c_t + \alpha_3\beta_3 x_{t+1} + \alpha_4\beta_3 E(c_{t+1} \mid x_t, x_{t+1}, c_t) \quad (1)$$

and

$$E(Y_t \mid x_t, x_{t+1}, c_t, c_{t+1}) = (\beta_0 + \alpha_0\beta_3) + (\beta_1 + \alpha_1\beta_3)x_t + (\beta_2 + \alpha_2\beta_3)c_t + \alpha_3\beta_3 x_{t+1} + \alpha_4\beta_3 c_{t+1}, \quad (2)$$

which can potentially reduce confounding bias compared with equation 1. Suppose we have obtained $\widehat{E}(c_{t+1} \mid x_t, x_{t+1}, c_t) = \delta_0 + \delta_1 x_t + \delta_2 c_t + \delta_3 x_{t+1}$ using least squares, then equation 1 results in an estimator for β_1 with bias $\alpha_1\beta_3 + \alpha_4\beta_3\delta_1$ but equation 2 with bias $\alpha_1\beta_3$. If $\alpha_1\beta_3$ and $\alpha_4\beta_3\delta_1$ have the same sign, then equation 2 leads to smaller bias than equation 1. Assuming that α_1 and α_3 have the same sign, signs of $\alpha_1\beta_3$ and $\alpha_4\beta_3$ can be identified from equation 2, so we recommend performing least squares both for equations 1 and 2 to assess whether inclusion of $\mathbf{C_{t+1}}$ can further reduce bias. Note that, conditional on $\mathbf{C_{t+1}}$, future covariates $\mathbf{C_{t+s}}$, $s \geq 2$ are independent of (X_t, Y_t) and thus need not be adjusted for.

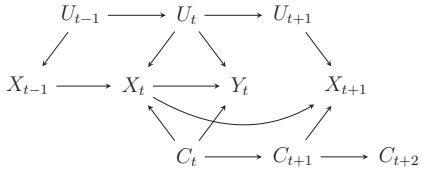

Figure 2. Directed acyclic graph for a time-series study.

Even if past exposure X_{t-1} is not a direct cause of Y_t, it is likely correlated with X_t in air pollution settings, and it is also likely correlated with Y_t through the back-door path $X_{t-1} \leftarrow U_{t-1} \rightarrow U_t \rightarrow Y_t$. Analogous to $\mathbf{C_{t+1}}$, inclusion of X_{t-1} may also reduce confounding bias. For notational convenience, we suppress covariates ($\mathbf{C_t}$, $\mathbf{C_{t+1}}$) in the rest of the commentary. Assuming

$$Y_t(x_t, u_t) = \beta_0 + \beta_1 x_t + \beta_2 u_t + \varepsilon_t, \quad (3)$$

we consider

$$E(U_t \mid x_t, x_{t+1}, x_{t-1}) = \alpha_0 + \alpha_1 x_t + \alpha_2 x_{t+1} + \alpha_3 x_{t-1};$$

then from equation 3, we have

$$E(Y_t \mid x_t, x_{t+1}) = (\beta_0 + \alpha_0\beta_2) + (\beta_1 + \alpha_1\beta_2)x_t + \alpha_2\beta_2 x_{t+1} + \alpha_3\beta_2 E(x_{t-1} \mid x_t, x_{t+1}) \quad (4)$$

and

$$E(Y_t \mid x_t, x_{t+1}, x_{t-1}) = (\beta_0 + \alpha_0\beta_2) + (\beta_1 + \alpha_1\beta_2)x_t + \alpha_2\beta_2 x_{t+1} + \alpha_3\beta_2 x_{t-1}, \quad (5)$$

which can potentially reduce confounding bias compared to equation 4. Suppose we have obtained $\widehat{E}(x_{t-1} \mid x_t, x_{t+1}) = \eta_0 + \eta_1 x_t + \eta_2 x_{t+1}$ using least squares; then equation 4 results in an estimator for β_1 with bias $\alpha_1\beta_2 + \alpha_3\beta_2\eta_1$ but equation 5 with bias $\alpha_1\beta_2$. If $\alpha_1\beta_2$ and $\alpha_3\beta_2\eta_1$ have the same sign, then equation 5 leads to smaller bias than equation 4. Assuming that α_1 and α_2 have the same sign, signs of $\alpha_1\beta_2$ and $\alpha_3\beta_2$ can be identified from equation 5, we recommend performing least squares both for equations 4 and 5 to check whether inclusion of X_{t-1} can further reduce bias. Conditional on X_{t-1}, past exposures X_{t-s}, $s \geq 2$ may still be correlated with Y_t due to residual confounding. However, one may not necessarily include them in the extended model, primarily due to efficiency considerations.

IDENTIFICATION WITH FUTURE EXPOSURES

Although the method of Flanders et al. (1) can reduce confounding bias under certain conditions, it cannot eliminate it (i.e., the causal effect of X_t on Y_t corresponding to β_1 in equation 3 cannot be identified by their method). We observe that β_1 in equation 3 can be identified by including future exposures X_{t+s}, $s \geq 2$. Assume equation 3 and

$$E(U_t \mid x_t, \ldots, x_{t+s}) = \alpha_0 + \sum_{i=0}^{s} \rho^i \alpha_1 x_{t+i}, \quad 0 < |\rho| < 1, \quad s \geq 2,$$

which is analogous to the familiar autoregressive model popular in time-series analysis, except that it describes the forward relation of current confounder with future exposures. This forward-in-time regression formalizes the idea that upon conditioning on exposures closer in time to the unmeasured confounder, the association of the latter with

future exposures decreases proportionally over time. Then from equation 3, we have

$$E(Y_t \mid x_t, \ldots, x_{t+s}) = (\beta_0 + \alpha_0\beta_2) + (\beta_1 + \alpha_1\beta_2)x_t + \sum_{i=1}^{s} \rho^i\alpha_1\beta_2 x_{t+i}, \quad (6)$$

which suffices to identify $\beta_1 + \alpha_1\beta_2$ and $\rho^i\alpha_1\beta_2$ for $i = 1, \ldots, s$, and thus ρ. As a result, we can identify $\alpha_1\beta_2$ and therefore β_1, which is the causal effect of interest. The approach of Flanders et al. (1) can still achieve bias attenuation under equation 6 when $\rho > 0$. But when $\rho < 0$, such bias attenuation is not guaranteed. Although we have focused on a linear model for the outcome, much of our results easily extends to a setting where the mean of Y_t is modeled with a log link, provided that the U_t follows a location-shift model. Such a model holds, for instance, if the error distribution of U_t is normal with homoscedastic variance but could hold more generally even if not normal.

IDENTIFICATION WITH BOTH NEGATIVE CONTROL EXPOSURE AND OUTCOME

The negative outcome control approach of Tchetgen Tchetgen (6) and Sofer et al. (7) and the multiple negative exposure approach of the previous section involve either negative control outcomes or negative control exposures, respectively, but not both and can eliminate confounding bias only under fairly stringent assumptions.

Nevertheless, we have recently developed a nonparametric method that employs both a negative control outcome and exposure to identify the causal effect without an additional parametric assumption. In the manuscript "Identifying causal effects with proxy variables of an unmeasured confounder" (8), we consider the model in Figure 3, where Z and W denote the negative control exposure and outcome, respectively. Suppose for the moment that (Z, X, U, Y, W) are categorical variables, each with k categories, and let $Y(x)$ denote the counterfactual outcome under exposure x; we have previously obtained the following identification formula for the potential outcome distribution

$$\text{pr}\{Y(x) = y\} = P(y \mid Z, x)P(W \mid Z, x)^{-1}P(W), \quad \text{for all } x, y \quad (7)$$

with matrices consisting of corresponding probabilities:

$$P(y \mid Z, x) = \{\text{pr}(y \mid z_1, x), \ldots, \text{pr}(y \mid z_k, x)\},$$

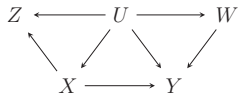

Figure 3. Directed acyclic graph with both a negative control exposure (Z) and outcome (W). Note: Independencies are invariant when the arrow $X \to Z$ is inverted.

$$P(W) = \begin{Bmatrix} \text{pr}(w_1) \\ \vdots \\ \text{pr}(w_k) \end{Bmatrix}, \; P(W \mid Z, x) = \begin{Bmatrix} \text{pr}(w_1 \mid z_1, x) & \cdots & \text{pr}(w_1 \mid z_k, x) \\ \vdots & \ddots & \vdots \\ \text{pr}(w_k \mid z_1, x) & \cdots & \text{pr}(w_k \mid z_k, x) \end{Bmatrix}.$$

Thus, using counterfactual notation, both the distributional causal effect $\text{pr}\{Y(x)\} - \text{pr}\{Y(x')\}$ and the average causal effect $E\{Y(x) - Y(x')\}$ can be identified according to formula 7, provided that $P(W|Z, x)$ is invertible. We have also generalized formula 7 to the continuous case, whereby the general identification strategy for nonparametric models involves solving an integral equation instead of the inverse of the matrix $P(W \mid Z, x)$. In the important special case of normal models for $\text{pr}(y \mid z, x)$ and $\text{pr}(w \mid z, x)$, our identification strategy entails the following: We first apply linear regression on the observed variables to obtain $E(y \mid z, x) = \gamma_0 + \gamma_1 z + \gamma_2 x$, and $E(w \mid z, x) = \gamma'_0 + \gamma'_1 z + \gamma'_2 x$; then one can verify that the path coefficient $\partial E(y|u, x) \div \partial x = \gamma_1\gamma'_2 \div \gamma'_1 - \gamma_2$. This result is consistent with previous work by Kuroki and Pearl (9), obtained via an analysis of variance under a joint normal model for (X, Y, U, Z, W). In contrast, our approach only requires normality of $\text{pr}(y \mid z, x)$ and $\text{pr}(w \mid z, x)$, not necessarily for the joint distribution $\text{pr}(x, y, u, z, w)$.

Time-series studies, such as the air pollution study considered by Flanders et al. (1), are particularly well-suited for joint negative exposure-outcome control of unmeasured confounding. Because, as one is interested in the causal effect of X_t on Y_t, future exposure X_{t+1} that does not causally affect Y_t, and past outcome Y_{t-1} that is not causally affected by X_{t+1} or X_t, are, respectively, valid candidates for negative control exposure and outcome, as they are expected to be associated with the unmeasured confounder and therefore to satisfy the structure of the graph model in Figure 3. We can apply our identification strategy with (X, Y, U, Z, W) replaced by $(X_t, Y_t, U_t, X_{t+1}, Y_{t-1})$ in Figure 3, to identify the causal effect of X_t on Y_t.

CLOSING REMARKS

Literature on negative control methods to address concerns about unmeasured confounding has been growing fast, and objectives have seemed to become more ambitious in recent years, evolving from methods for confounding bias detection (under assumptions formalized by Lipsitch et al. (10)) into methods for bias correction using one or more negative control variates. In their paper, Flanders et al. exploit the inherent time-series structure of air pollution studies to identify compelling negative control exposures, which in principle can be used to obtain a nonparametric test of confounding bias. Under more involved (untestable) parametric assumptions, they demonstrate that bias attenuation is possible. In this discussion, we have described a number of possible extensions to the work of Flanders et al., all of which rely on imposing one or more untestable assumptions to obtain further bias attenuation or even sometimes to recover identification. Another approach, which may be less ambitious and deserves more attention, is to explore nonparametric bounds and informed sensitivity analyses using negative control variates, possibly in conjunction with other design-based analyses such as instrumental variable techniques. We

expect to see more developments in these directions in the not-so-distant future.

ACKNOWLEDGMENTS

Author affiliations: Beijing International Center for Mathematical Research, Peking University, Beijing, People's Republic of China (Wang Miao); and Department of Biostatistics, T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts (Eric J. Tchetgen Tchetgen).

This work was partially supported by National Institutes of Health (grants R01AI104459 and U54 GM08858).

Conflict of interest: none declared.

REFERENCES

1. Flanders WD, Strickland MJ, Klein M. A new method for partial correction of residual confounding in time-series and other observational studies. *Am J Epidemiol*. 2017;185(10):941–949.
2. Greenland S. The effect of misclassification in the presence of covariates. *Am J Epidemiol*. 1980;112(4):564–569.
3. Greenland S, Lash TL. Bias analysis. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams and Wilkins; 2008: 345–380.
4. Ogburn EL, VanderWeele TJ. On the nondifferential misclassification of a binary confounder. *Epidemiology*. 2012; 23(3):433–439.
5. Ogburn EL, Vanderweele TJ. Bias attenuation results for nondifferentially mismeasured ordinal and coarsened confounders. *Biometrika*. 2013;100(1):241–248.
6. Tchetgen Tchetgen E. The control outcome calibration approach for causal inference with unobserved confounding. *Am J Epidemiol*. 2014;179(5):633–640.
7. Sofer T, Richardson DB, Colicino E, et al. On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Stat Sci*. 2016; 31(3):348–361.
8. Miao W, Geng Z, Tchetgen Tchetgen E. Identifying causal effects with proxy variables of an unmeasured confounder. *arXiv preprint*, arXiv:1609.08816, 2016. https://arxiv.org/abs/1609.08816.
9. Kuroki M, Pearl J. Measurement bias and effect restoration in causal inference. *Biometrika*. 2014;101(2):423–437.
10. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*. 2010;21(3): 383–388.