American Journal of Epidemiology

Vol. 185, No. 10
DOI: 10.1093/aje/kwx013
Advance Access publication:
April 18, 2017

Practice of Epidemiology

A New Method for Partial Correction of Residual Confounding in Time-Series and Other Observational Studies

W. Dana Flanders*, Matthew J. Strickland, and Mitchel Klein

* Correspondence to Dr. W. Dana Flanders, Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Road, Atlanta, GA 30322 (e-mail: wflande@emory.edu).

Initially submitted October 23, 2015; accepted for publication June 24, 2016.

Methods exist to detect residual confounding in epidemiologic studies. One requires a negative control exposure with 2 key properties: 1) conditional independence of the negative control and the outcome (given modeled variables) absent confounding and other model misspecification, and 2) associations of the negative control with uncontrolled confounders and the outcome. We present a new method to partially correct for residual confounding: When confounding is present and our assumptions hold, we argue that estimators from models that include a negative control exposure with these 2 properties tend to be less biased than those from models without it. Using regression theory, we provide theoretical arguments that support our claims. In simulations, we empirically evaluated the approach using a time-series study of ozone effects on asthma emergency department visits. In simulations, effect estimators from models that included the negative control exposure (ozone concentrations 1 day after the emergency department visit) had slightly or modestly less residual confounding than those from models without it. Theory and simulations show that including the negative control can reduce residual confounding, if our assumptions hold. Our method differs from available methods because it uses a regression approach involving an exposure-based indicator rather than a negative control outcome to partially correct for confounding.

bias; confounding; environmental epidemiologic methods; model misspecification; negative control exposure; time-series

Abbreviations: DAG, directed acyclic graph; EDV, emergency department visit.

Editor's note: *An invited response appears on page 950.*

Confounding is an important threat to the validity of observational studies. It is a mixing of the effects of an extraneous factor with those of the exposure of interest so as to distort observed associations (1, 2). It is expected if certain causal patterns (confounding paths) are present in a causal graph that reflects the causal relationships (2–4). Confounding is controlled analytically by adequate stratification or modeling covariate effects so as to block the confounding path. Residual confounding is confounding that remains even after attempts to control it. It can be present if confounders are unmeasured or mismeasured or their form is misspecified (2, 5, 6). Thus, sensitivity analyses to assess the potential impact of residual confounding can be important (2, 7, 8). In time-series studies of the short-term health effects of environmental exposures, confounding is often controlled, as least partly, by including in the model covariates such as day of week, temperature, and humidity and by including splines or other terms for time that can control for unmeasured factors that covary smoothly over time.

By considering temporality and causal relationships, we previously showed (9–11) how to test for residual confounding or other model misspecification. The test involves adding to the final model a residual-confounding indicator variable with 2 properties: P1) The indicator should be independent of disease in a correctly specified model; in particular, it should neither cause (after blocking effects of controlled factors) nor be caused by the disease. P2) It should be associated with the exposure of interest and, like the exposure, with unmeasured confounders. Plausibility of these properties can be evaluated

by considering causal diagrams that summarize tenable, causal relationships.

The test for confounding is conducted by adding the residual-confounding indicator to the model to be evaluated. We refer to the final model but with the indicator added as the "extended" model. An association between the indicator and disease in the extended model suggests model misspecification. Ability of the indicator to detect confounding was justified using causal considerations, much like those for negative control exposures (12) and Granger causality (13). In environmental time-series studies, we argued that the exposure level after the event has already occurred is a candidate for such an indicator because of the requirement that a cause precede the disease, and because health events do not, at least in the short term, affect ambient levels of many environmental exposures, such as air pollution. Thus, future levels of the exposure cannot cause disease and should not be affected by it, partially justifying one of the key properties needed for the indicator (9, 11).

Indicators with properties P1–P2 will be called "negative control exposures" (12). To detect (classical) confounding, negative control exposures and the outcome should have the same common causes as the exposure and the outcome (12) ("U-comparability"; Appendix 1). The indicators used here are expected to be U-comparable, but more generally indicators with properties P1–P2 (9–11) can also detect additional biases and need not be U-comparable (examples in Appendix 1). Lipsitch et al. (12) described a method to correct for residual confounding in sensitivity analyses when stratification, though not a regression model, is used to control for known confounders. Tchetgen Tchetgen (14) also suggested a method to correct for confounding using negative control outcomes, although it potentially involves assumption of a deterministic relationship between the dependent variable and its causes. Nevertheless, the problem of how to correct for residual confounding in regression analyses by using a negative control exposure remains open.

Our goal is to present and justify a regression-based method for reducing residual confounding in observational studies based on negative control exposures. Our main result has been that the effect estimator from the "extended" model (with the negative control exposure) tends to be less biased than that from the "final" model (without the negative control). That is, adding a negative control exposure to the model is expected to reduce residual confounding, given our assumptions. An important assumption is that measurement error is relatively unimportant. If this assumption is uncertain, the method can be used as the basis for sensitivity analyses to partially correct for bias. Our approach differs from others because it uses regression analyses with a negative control exposure rather than a negative control outcome or stratification.

METHODS

We state our assumptions and then show that effect estimators should be less biased when the negative control exposure is included in the model than when it is not included.

Assumptions

We used directed acyclic graphs (DAGs) to summarize assumed causal relationships. Many syntheses of the terminology, construction, and use of DAGs are available, so we do not repeat them here (4, 15–18). As in previous work (9), we illustrate using a time-series study with emergency department visits (EDVs) on day t as the health event (Y_t) and ambient air pollution on day t or earlier as the exposure of interest (X_t), although results apply more generally. We let C_t represent controlled confounders (measured on day t or before). The concern is that an unmeasured, perhaps unrecognized, confounder or covariate is associated with X_t and is a cause of Y_t (e.g., U_t in Figure 1). We assume that any such covariate is associated with the exposure both on day t (X_t) and the day after the health event (X_{t+1}). Throughout, we assume that the measurement error and misspecification of functional form are negligible, so any model "misspecification" stems primarily from omitting 1 or more confounders (U_t, possibly vector-valued).

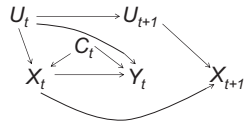


Figure 1. Directed acyclic graph summarizing assumed causal relationships; in particular, X_{t+1} is independent of Y_t, conditional on X_t, C_t, and U_t. Like X_t, X_{t+1} is associated with U_t (properties P1–P2).

We based derivation of our results on 2 key assumptions (A1 and A2) about the pattern of causal effects, consistent with those summarized in Figure 1:

Assumption 1 (A1): $Y_t(x_t, c_t, u_t) = \beta_0 + \beta_1 x_t + \beta_2 c_t + \beta_3 u_t + \varepsilon_t$, where $E[\varepsilon_t \mid X_t = x_t, C_t = c_t, U_t = u_t] = 0$, where β_0–β_3 are parameters.

Assumption 2 (A2): The DAG in Figure 1 is correct; in particular, X_{t+1} is independent of Y_t, conditional on X_t, C_t and U_t, and like X_t, X_{t+1} is associated with U_t (properties P1–P2).

The expression in A1 is interpretable as a structural equation, where $Y_t(x_t, c_t, u_t,)$ is the counterfactual value of Y_t, if X_t were set to x_t, C_t to c_t, and U_t to u_t. If U_t were measured, we could fit the regression model $Y_t = b_0 + b_1 x_t + b_2 c_t + b_3 u_t + \varepsilon_t$, to consistently estimate β_1, the effect of X_t on Y_t, the analytical objective.

Our other assumptions are:

Assumption 3 (A3): Model misspecification is due primarily to omission of the confounder U_t; other types of misspecification are negligible.

Assumption 4 (A4): The joint distribution of (ε_t, X_t, C_t, X_{t+1}, U_t) is a stationary, ergotic process, and regularity conditions (19–21) hold.

Assumption 5 (A5): $E[U_t \mid X_t = x_t, C_t = c_t, X_{t+1} = x_{t+1}] = \alpha_0 + \alpha_1 x_t + \alpha_2 c_t + \alpha_3 x_{t+1}$.

Assumption 6 (A6): $E[X_{t+1} \mid X_t = x_t, C_t = c_t] = \gamma_0 + \gamma_1 x_t + \gamma_2 c$;

where α_1–α_3 and γ_1, γ_2 are parameters.

We now derive expressions for bias when the confounder U_t is omitted from the regression model.

Mean of Y_t conditional on X_t, C_t, and X_{t+1} or on X_t and C_t

To derive formulas for the magnitude of residual confounding due to omission of one or more covariates (U_t), we first consider the mean of Y_t conditional on X_t, C_t, and X_{t+1} or on X_t and C_t.

To find an expression for $\mathrm{E}[Y_t \mid X_t = x_t, C_t = c_t, X_{t+1} = x_{t+1}]$, we first note that, in general, $\mathrm{E}[Y_t \mid X_t = x_t, C_t = c_t, X_{t+1} = x_{t+1}] = E_{U_t \mid X_t, C_t, X_{t+1}}[\mathrm{E}[Y_t \mid X_t = x_t, C_t = c_t, U_t = u_t, X_{t+1} = x_{t+1}]]$. Furthermore, $\mathrm{E}[Y_t \mid X_t = x_t, C_t = c_t, U_t = u_t, X_{t+1} = x_{t+1}]$ does not depend on X_{t+1}, given X_t, C_t, and U_t by A1–A2 so we can simplify by dropping X_{t+1} and write:

$$\begin{aligned}
&\mathrm{E}\left[Y_t \mid X_t = x_t, C_t = c_t, X_{t+1} = x_{t+1}\right] \\
&= E_{U_t \mid X_t, C_t, X_{t+1}}[\mathrm{E}[Y_t \mid X_t = x_t, C_t = c_t, U_t = u_t]] \\
&= E_{U_t \mid X_t, C_t, X_{t+1}}[\beta_0 + \beta_1 x_t + \beta_2 c_t + \beta_3 u_t] \\
&= \beta_0 + \beta_1 x_t + \beta_2 c_t + \beta_3 E_{U_t \mid X_t, C_t, X_{t+1}}[u_t] \\
&= \beta_0 + \beta_1 x_t + \beta_2 c_t + \beta_3(\alpha_0 + \alpha_1 x_t + \alpha_2 c_t + \alpha_3 x_{t+1}) \\
&= \beta_0 + \beta_3\alpha_1 + \left(\beta_1 + \beta_3\alpha_1\right)x_t + \left(\beta_2 + \beta_3\alpha_2\right)c_t + \beta_3\alpha_3 x_{t+1}
\end{aligned} \tag{1}$$

where the fourth equality follows by substitution of A5. A similar evaluation gives:

$$\begin{aligned}
&\mathrm{E}\left[Y_t \mid X_t = x_t, C_t = c\right] \\
&= \beta_0 + \left(\beta_1 + \beta_3\alpha_1\right)x_t + \left(\beta_2 + \beta_3\alpha_2\right)c_t \\
&\quad + \beta_3\alpha_3 E_{X_{t+1} \mid X_t, C_t}[x_{t+1}] \\
&= \beta_0 + \beta_3\alpha_0 + \beta_3\alpha_3\gamma_0 + \left(\beta_1 + \beta_3\alpha_1 + \beta_3\alpha_3\gamma_1\right)x_t \\
&\quad + (\beta_2 + \beta_3\alpha_2 + \beta_3\alpha_3\gamma_2)c_t
\end{aligned} \tag{2}$$

Magnitude of confounding

If we fit the "extended" model, $\mathrm{E}[Y_t \mid X_t = x_t, C_t = c_t, X_{t+1} = x_{t+1}] = b_0 + b_1 x_t + b_2 c_t + b_3 x_{t+1}$, say using least squares, the method of maximum likelihood, or generalized method of moments then the estimator $\widehat{b_1}$ satisfies

$$\widehat{b_1} \rightarrow \beta_1 + \beta_3\alpha_1 \tag{3}$$

where $\rightarrow$ means convergence in probability. This follows from Equation 2 and usual regression results (e.g., under assumptions A1–A6). Similarly, if we fit the "final" model, $\mathrm{E}[Y_t \mid X_t = x_t, C_t = c] = d_0 + d_1 x_t + d_2 c_t$, then:

$$\widehat{d_1} \rightarrow \beta_1 + \beta_3\alpha_1 + \beta_3\alpha_3\gamma_1. \tag{4}$$

These results show that $\widehat{b_1}$ and $\widehat{d_1}$ are not consistent estimators of parameter β_1, but have biases:

$$\mathrm{B}_1 = \beta_3\alpha_1 \text{ for the extended model and} \tag{5}$$

$$\mathrm{B}_2 = \beta_3\alpha_1 + \beta_3\alpha_3\gamma_1 \text{ for the final model.} \tag{6}$$

Here, bias is the difference between β_1 and the large sample limit of its estimator. Compared with the bias B_1 for the extended model, B_2 includes the additional term $\beta_3\alpha_3\gamma_1$. Thus, bias for the final model will be in the same direction but larger than that for the extended model, provided that the change of X_t is in the same direction, on average, as the change of X_{t+1} for an increase in U_t (measured by signs of α_1 and α_3) and that X_t is positively associated with X_{t+1} ($\gamma_1 > 0$). When these 2 conditions are plausible, then we expect the magnitude of confounding in the final model to exceed that in the extended model by $\beta_3\alpha_3\gamma_1$.

Alternative assumptions

These bias equations also hold in the log-linear case, when Y_t satisfies the assumptions of the Poisson distribution, and X_{t+1} and U_t have Gaussian distributions (Appendix 2).

Simulations

To assess the performance of our confounding-corrected estimator in finite samples we conducted a series of simulations. We used Poisson additive (linear) models but also evaluated log-linear models with generally similar results (Web Appendix 1; simulation results in Web Table 1 for null, and Web Table 2 for non-null available at http://aje.oxfordjournals.org/). We assessed the ability of this approach to partially correct for residual confounding using data from ongoing time-series studies of air pollution and daily EDVs analyzed using Poisson regression. We used simulations so that the true causal relationships will be known. To make the simulations realistic, the "true" expected counts were the model-predicted counts of daily EDVs for asthma (the health event) obtained by fitting a Poisson model to real, observed data for a recent 10-year period (1995–2004) in Atlanta, Georgia. We use observed, 8-hour maximum ozone levels lagged 1 day as the air pollutant of interest (Table 1). To reduce heterogeneity, we restricted analyses to the warm season (May–October).

Analyses used the following linear Poisson model:

$$E[Y_t] = b_0 + b_1 x_{t-1} + b_2 c_t \tag{7}$$

where b_0, b_1, and b_2 (a vector) are parameters in the model and c_t is a vector of controlled covariates including linear spline terms for time t (day, numbered from 1 to 185, with 2 knots for each 6-month period), terms for the moving average of maximum temperature and dew point, indicators for weekend, and indicators for year. EDV counts (Y_t) are assumed to be Poisson with mean given by equation 7. We have controlled for these covariates previously (9, 22), although sometimes including higher-order terms.

We fitted this Poisson model using Proc Genmod in SAS, version 9.3 (SAS Institute, Inc., Cary, North Carolina), to the observed counts to obtain model-predicted daily counts, which we treated as the truth. For simulations with an assumed non-null air pollution effect (Scenarios 1B-6B), the coefficient (b_1) of X_{t-1} was 1.0 to calculate the model-predicted counts; for simulations with no assumed air pollution effect (scenarios 1A–6A), we set $b_1 = 0$. We next generated simulated daily counts of EDVs using a Poisson distribution with mean equal to model-predicted counts. We then analyzed each simulated data set using models that included exposure

Table 1. Description of Observed Data for Emergency Department Visits and Environmental Conditions, Atlanta, Georgia, 1995–2004

Variable	Mean (SD)	Median	Minimum	Maximum
Daily asthma EDVs	50.2 (21.2)	46	6	144
Daily 8-hour ozone, per 25-ppb increment[a]	2.28 (0.93)	2.22	0.11	5.56
Daily maximum temperature, °C	28.4 (4.44)	29	11	39
Daily minimum temperature, °F	18.2 (4.35)	19	1	26

Abbreviations: EDV, emergency department visit; ppb; parts per billion; SD, standard deviation.
[a] Ozone concentration was measured in units of 25 ppb, approximately equal to its standard deviation.

X_{t-1} but not the negative control exposure X_{t+1}, as well as models with both.

Next, we misspecified the analytical model by intentionally omitting 1 or more covariates (scenarios 2–6). By omitting known covariates, we simulated bias that would occur if a covariate with a realistic distribution and realistic associations with exposure and outcome were inadvertently omitted. Scenarios 2–6 differed because each omits a different covariate from the full, correct model (Tables 2 and 3, second column). In our simulations $\widehat{d_1}$ refers to the X_t-coefficient estimator from the model without the negative control exposure and $\widehat{b_1}$ from the model with it. We calculated the bias in our estimators as the (median) rate ($\widehat{b_1}$ or $\widehat{d_1}$) estimated with the misspecified model (e.g., a covariate omitted) minus the true X_t-coefficient in the model used to generate the simulated data. Finally, we compared the 2 estimators $\widehat{b_1}$ and $\widehat{d_1}$ using the median bias and the mean squared error.

RESULTS

As shown in Table 2, a small to moderate bias in the rate difference was introduced in scenarios 2A–6A by dropping variables for the time spline, dew point, maximum temperature, weekend indicator, and all simultaneously, respectively. In scenarios 2A–6A, the median bias was slightly to moderately reduced when the negative control exposure was included in the regression model. However, the mean squared error was not reduced when the bias reduction was small (5A).

In scenario 2A, the bias was slightly worse when the negative control exposure was included; however, this is expected because the omitted weekend indicator was positively associated with exposure but negatively with the negative control exposure after adjustment for other modeled covariates. We found similar results under the non-null condition (Table 3).

Thus, under both the null and non-null conditions, inclusion of the ozone level 1 day after the health event tended to at least somewhat reduce the bias due to confounding when the exposure and negative control exposure were similarly associated with the omitted confounder.

EXAMPLE

We illustrate our approach using data from the ongoing time-series studies of air pollution and daily EDVs used in the simulations. We found little indication of residual confounding for ozone and asthma EDVs. However, for upper respiratory infection EDVs and same-day carbon monoxide pollution levels, the negative control exposure was significantly associated with the outcome ($P = 0.022$, $\beta = 0.81$) (Table 4). The naive estimate of the effect of carbon monoxide on upper respiratory infection visits was −0.83 ($P = 0.024$). After more complete control for time, both the naive and partially-adjusted estimates of the effect of carbon monoxide on upper respiratory infection visits were close to the null, not significant ($P = 0.598$), with little indication of residual confounding (Table 4).

Table 2. Simulation Results Under Null Hypothesis, True Effect of Air Pollutant, $b_1 = 0$

Scenario and Description		Uncorrected Estimator		Corrected Estimator	
Scenario	Type of Analytical Error/Misspecification	Bias: Median $\widehat{d_1}$ − True b_1(SE ($\widehat{d_1}$))	MSE	Bias: Median $\widehat{b_1}$ − True b_1(SE ($\widehat{b_1}$))	MSE
1A	None	0.0156 (0.220)	0.046	0.0195 (0.223)	0.048
2A	Omit weekend indicator	0.276 (0.219)	0.123	0.325 (0.222)	0.151
3A	Omit time spline variables (*t*1–*t*3[a])	−1.101 (0.216)	1.251	−0.712 (0.223)	0.560
4A	Omit dew point variable	0.214 (0.206)	0.089	0.192 (0.213)	0.080
5A	Omit maximum temperature variable	0.048 (0.191)	0.037	0.036 (0.202)	0.041
6A	Omit all the preceding variables	−4.419 (0.169)	19.5	−2.660 (0.198)	7.13

Abbreviations: MSE, mean squared error; SE, standard error.
[a] *t*1–*t*3 are continuous variables used in time splines.

Table 3. Simulation Results Under Non-Null Hypothesis, True Effect of Air Pollutant, $b_1 = 1.0$

Scenario and Description		Uncorrected Estimator		Corrected Estimator	
Scenario	Type of Analytical Error/Misspecification	Bias: Median $\widehat{d_1}$ − True b_1(SE ($\widehat{d_1}$))	MSE	Bias: Median $\widehat{b_1}$ − True b_1(SE ($\widehat{b_1}$))	MSE
1B	None	0.0027 (0.228)	1.052	−0.0100 (0.232)	1.056
2B	Omit weekend indicator	0.2526 (0.225)	0.611	0.2830 (0.231)	0.550
3B	Omit time spline variables (*t*1–*t*3[a])	−1.174 (0.223)	4.813	−0.7857 (0.231)	3.254
4B	Omit dew point variable	0.2122 (0.213)	0.677	0.1815 (0.221)	0.728
5B	Omit maximum temperature variable	0.0276 (0.199)	0.967	0.0204 (0.209)	0.980
6B	Omit all the preceding variables	−4.730 (0.174)	29.953	−2.706 (0.176)	13.721

Abbreviations: MSE, mean squared error; SE, standard error.
[a] *t*1–*t*3 are continuous variables used in time splines.

DISCUSSION

We have shown that in linear and log-linear (Appendix 2; Web Appendix 1) models, estimators in the extended model should often be less biased than those in the "final" model without the negative control exposure when in the presence of residual confounders like those we considered. Our simulations demonstrated this result empirically for linear models using actual covariates that we intentionally omitted, suggesting that one can often, but not invariably, expect use of the extended model to produce less biased estimates than the naive model under often plausible assumptions. We also found a similar reduction of bias in the log-linear regression context (results not shown), which are also common in practice.

To our knowledge, this approach is new: It differs from currently available approaches for correcting confounding. First, our approach is based on a negative control exposure with properties (P1 and P2), whereas most others use an outcome-based negative control (12, 14, 23). This difference is important; assumptions for using a negative control exposure may hold whereas those for a negative control outcome could fail, as well as the converse. Here, we emphasized temporality considerations because they help make needed assumptions plausible in environmental time-series studies. Second, our approach is regression-based and involves partial correction under plausible assumptions, whereas one published exposure-based approach (12) is primarily for sensitivity analyses and uses only stratification. The other published negative control exposure–correction approach method (24) is based on modeling the null distribution and calibrating *P* values, not correcting effect estimates.

Our approach extends to a sensitivity analysis (Web Appendix 1). Because the magnitude of the association of U_t with X_t, and X_{t+1} is generally unknown (λ not identified), sensitivity analyses can be important. Absent a correction, sensitivity analyses, or other allowance for residual confounding, the assumption is that there is no residual confounding—a rather strong assumption. Sensitivity analyses using our approach are easily implemented; they use one parameter, say λ, which is interpretable as a measure of the association of U_t with X_t compared with that with X_{t+1} and the linear association between X_t and X_{t+1}. Although somewhat like the sensitivity analysis of Lipsitch et al. (12), ours

Table 4. Results for Example Using Emergency Department Visits for Upper Respiratory Infection and Carbon Monoxide Pollution Levels

Analysis Model	Estimated Coefficient of x_t in Analysis Model	*P* Value for Null: Coefficient of $x_t = 0$ in Analysis Model	*P* Value for Negative Control Exposure	Comments
Final model[a]	−0.8323 (0.370)	0.024		
Extended model[b]	−0.7226 (0.373)	0.053	0.022	Indication of confounding; $\widehat{b_1}$ closer to null in extended model
"New" final model, time spline × year interactions[c]	0.0147 (0.380)	0.969		Include additional control of time patterns by adding interaction of time spline variables with year indicators
New extended model[d]	0.0327 (0.382)	0.931	0.598	Little change from final model; little indication of confounding

Abbreviation: SE, standard error.
[a] Final model is the additive Poisson regression model given by equation 7 and used for analysis, with the outcome being cardiovascular emergency department visits; c_t is the same covariate vector used for the simulations; and the exposure is carbon monoxide (lag 0).
[b] Extended model is the final model plus the negative control exposure.
[c] "New" final model is the final model used for analysis, with time spline × year interactions, added because of the indication of residual confounding.
[d] New extended model is the new final model with the negative control exposure added.

applies when known confounders are controlled by use of a regression model whereas theirs uses stratification (Web Appendix 1).

Although we found a scenario under the null and one under the non-null (2A or 2B) where inclusion of the future indicator did not lead to reduced bias, these 2 scenarios are consistent with expectations. In particular, a key assumption in our proof that the extended model should yield less biased estimators was that the uncontrolled (omitted) confounder be positively (or negatively) associated with both exposure and negative control exposure. However, the weekend indicator (scenarios 2A and 2B) is perhaps somewhat unusual as those associations are in the opposite directions. For example, Sundays are associated with lower ozone levels the previous day (X_{t-1}) but with higher levels in the future (X_{t+1})—a Monday when traffic is back to weekday levels. Theoretically we should expect more confounding by including the future indicator if confounding is due to omission of variables like this. Thus, consideration of substantive issues and prior knowledge is important (25). We also note that improvement in mean squared error was not as consistently attained as the improvement in bias, reflecting the slight but noticeable increase in variance of the corrected estimator. However, when bias was large (e.g., 3A–3B and 6A–6B) the mean squared error was meaningfully reduced.

We derived results for the future value of the exposure X_{t+1}, but the derivation relied mainly on properties P1 and P2 and the negative control exposure being associated with U_t in the same direction as X_t. Thus, the approach can be extended to justify the usual approach to correct for confounding given a surrogate variable for a confounder U_t. In other words, an extension can provide an alternative justification of the common practice of controlling for a surrogate of a confounder when the confounder itself is not measured directly, and suggests a formula for further correcting for confounding in sensitivity analyses.

The approach we have described emphasizes confounding and is based on the assumption that confounding is the predominant source of bias. However, if exposure-measurement error is important so that the negative control exposure provides information about the true exposure, or if properties P1 and P2 fail to hold, then inclusion of the negative control exposure could worsen bias (see also Web Appendix 1). Thus, minimizing measurement error, prior knowledge about substantive issues, and consideration of causal relationships are important.

In summary, we have presented a method to reduce residual confounding of effect estimators, even if the uncontrolled confounders are unmeasured. The approach assumes availability of a negative control exposure with 2 key properties and that the unmeasured confounders have the same direction of association with both the negative control and actual exposures. We have argued that, for environmental studies such as those of air pollution and health effects, the air pollutant level after the health effect has occurred may be have the properties needed for such an indicator. Although not the primary focus, we have also suggested a method for sensitivity analyses that extends that of Lipsitch et al. (12) because it can be used when the known confounders are controlled by regression models.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, Georgia (W. Dana Flanders, Mitchel Klein); Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, Georgia (W. Dana Flanders); Department of Environmental Health, School of Community Health Sciences, Division of Health Sciences, University of Nevada, Reno, Reno, Nevada (Matthew Strickland); and Department of Environmental and Occupational Health, Rollins School of Public Health, Emory University, Atlanta, Georgia (Mitchel Klein).

This publication was supported by the Environmental Protection Agency (grant R834799).

This publication's contents are solely the responsibility of the grantee and do not necessarily represent the official views of the US Environmental Protection Agency. Further, Environmental Protection Agency does not endorse the purchase of any commercial products or services mentioned in the publication.

Conflict of interest: none declared.

REFERENCES

1. Rothman KJ. *Modern Epidemiology*. Boston, MA: Little, Brown and Co; 1986.
2. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.
3. Greenland S, Pearl J. Causal diagrams. In: Boslaugh S, ed. *Encyclopedia of Epidemiology*. Thousand Oaks, CA: Sage Publications; 2007:149–156.
4. Flanders WD, Eldridge RC. Summary of relationships between exchangeability, biasing paths and bias. *Eur J Epidemiol*. 2015;30(10):1089–1099.
5. Fewell Z, Davey Smith G, Sterne JA. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. *Am J Epidemiol*. 2007;166(6):646–655.
6. Rosenbaum PR. *Design of Observational Studies*. New York, NY: Springer; 2010.
7. Goodman M, Barraj LM, Mink PJ, et al. Estimating uncertainty in observational studies of associations between continuous variables: example of methylmercury and neuropsychological testing in children. *Epidemiol Perspect Innov*. 2007;4:9.
8. Greenland S. Multiple-bias modelling for analysis of observational data. *J R Stat Soc Ser A Stat Soc*. 2005;168(2):267–306.
9. Flanders WD, Klein M, Darrow LA, et al. A method for detection of residual confounding in time-series and other observational studies. *Epidemiology*. 2011;22(1):59–67.
10. Flanders WD, Klein M, Darrow LA, et al. A method to detect residual confounding in spatial and other observational studies. *Epidemiology*. 2011;22(6):823–826.
11. Flanders WD, Klein M, Strickland M, et al. A method of identifying residual confounding and other violations of model assumptions. *Epidemiology*. 2009;20(6):S44–S45.
12. 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.

13. Granger CW. Some recent development in a concept of causality. *J Econom*. 1988;39(1-2):199–211.
14. Tchetgen Tchetgen E. The control outcome calibration approach for causal inference with unobserved confounding. *Am J Epidemiol*. 2014;179(5):633–640.
15. Pearl J. *Causality*. 2nd ed. Cambridg, UK: Cambridge University Press; 2009.
16. Greenland S, Brumback B. An overview of relations among causal modelling methods. *Int J Epidemiol*. 2002;31(5): 1030–1037.
17. Glymour MM, Greenland S. Causal diagrams. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott, Williams & Wilkins; 2008: 183–209.
18. Pearl J. Causal diagrams for empirical research (with discussion). *Biometrika*. 1995;82(4):669–710.
19. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*. 1980;48(4):817–838.
20. Hansen LP. Large sample properties of generalized method of moments estimators. *Econometrica*. 1982;50(4):1029–1054.
21. Newey WK, West KD. A simple, positive semi-definite, heteroskedastcity and autocorrelation consistent covariance matrix. *Econometrica*. 1987;55(3):703–708.
22. Peel JL, Tolbert PE, Klein M, et al. Ambient air pollution and respiratory emergency department visits. *Epidemiology*. 2005; 16(2):164–174.
23. Richardson DB, Keil AP, Tchetgen Tchetgen E, et al. Negative control outcomes and the analysis of standardized mortality ratios. *Epidemiology*. 2015;26(5):727–732.
24. Schuemie MJ, Ryan PB, DuMouchel W, et al. Interpreting observational studies: why empirical calibration is needed to correct p-values. *Stat Med*. 2014;33(2):209–218.
25. Robins JM, Wasserman L. On the impossibility of inferring causation from association without background knowledge. In: Glymour CN, Cooper GF, eds. *Computation, Causation, and Discovery*. Cambridge, MA: MIT Press; 1999:305–321.
26. Johnson PD, Besselsen DG. Practical aspects of experimental design in animal research. *ILAR J*. 2002;43(4):202–206.
27. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15(5): 615–625.

(Appendix follows)

APPENDIX 1

Lipsitch et al. (12) encouraged use of negative controls analogous to those in experiments (26) to detect bias in epidemiologic studies. To detect residual confounding using negative control exposures (B), they emphasized the importance of U-comparability which represents overlap between "...the set of unobserved common causes U of A and Y and the set of unobserved common causes of B and Y.... If A and B are perfectly U-comparable and B does not cause Y, then an association B-Y when analyzed according to the same model used to analyze A-Y would indicate bias in the association A-Y" (12, p. 386). Here, A is the exposure and Y the outcome.

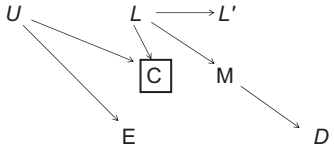
Flanders et al. (9–11) characterized the indicator for detecting model misspecification as having 2 basic properties:

P1: In correctly specified models, the indicator must be conditionally independent of the outcome, given exposure and other modeled covariates;

P2: It should be associated with the exposure and, like that exposure, with any unmeasured confounders, where a confounder is a variable on a confounding path (2, 4).

We argue that useful, perfectly U-comparable negative control exposures should have properties P1–P2, whereas some misspecification indicators with properties P1–P2 have no U-comparability. Property P1, although not assured by U-comparability, holds for the negative control exposure in the illustrative DAG (12, Figure 2) and is seemingly assumed to hold in proving that a test for confounding can be based on U-comparable negative control exposures (12, eAppendix 2). Property P2 should also hold for useful, U-comparable negative control exposures, because an unmeasured common cause will induce an association (assuming the causal path is not blocked, e.g., by stratification, making the control-negative exposure not useful). Thus, properties P1 and P2 hold or are implicitly assumed to hold for useful, perfectly U-comparable negative control exposures.

On the other hand, in some situations, a useful indicator with properties P1 and P2 need not have any U-comparability. Appendix Figure 1 illustrates this. It is a modification of Hernán's example of selection bias due to selection of firefighters (C) for study (27). Absent control for diet (M, or L itself), physical activity (E) is associated with cardiovascular disease (D) through the biasing path involving parental socioeconomic status (L). With control for diet, the biasing path is blocked and no bias is expected. L has properties P1 and P2 and should correctly indicate bias. However, indicator L (alternatively L′) and E have no U-comparability. In Appendix Figure 1 the biasing path is a confounding path (2) not classical confounding.



Appendix Figure 1.

APPENDIX 2

LOG-LINEAR MODELS

We can derive results similar to those in the main text, based on alternative assumptions:

A1*. Y_t has mean $\exp(\beta_0 + \beta_1 X_t + \beta_2 C_t + \beta_3 U_t)$, conditional on $X_t = x$, $C_t = c$, $U_t = u$.

This assumption is commonly used in modeling health outcomes measured as counts where an additional assumption is often that Y_t is Poisson.

A2*. the distribution of U_t given $X_t = x$, $C_t = c_t$ and $X_{t+1} = x_{t+1}$ denoted by $f_U(U_t \mid \beta, X_t = x_t, C_t = c_t, X_{t+1} = x_{t+1})$: has moment generating function $\Psi_U(\beta,x,c,f)$.

Claim 1B. Assumptions A1* and A2* imply:

$$\begin{aligned} &E[Y_t \mid X_t = x_t, C_t = c_t, X_{t+1} = x_{t+1}] \\ &\quad = \exp(\beta_0 + \beta_1 x_t + \beta_2 c_t + log(\Psi_U(\beta_3; x_t, c_t, x_{t+1})), \end{aligned} \tag{A1}$$

where $log(\Psi_U(\beta_3; x_t,c_t,x_{t+1}))$ is the moment generating function of the conditional distribution of U_t given $X_t = x_t$, $C_t = c_t$, $X_{t+1} = x_{t+1}$.

Proof of claim 1B: As in the main text (equations 1 and 2), we evaluate the expectation of Y_t by integrating the joint distribution of Y_t and U_t given $X_t = x$, $C_t = c$, $X_{t+1} = x_{t+1}$ over U_t to obtain:

$$\begin{aligned} &E\left[Y_t \mid X_t = x, C_t = c, X_t = x_{t+1}\right] \\ &= \int\int y f_Y\left(y \mid X_t = x, C_t = c, U_t = u\right) \\ &\qquad f_U\left(u \mid X_t = x_t, C_t = c_t, X_{t+1} = x_{t+1}\right) dy du \\ &= \int \exp\left(\beta_0 + \beta_1 x + \beta_2 c + \beta_3 u\right) \\ &\qquad f_U\left(u \mid X_t = x_t, C_t = c_t, X_{t+1} = x_{t+1}\right) du. \end{aligned} \tag{A2}$$

The second equality follows from assumption (A1*) by carrying out the integration with respect to y and where $f_U(u \mid X_t = x, C_t = c, X_{t+1} = x_{t+1})$ is the conditional probability density function of U_t. Rearrangement now shows that the moment generating function of $f_U(u \mid X_t = x_t, C_t = c_t, X_{t+1} = x_{t+1})$ appears as a factor on the right hand side:

$$\begin{aligned} &E\left[Y_t \mid X_t = x_t, C_t = c_t, X_{t+1} = x_{t+1}\right] \\ &= \exp\left(\beta_0 + \beta_1 x + \beta_2 c\right) \int \exp\left(\beta_3 u\right) \\ &\qquad f_U\left(u \mid X_t = x_t, C_t = c_t, X_{t+1} = x_{t+1}\right) du. \end{aligned} \tag{A3}$$

Substituting $\Psi_U(\beta_3;x,c,f)$ for its defining integral gives:

$$\begin{aligned} &E\left[Y_t \mid X_t = x_t, C_t = c_t, X_{t+1} = x_{t+1}\right] \\ &= \exp\left(\beta_0 + \beta_1 x_t + \beta_2 c_t\right) \Psi_U\left(\beta_3; x_t, c_t, x_{t+1}\right) \\ &= \exp(\beta_0 + \beta_1 x + \beta_2 c + \log\left(\Psi_U\left(\beta_3; x_t, c_t, x_{t+1}\right)\right) \end{aligned} \tag{A4}$$

proving claim 1B.

Corollary 1: If $U_t = \alpha_0 + \alpha_1 X_t + \alpha_2 C_t + \alpha_3 X_{t+1} + \varepsilon_{U,t}$ with $\varepsilon_{U,t}$ Gaussian, then $E[Y_t \mid X_t = x_t, C_t = c_t, X_{t+1} = x_{t+1}] = \exp(\beta_0^*(\beta_1 + \beta_3\alpha_1)x_t + (\beta_2 + \beta_3\alpha_2)c_t + (\alpha_3 + \beta_3\alpha_3)x_{t+1})$

Proof: The result follows by substitution, since $\Psi_U(\beta_3; x_t, c_t, x_{t+1}) = e^{\beta_3(\alpha_0+\alpha_1 x_t+\alpha_2 c_t+\alpha_3 x_{t+1})} + \frac{1}{2}\sigma^2\beta_3^2$.

Claim 1C. Assumptions (A1*) and (A2*), and those for claim 1B and corollary 1 imply:

$$\begin{aligned}\mathrm{E}\big[Y_t \mid X_t = x_t,\, C_t = c_t\big] &= \exp(\beta_0^* + (\beta_1 + \beta_3\alpha_1)x_t + (\beta_2 + \beta_3\alpha_2)c_t \\ &\quad + \log(\Psi_X(\alpha_3 + \beta_3\alpha_3);\, x_t,\, c_t))\end{aligned} \tag{A5}$$

where $(\Psi_X((\alpha_3 + \beta_3\alpha_3); x_t, c_t, x_{t+1})$ is the moment generating function of the conditional distribution of X_{t+1} given $X_t = x_t, C_t = c_t$ with variable $(\alpha_3 + \beta_3\alpha_3)$.

Proof of claim 1C: This follows from claim 1B and corollary 1, by a proof similar to that used for claim 1B.

Corollary 2: If, in addition to the assumptions of corollary 1, $X_{t+1} = \gamma_0 + \gamma_1 X_t + \gamma_2 C_t + \varepsilon_{Xt+1}$ with ε_{Xt+1} Gaussian, then $\mathrm{E}[Y_t \mid X_t = x_t,\, C_t = c_t] = \exp(\beta_0^{**}(\beta_1 + \beta_3\alpha_1 + \beta_3\alpha_3\gamma_1)x_t + (\beta_2 + \beta_3\alpha_2 + \beta_3\alpha_3\gamma_2)c_t)$.

Proof: The result follows from claim 1C by substitution of the moment generating function of X_{t+1}.

Note: Under the assumptions corollaries 1 and 2, the mean of Y is log-linear in the variables X_t, C_t, X_{t+1} or X_t, C_t. Thus, if Y is Poisson, the maximum likelihood estimators of β_1 will converge as indicated in equations 3 and 4 of the main text. Thus, the bias of equations 5 and 6 will hold under the alternative assumptions A1*, A2*, corollaries 1 and 2, and A3–A6.