



Practice of Epidemiology

The Control Outcome Calibration Approach for Causal Inference With Unobserved Confounding

Eric Tchetgen Tchetgen*

* Correspondence to Dr. Eric Tchetgen Tchetgen, Department of Biostatistics, Harvard University, 677 Huntington Avenue, Kresge, Room 822, Boston, MA 02115 (e-mail: etchetge@hsph.harvard.edu).

Initially submitted May 29, 2013; accepted for publication November 13, 2013.

Unobserved confounding can seldom be ruled out with certainty in nonexperimental studies. Negative controls are sometimes used in epidemiologic practice to detect the presence of unobserved confounding. An outcome is said to be a valid negative control variable to the extent that it is influenced by unobserved confounders of the exposure effects on the outcome in view, although not directly influenced by the exposure. Thus, a negative control outcome found to be empirically associated with the exposure after adjustment for observed confounders indicates that unobserved confounding may be present. In this paper, we go beyond the use of control outcomes to detect possible unobserved confounding and propose to use control outcomes in a simple but formal counterfactual-based approach to correct causal effect estimates for bias due to unobserved confounding. The proposed control outcome calibration approach is developed in the context of a continuous or binary outcome, and the control outcome and the exposure can be discrete or continuous. A sensitivity analysis technique is also developed, which can be used to assess the degree to which a violation of the main identifying assumption of the control outcome calibration approach might impact inference about the effect of the exposure on the outcome in view.

bias; case-control study; counterfactual; negative control outcome; observational study; unobserved confounding

Abbreviations: COCA, control outcome calibration approach; ETT, effect of treatment on the treated; OLS, ordinary least squares.

Unobserved confounding is a well-known threat to valid causal inference, which can seldom be ruled out with certainty in an observational study. An approach that is sometimes used in epidemiologic practice to evaluate whether empirical results are subject to confounding bias entails evaluating whether the treatment or exposure in view is associated with a so-called negative control outcome upon adjustment for observed confounders (1–4). An outcome is said to be a valid negative control variable to the extent that it is influenced by unobserved confounders of the exposure effects on the outcome in view, although not directly influenced by the exposure (3). Thus, a negative control outcome found to be empirically associated with the exposure indicates that unobserved confounding may be present for the primary outcome provided that, upon adjustment for observed covariates, there is no unobserved confounder of the negative control outcome that does not also confound the primary outcome (3).

Suppose that in an application, a negative control outcome is found to be associated with the treatment in view, thus correctly indicating the presence of unobserved confounding. Then, it may seem natural to consider the observed association between the exposure and the control outcome as an estimate of bias due to unmeasured confounding, and one may be tempted to simply correct the confounded estimate of the exposure-outcome association by subtracting the estimated bias. Although this ad hoc bias correction approach may sometimes be appropriate, it often is not. A difficulty with the approach is that it relies on the key assumption that the bias observed for the negative control outcome is somehow equivalent to the bias one would have observed between the exposure and the primary outcome under the null hypothesis of no causal effect of the exposure. A natural prerequisite for this “bias equivalence” assumption is that the outcomes are measured on comparable scales, which would be the case if, for example, the control outcome was a preexposure

measure of the outcome process. However, outside of this special case, the assumption may not be appropriate if the outcomes are clearly measured on different scales, such as, for example, if the negative control outcome were dichotomous and the primary outcome were continuous. The assumption would then likely be violated because an additive association of the exposure with the control outcome would a priori be restricted by the binary nature of the outcome, whereas the additive association of the outcome in view with the exposure would not.

In this paper, we propose to use control outcomes in a simple but formal counterfactual or potential outcome-based approach to correct causal effect estimates for bias due to unobserved confounding, while avoiding the above assumption of bias equivalence. The proposed control outcome calibration approach (COCA) is motivated by noting that the ultimate set of unobserved confounders in an analysis relating the exposure to the primary outcome entails the set of counterfactuals for the outcome of interest under all possible treatment values. This is because conditioning on the set of counterfactuals for the primary outcome renders the latter constant and, thus, independent of the treatment assignment, a sufficient requirement for identification of a causal effect of treatment on the outcome. Furthermore, when, as we have assumed, there is no unobserved confounder of the treatment effect on the control outcome that does not confound the treatment effect on the primary outcome, it is natural to expect that the set of the observed covariates together with the counterfactuals for the primary outcome under all possible treatment values also suffices to identify the effects of treatment on the control outcome. This is the fundamental assumption made when using the COCA. In the context of a negative control outcome, the COCA produces an effect estimate of the association between treatment and the primary outcome under an assumed causal model. This effect estimate is obtained upon calibrating the parameters of the causal model so that the set of all counterfactuals for the primary outcome recovered from the observed data under the calibrated model suffices, together with the observed covariates, to fully adjust for confounding in a regression analysis for the negative control outcome and correctly recovers the null association between the exposure and the control outcome. The COCA is separately developed for a continuous outcome and a binary outcome, and the control outcome and the exposure in view can be either binary, a count, or continuous. Finally, a sensitivity analysis technique is developed in the Appendix to assess the extent to which a violation of the main identifying assumption of the COCA might affect inference about the effect of the exposure on the outcome in view.

THE COCA FOR ADDITIVE CAUSAL EFFECTS

We introduce the notation and definitions we will be using throughout. Let A denote the exposure or treatment received by an individual, let Y denote a posttreatment outcome, and let C denote the value of a set of observed preexposure confounding variables of the effects of A on Y . Let U denote a set of unmeasured preexposure confounders of the effects of A . Let Z denote a negative control outcome variable. Then, the

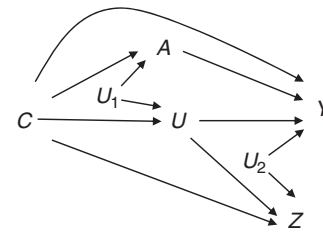


Figure 1. Causal diagram depicting unmeasured confounding of the $A - Y$ association and the negative control outcome Z , where A represents exposure; Y represents outcome; U , U_1 , and U_2 represent unobserved confounders; and C represents observed confounders.

relationships between these variables may be depicted as in the causal diagram in Figure 1.

Figure 1 gives a graphical representation of the assumption that adjustment for both C and U would suffice to account for confounding of the causal effects of A on Y and on Z , respectively. The variables U_1 and U_2 on the graph represent the possible presence of unobserved factors that correlate U with A and Y with Z , respectively. Formally, this graph is a causal directed acyclic graph representing the observed variables together with both observed and unobserved common causes. As shown in Figure 1, Z is an ideal negative control outcome because it is not directly influenced by exposure, but it is influenced by the unmeasured confounders of the exposure-outcome association (3).

We also consider counterfactuals or potential outcomes under possible interventions on the treatment. Let Y_a denote a subject's outcome if treatment A were set, possibly contrary to fact, to a . Also, let Z_a denote a subject's counterfactual value for Z if A were set to a . By assumption, $Z_a = Z$, $a = 0, 1$, for a negative control outcome and, by the consistency assumption usually made in the causal inference literature, $Y_a = Y$ if $A = a$. The assumption encoded in Figure 1, that $\{U, C\}$ suffices to account for confounding of the causal associations between A and Y and between A and Z , respectively, is equally expressed using counterfactuals

$$Y_a \perp\!\!\!\perp A \mid \{C, U\} \text{ and} \tag{1}$$

$$Z_a \perp\!\!\!\perp A \mid \{C, U\}; \quad a = 0, 1. \tag{2}$$

Note that U is an unobserved confounder for the effects of A on Y in the sense that, although equation 1 is satisfied, it is also the case that

$$Y_a \not\perp\!\!\!\perp A \mid C, \quad a = 0, 1,$$

so that C alone does not suffice to adjust for confounding, whereas $\{U, C\}$ does.

Focusing on negative control outcomes, one may formalize its definition as follows.

Definition 1. Z is said to be a negative control outcome if

$$Z_a = Z \text{ for all individuals, and}$$

$$Z \not\perp\!\!\!\perp A \mid C \Leftrightarrow Y_a \not\perp\!\!\!\perp A \mid C \text{ for all } a.$$

Definition 1 formalizes the idea that the exposure-negative control outcome association cannot be confounded by a

variable that does not also confound the exposure-outcome association. Although this assumption may suffice to detect the presence of unobserved confounding, it does not suffice to identify the causal effect of A on Y . To make progress, we make an additional identifying assumption, depicted in the graph in Figure 2, which is similar but more elaborate than the graph in Figure 1, and which encodes the following assumption.

Assumption 1: Let $\mathcal{Y}_A = \{Y_a : a \in \mathcal{A}\}$ denote the set of all counterfactuals for the primary outcome under all possible values of treatment in the set \mathcal{A} . Then, the treatment assignment is independent of Z_a conditional on $\{C, \mathcal{Y}_A\}$, or

$$Z = Z_a \perp\!\!\!\perp A \mid \{C, \mathcal{Y}_A\}.$$

Assumption 1 states that, even though C may not suffice to account for unobserved confounding to make correct inferences about the relation between A and Z , that is,

$$Z = Z_a \not\perp\!\!\!\perp A \mid C,$$

enriching the adjustment set of covariates with the set \mathcal{Y}_A suffices to adjust for confounding. Note that in Figure 2, U represents unobserved predictors of the outcomes, whereas W represents unobserved factors that may have influenced treatment selection. Our causal model encodes an assumption that these 2 factors are independent conditional on (Y_0, Y_1, C) and not otherwise. Thus, if observed, (W, C) and (U, C) would also suffice to account for confounding of the $A - Y$ and $A - Z$ associations, respectively. An intuitive explanation of assumption 1 is obtained upon noting that, when A is binary, Y_0 and Y_1 can be viewed as baseline covariates that capture all relevant information about an individual's health status prior to treatment assignment; thus, such variables are ideal proxy measures of unobserved factors that may influence treatment selection (i.e., W) and unobserved risk factors of the outcomes (i.e., U).

To further ground ideas, it is helpful to consider the familiar context of a point exposure study with A binary and where we assume that the observed data (Y, Z, A, C) follow the model

$$Z = g_1(U, C), \tag{3}$$

$$A = g_2(W, C), \text{ and} \tag{4}$$

$$Y = Y_0 + \psi_0 A. \tag{5}$$

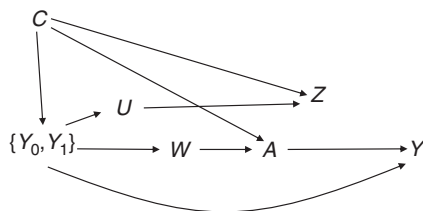


Figure 2. Causal diagram depicting unobserved confounding by U and W and the negative control outcome Z , where A represents exposure; Y represents outcome; U and W represent unobserved predictors of Z and A , respectively; C represents observed confounders; and $\{Y_0, Y_1\}$ are counterfactual outcomes for different exposure values.

This model is consistent with the graph depicted in Figure 2. The variables (U, W) are not observed, and under the consistency assumption $Y = Y_A$, so that Y_a is observed only for persons with $A = a$, and Y_{1-A} remains unobserved. The model allows the relation between (U, C) and Z encoded by the function g_1 to remain unrestricted and encodes the fact that A does not directly influence Z . Likewise, the model allows the relation between (W, C) and A encoded by the function g_2 to remain unrestricted. The parameter

$$\psi_0 = Y_1 - Y_0$$

encodes a constant additive individual causal effect of A on Y . This is a strong assumption because it implies so-called rank-preservation of individuals' counterfactuals under treatment versus control conditions. The assumption can be relaxed somewhat by incorporating interactions of A with components of C , and the assumption can be dropped entirely for binary outcomes, as we later demonstrate. The model is consistent with the graph in Figure 2, because $Y_a \perp\!\!\!\perp A \mid C$, and therefore, the effect of A on Y is confounded due to dependence between Y_a , U , and W , even if one conditions on C . Note that although the model specifies an additive causal effect, the relation between Y_0 and (A, Z, C) is otherwise unrestricted.

To describe the COCA, let $Y(\psi) = Y - \psi A$, and note that $Y_0 = Y(\psi_0)$ and $Y_1 = Y(\psi_0) + \psi_0$. Further note that under the model, conditioning on Y_0 is equivalent to conditioning on the set $\mathcal{Y}_A = \{Y_0, Y_1\}$, because the 2 counterfactuals in the set are deterministically related. This, in turn, implies that under assumption 1,

$$Z \perp\!\!\!\perp A \mid \{C, Y(\psi)\}$$

if and only if $\psi = \psi_0$. This is the key insight by which the COCA identifies ψ_0 . A regression-based approach to implement the COCA then entails searching for the parameter ψ such that

$$E(Z \mid A, C, Y(\psi)) = E(Z \mid C, Y(\psi)). \tag{6}$$

For example, a simple implementation of the approach uses linear models, whereby for each value of ψ , one obtains an estimate of the regression model by using ordinary least squares (OLS).

$$E(Z \mid A, C, Y(\psi)) = \beta_1 + \beta_2' C + \beta_3 Y(\psi) + \beta_4 A, \tag{7}$$

with estimated coefficients $(\hat{\beta}_1(\psi), \hat{\beta}_2(\psi), \hat{\beta}_4(\psi))$. Then, a 95% confidence interval for ψ_0 consists of all values of ψ for which a valid test of the null hypothesis $\hat{\beta}_4(\psi) = 0$ fails to reject at the 0.05 type 1 error level. The latter hypothesis test may be performed by verifying whether the interval $\hat{\beta}_4(\psi) \pm 1.96 \widehat{SE}(\hat{\beta}_4(\psi))$ contains 0, with $\widehat{SE}(\hat{\beta}_4(\psi))$ the OLS estimate of the standard error (SE) of $\hat{\beta}_4(\psi)$. An alternative, potentially simpler, approach is obtained by evaluating

the regression model 7 at ψ_0 under model 5,

$$\begin{aligned} E(Z|A, C, Y(\psi_0)) &= E(Z|C, Y(\psi_0)) \\ &= \beta_1 + \hat{\beta}'_2 C + \beta_3 Y(\psi_0), \quad (8) \\ &= \beta_1 + \hat{\beta}'_2 C + \beta_3 Y + \hat{\beta}'_4 A \end{aligned}$$

where $\hat{\beta}'_4 = -\beta_3 \psi_0$, assuming that $\beta_3 \neq 0$. The parameter ψ_0 is then identified by regressing Z on (C, Y, A) via OLS, which produces the estimate $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$ and

$$\hat{\psi} = -\hat{\beta}'_4 / \hat{\beta}_3.$$

A corresponding variance estimate can be obtained by a straightforward application of the delta method, giving

$$\text{var}(\hat{\psi}) = \frac{\hat{\sigma}_3^2}{\hat{\beta}_3^2} - 2\hat{\beta}_4^* \frac{\hat{\sigma}_{34}}{\hat{\beta}_3^3} + \hat{\beta}_4^{*2} \frac{\hat{\sigma}_4^2}{\hat{\beta}_3^4},$$

where

$$\begin{pmatrix} \hat{\sigma}_3^2 & \hat{\sigma}_{34} \\ \hat{\sigma}_{34} & \hat{\sigma}_4^2 \end{pmatrix}$$

is the OLS estimate of the variance covariance matrix of $(\hat{\beta}_3, \hat{\beta}_4)'$.

The first COCA strategy described in the previous paragraph is quite general in the sense that the regression model $E(Z|A, C, Y(\psi))$ can be estimated using any appropriate regression approach, including any generalized linear model with appropriate link function, say, the logit link or the log link for binary or count Z , respectively. Furthermore, for a given choice of model, a statistical test of the null hypothesis displayed in equation 6 can be performed using a standard likelihood ratio test, a score test, or a Wald test statistic, regardless of the underlying functional form of the regression. In principle, a more flexible model could be used to estimate the left-hand side of equation 6, including nonlinear terms and interactions to improve the fit of the model. Note also that our choice of a constant additive causal model (model 5) is made mainly for convenience, and that the underlying causal model can be easily modified to incorporate possible effect heterogeneity with observed covariates. For instance, model 5 can be replaced with $Y = Y_0 + (A, A \times C')\psi_0$, thus incorporating effect modification of the causal effect of A on Y with respect to C .

Note also that the simplified second COCA strategy described above is tailored to the linear functional form of both models 5 and 6. Although the models make some simplifying assumptions, the approach reveals a simple strategy to test and correct for unmeasured confounding using the COCA under the foregoing formalization. Under the sharp null of no causal effect of A on Y , that is, $\psi_0 = 0$, a straightforward test of no unmeasured confounding then entails assessing whether Z and A are additively associated conditional on Y and C . This strategy is reasonable, because under the sharp null, $Y = Y_0$ is a proxy of unmeasured common causes of Y

and A and therefore, adjustment for Y in the regression of Z on A essentially amounts to adjustment for unobserved confounding to the extent that Z is a valid negative control outcome for the effects of A on Y . The COCA formalizes this basic idea so that it may be used equally both under and away from the sharp null hypothesis, that is, even when $\psi_0 \neq 0$, by leveraging the causal model to recover the proxy measure of unobserved confounding Y_0 to use for adjustment in the negative control outcome regression model. This essentially describes the COCA, which accomplishes the above task by calibrating the causal model by varying the value of ψ until confounding control based on $Y(\psi)$ in the control outcome regression is satisfactory.

DATA EXAMPLE: CHROMOSOME DAMAGE FROM CONTAMINATED FISH

We use the proposed approach in a reanalysis for the purpose of illustration of a simplified version of a study conducted by Skerfving et al. (5) on the relation between consumption of contaminated fish and chromosome damage. The authors studied 23 subjects who had eaten large quantities of fish contaminated with methylmercury ($A = 1$). These subjects lived in different areas in Sweden and included fishermen, fishermen's wives, workmen, farmers, and clerks. Each of the 23 exposed subjects reported eating at least 3 meals a week of contaminated fish for more than 3 years. The comparison group included 16 subjects who were exposed to substantially lower amounts of contaminated fish and who reported consuming less fish of all kinds ($A = 0$). These subjects were from the Stockholm metropolitan area and included clerks, craftsmen, porters, workmen, and a glass washer. The 2 outcomes of primary interest consist of the amount of mercury found in the person's blood, recorded in ng/g and log transformed for the analysis (Y), and the percent of cells exhibiting a particular chromosome abnormality called C_u cells (Y^*). Although the original study considered a variety of chromosome abnormalities, we proceed as in the report by Rosenbaum (2), who focused on these particular outcomes to illustrate the use of negative control outcomes to detect the presence of unobserved confounding. The negative control outcome in this example consists of a count of other health conditions experienced by each of the 39 subjects enrolled in the study (Z). This composite outcome includes other diseases such as hypertension and asthma, drugs taken regularly, diagnostic radiography over the previous 3 years, and viral diseases such as influenza. Although these outcomes were observed during the period when exposed individuals consumed contaminated fish, one does not expect that eating fish contaminated with methylmercury causes influenza or asthma or prompts radiography of the hip or lumbar spine. We make the additional assumption 1 (with C the empty set), and thus assume that Z may be used to detect and correct for unobserved confounding for the association between A and Y using the COCA. Referring back to Figure 1, our assumption is thus that there is no unobserved common cause of A , and any chronic condition used to define Z that does not also confound the relation between fish consumption and mercury in the blood Y , and thus Z , may be used to account for unobserved confounding for the

association between A and Y using the COCA. Similar assumptions are made about Y^* . For each outcome, we assume the constant additive effect model 5, that is, $Y = Y_0 + \psi_0 A$ and $Y^* = Y_0^* + \psi_0^* A$, so that ψ_0 encodes the causal effect of A on Y and likewise for ψ_0^* . For the COCA, we assume events contributing to the count Z are mutually independent and take Z to be Poisson distributed with conditional mean

$$\begin{aligned}\log E(Z|A, Y(\psi_0)) &= \beta_1 + \beta_3 Y(\psi_0) \\ &= \beta_1 + \beta_3 Y + \beta_4^\dagger A\end{aligned}$$

where, as before, $\beta_4^\dagger = -\beta_3 \psi_0$. Thus, we compute the COCA estimator $\hat{\psi} = -\hat{\beta}_4^\dagger / \hat{\beta}_3$, where $(\hat{\beta}_1, \hat{\beta}_3, \hat{\beta}_4^\dagger)$ is obtained by maximum likelihood. For comparison, we also compute the standard OLS estimator $\tilde{\psi}$ of the linear (crude) association between A and Y^* . Similar models were used for Y^* .

The OLS crude estimate of ψ_0 was 2.77 for Y (95% confidence interval: 2.26, 3.27), and was comparable to the COCA estimate of 2.32 (95% confidence interval: 1.36, 3.28), thus indicating little empirical evidence of unobserved confounding. In contrast, the OLS crude estimate of ψ_0 was 1.70 for Y^* (95% confidence interval: 0.426, 2.97) and was considerably smaller than the COCA estimate of 4.14 (95% confidence interval: 0.08, 8.19). The large difference between the COCA estimate and the OLS estimate is suggestive of unobserved confounding; however, the COCA estimate of the causal effect was also considerably more variable than the OLS estimate. To formally assess whether the OLS and COCA estimates are within sampling variability of each other, that is, that there was no bias due to unobserved confounding, we implemented a Hausman test (6), which entails computing a confidence interval for the limiting value of $\hat{\psi} - \tilde{\psi}$ using the simple formula

$$\hat{\psi} - \tilde{\psi} \pm 1.96 \times \{\hat{\sigma}^2(\hat{\psi}) - \hat{\sigma}^2(\tilde{\psi})\}^{1/2}$$

and verifying whether 0 falls in the above interval as would be consistent with the null hypothesis of no confounding, where $\hat{\sigma}^2(\hat{\psi})$ and $\hat{\sigma}^2(\tilde{\psi})$ are consistent estimates of the asymptotic variance of $\hat{\psi}$ and $\tilde{\psi}$, respectively (6). Note that, although under the null hypothesis of no unobserved confounding, $\hat{\sigma}^2(\hat{\psi}) - \hat{\sigma}^2(\tilde{\psi})$ converges to a positive number with increasing sample size, it can be negative in the observed finite sample or if the null hypothesis is false, in which case its square root is not a real number. In such cases, it is recommended to instead use the nonparametric bootstrap approach to estimate the variance of $\hat{\psi} - \tilde{\psi}$. The above 95% confidence intervals were (-0.36, 1.268) for Y , indicating no statistically significant evidence of bias due to unobserved confounding for the crude association between consumption of contaminated fish and level of mercury in the blood; and (-1.40, 6.28) for Y^* , indicating no statistically significant evidence of bias due to unobserved confounding for the crude association between consumption of large quantities of fish contaminated with methylmercury and percent chromosome abnormality. In closing, we should note that the foregoing analysis and its conclusions may dismiss unobserved confounding by certain, but not all, hidden variables. Assumption 1 may not be entirely credible if, say, an

ingredient other than methylmercury in contaminated fish caused the chromosomal abnormalities, or if lack of eating meat by fish consumers were the culprit. This is because the unobserved confounder may no longer be shared between the outcome and the negative control outcome, so that the negative control outcome would have no power to detect unobserved confounding, let alone correct for it. The analysis should be interpreted with caution, particularly because no additional covariates C were available for adjustment, which would have helped to make the identifying assumption more credible.

THE COCA FOR A DICHOTOMOUS OUTCOME

The foregoing presentation focused primarily on settings in which the outcome in view is continuous. Dichotomous outcomes are also quite common in epidemiologic practice; thus, in this section, we extend the COCA to the context of a binary Y , and we present similar methodology to estimate the effect of treatment on the treated (ETT), with

$$ETT = E(Y_1 - Y_0 | A = 1).$$

To proceed, one may note that the observed crude difference $E(Y = 1 | A = 1) - E(Y | A = 0)$ is biased for the ETT, with

$$\begin{aligned}\text{Bias} &= E(Y = 1 | A = 1) - E(Y | A = 0) - E(Y_1 - Y_0 | A = 1) \\ &= \Pr(Y_0 = 1 | A = 0) - \Pr(Y_0 = 1 | A = 1).\end{aligned}$$

Therefore, to nonparametrically identify the ETT, one must identify $\Pr(Y_0 = 1 | A = 1)$. Suppose that Z satisfies assumption 1, with $\mathcal{A} = \{0\}$, that is,

$$Z \perp\!\!\!\perp A | \{C, Y_0\}.$$

Under the assumption, the conditional mean $E(Z|A)$ may be written as

$$\begin{aligned}E(Z|A) &= E(Z|A, Y_0 = 1)\Pr(Y_0 = 1|A) \\ &\quad + E(Z|A, Y_0 = 0)\Pr(Y_0 = 0|A) \\ &= E(Z|A = 0, Y = 1)\Pr(Y_0 = 1|A) \\ &\quad + E(Z|A = 0, Y = 0)(1 - \Pr(Y_0 = 0|A)) \\ &= \Pr(Y_0 = 1|A)\{E(Z|A = 0, Y = 1) \\ &\quad - E(Z|A = 0, Y = 0)\} \\ &\quad + E(Z|A = 0, Y = 0),\end{aligned}$$

which gives

$$\begin{aligned}\Pr(Y_0 = 1|A) &= \frac{E(Z|A) - E(Z|A = 0, Y = 0)}{E(Z|A = 0, Y = 1) - E(Z|A = 0, Y = 0)}, \quad (9)\end{aligned}$$

provided that

$$E(Z|A = 0, Y = 1) - E(Z|A = 0, Y = 0) \neq 0.$$

Thus,

$$\text{Bias} = \Pr(Y = 1|A = 0) - \frac{E(Z|A = 1) - E(Z|A = 0, Y = 0)}{E(Z|A = 0, Y = 1) - E(Z|A = 0, Y = 0)}$$

and

$$\text{ETT} = \Pr(Y = 1|A = 1) - \frac{E(Z|A = 1) - E(Z|A = 0, Y = 0)}{E(Z|A = 0, Y = 1) - E(Z|A = 0, Y = 0)}.$$

The result states that $\Pr(Y_0 = 1|A = 1)$ is nonparametrically identified by the ratio of differences displayed in equation 9, and because $\Pr(Y_1 = 1|A = 1) = \Pr(Y = 1|A = 1)$ by the consistency assumption, this in turn implies that the ETT is nonparametrically identified. Note that Z can be either discrete or continuous, and that the approach easily incorporates observed confounders C . In fact, by following similar steps as above, one can show that

$$\begin{aligned} \text{ETT}(c) &= E(Y_1 - Y_0|A = 1, c) \\ &= \Pr(Y = 1|A = 1, c) \\ &\quad - \frac{E(Z|A = 1, c) - E(Z|A = 0, Y = 0, c)}{E(Z|A = 0, Y = 1, c) - E(Z|A = 0, Y = 0, c)}, \end{aligned} \quad (10)$$

and the marginal ETT is given by

$$\text{ETT} = \sum_c \text{ETT}(c)\Pr(c|A = 1).$$

Estimation could then proceed by fitting using standard maximum likelihood, parametric models for $\Pr(Y = 1|A, C)$ and $E(Z|A, C)$ and plugging the latter into equation 10. A straightforward application of the delta method could be used to obtain standard errors for the resulting estimator, or alternatively, the nonparametric bootstrap could also be used.

Note that when C is not empty, one may also write

$$\begin{aligned} E(Z|A = 1, c) &= E(Z|A = 1, Y = 1, c)\Pr(Y = 1|A = 1, c) \\ &\quad + E(Z|A = 1, Y = 0, c) \\ &\quad \times \{1 - \Pr(Y = 1|A = 1, c)\}, \end{aligned} \quad (11)$$

which may be used to evaluate equation 10. This would simplify estimation by allowing the analyst to fit separate regression models for $E(Z|A, Y, C)$ and $\Pr(Y|A, C)$, say, standard logistic regression models if Z and Y are both binary, which are ensured not to conflict with a model for $E(Z|A = 1, C)$ obtained using equation 11. Inference for the causal risk ratio parameter

$$\gamma(a, c) = \frac{\Pr(Y_a = 1|A = 1, c)}{\Pr(Y_0 = 1|A = 1, c)}$$

or for the causal odds ratio parameter

$$\gamma(a, c) = \frac{\Pr(Y_a = 1|A = 1, c)\Pr(Y_0 = 0|A = 1, c)}{\Pr(Y_a = 0|A = 1, c)\Pr(Y_0 = 1|A = 1, c)}$$

can likewise be obtained by simply using the above expression for $\Pr(Y_0 = 1|A = 1, C)$ as a baseline risk in a standard (multiplicative or logistic) regression model. To fix ideas, suppose that $\gamma(a, c) = \psi_0 a$ on a given scale (either risk ratio or odds ratio scale), so that we assume that the effect of treatment is constant in the treated across levels of c . Then, one can estimate ψ_0 by fitting the regression model

$$g\{\Pr(Y_a = 1|A = a, c)\} = \psi_0 a + g\{\Pr(Y_0|A = a, c)\}, \quad (12)$$

where g is either the logit link function or the log link function, and $\Pr(Y_0 = 1|A = a, c)$ is estimated by evaluating equation 9. The “no interaction” assumption is easily relaxed by replacing the causal model with a model incorporating interactions between A and C .

Case-control studies are quite common in epidemiologic practice, and the COCA extends to this context but requires some modification to appropriately account for the study design, which is provided in the Appendix.

DISCUSSION

Some degree of unobserved confounding is almost certainly present in most observational studies. For this reason, it was recently argued that researchers should routinely supplement the primary analysis of such observational studies with some form of negative control outcome (or negative control exposure) analysis to demonstrate that exposure effects known not to be present in the population are in fact not observed in the study sample (1, 3). The extent to which such an analysis may reveal unobserved confounding bias relies on the non-empirically verifiable assumption that the negative control outcome is carefully chosen so that it is solely influenced by observed and unobserved confounders of the exposure-outcome relation in view. Here, we propose to use a negative control outcome not only to detect, but also to correct for unmeasured confounding bias. Some analytical strategies are described for continuous and binary outcomes, under the assumption that the primary outcome that would be observed were exposure withheld in the population suffices together with observed confounders to completely account for confounding of the exposure-negative control outcome association. We leverage this assumption to calibrate the causal effect, so that the assumption is empirically met. A sensitivity analysis technique is also described in the Appendix, which allows one to assess the degree to which a violation of the main identifying assumption, assumption 1, could alter the results.

Though a regression-based calibration approach is emphasized, in the context of a continuous outcome, in principle, upon obtaining the proxy measure of unobserved confounding, one could evaluate the adjusted association between the exposure and the control outcome using alternative approaches to the regression approach taken here without

additional difficulty, for example, propensity score methods or doubly robust estimation (7–9).

Time-to-event outcomes are also common in epidemiologic practice, and the methods developed in this paper can, in principle, be extended to allow for a censored time-to-event outcome. For example, the standard rank preserving structural accelerated failure time model (10) relates the log event time to the treatment using an additive model of the form given by model 5 and, therefore, the methodology described herein immediately applies for this model. However, one would have to ensure that the negative control outcome and the primary outcome are not competing risks, and one would also need to appropriately account for censoring. Similar methodology for the Cox proportional hazards model (11) or for the Aalen additive hazards model (12) still needs to be developed.

A positive control outcome can be defined for an outcome with a well-established nonnull causal association with the exposure, which is confounded in the observed sample by a subset of unobserved confounders for the exposure effects on the primary outcome. Positive control outcomes can, in a manner similar to negative controls, be used to detect unobserved confounding by verifying whether the known association is replicated in the observed sample. The methods described in this paper could be extended for use with positive control outcomes.

Negative control exposures are also quite common in epidemiologic practice (3, 13, 14). These are observed exposures known not to causally influence the primary outcome. It may be possible to also develop an approach similar to that given in this paper to leverage negative control exposures to correct for unobserved confounding bias. This will be investigated elsewhere.

ACKNOWLEDGMENTS

Author affiliation: Department of Biostatistics, Harvard University, Boston, Massachusetts (Eric Tchetgen Tchetgen).

This work was supported by the National Institutes of Health (grants AI104459, ES019712, GM088558, and AI51164).

The author thanks Drs. Stephen R. Cole and David Richardson for helpful comments on an earlier version of the manuscript.

Conflict of interest: none declared.

REFERENCES

- Rosenbaum P. The role of known effects in observational studies. *Biometrics*. 1989;45(2):557–569.
- Rosenbaum P. *Observational Studies*. 2nd ed. New York, NY: Springer-Verlag; 2002.
- 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.
- Flanders WD, Klein M, Strickland M, et al. A method for detection of residual confounding in time-series and other observational studies. *Epidemiology*. 2011;22(1):59–67.
- Skerfving S, Hansson K, Mangs C, et al. Methylmercury-induced chromosome damage in man. *Environ Res*. 1974;7(1):83–98.
- Hausman JA. Specification tests in econometrics. *Econometrica*. 1978;46(6):1251–1271.
- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41–55.
- Robins J, Rotnitzky A. On double robustness. Discussion of a paper by Peter Bickel and Jaimyoung Kwon. *Stat Sin*. 2001;4:920–936.
- Robins JM, Mark SD, Newey WK. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*. 1992;48(2):479–495.
- Robins JM, Tsiatis AA. Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika*. 1992;79(2):311–319.
- Cox DR. Regression models and life tables (with discussion). *J R Stat Soc Series B*. 1972;34(2):187–220.
- Aalen OO. A linear regression model for the analysis of life times. *Stat Med*. 1989;8(8):907–925.
- Davey Smith G. Re: “Negative exposure controls in epidemiologic studies” [letter]. *Epidemiology*. 2011;23(2):351–352.
- Lipsitch M, Tchetgen Tchetgen E, Cohen T. Re: “Negative control exposures in epidemiologic studies” [letter]. *Epidemiology*. 2012;23(2):351–352.

APPENDIX

Extension to case-control design

Case-control studies are quite common in epidemiologic practice, and COCA extends to this context but requires some modification to appropriately account for the study design. Thus, suppose cases (with $Y = 1$) and controls (with $Y = 0$) are obtained in a population with rare disease rate, and let S denote the indicator of selection into the case-control sample. The case-control design typically oversamples cases for more cost-effective and statistically efficient inference. We propose to estimate the causal effect of A via case-control COCA for a logistic regression model of the form given by equation 12 in the main text, upon redefining $\Pr(Y_0 = 1|A = a, c)$ as $\Pr(Y_0 = 1|A = a, c, S = 1)$, where

$$\Pr(Y_0 = 1|A = a, c, S = 1) = \begin{cases} \Pr(Y = 1|A = 0, c, S = 1) = \frac{E(Z|A = 0, S = 1, c) - E(Z|A = 0, Y = 0, c)}{E(Z|A = 0, Y = 1, c) - E(Z|A = 0, Y = 0, c)} & \text{if } a = 0 \\ \frac{E(Z|A = 1, Y = 0, c) - E(Z|A = 0, Y = 0, c)}{E(Z|A = 0, Y = 1, c) - E(Z|A = 0, Y = 0, c)} & \text{if } a = 1. \end{cases}$$

The key insight justifying the approach is that $E(Z|A = 1, Y = 0, c)$ approximates $E(Z|A = 1, c)$ under the rare disease assumption, and therefore $\Pr(Y_0 = 1|A = 1, c, S = 1)$ approximates $\Pr(Y_0 = 1|A = a, c)$, which suffices for identification of ψ , and furthermore, this is the case even though $\Pr(Y = 1|A = 0, c, S = 1)$ fails to identify $\Pr(Y = 1|A = 0, c)$. If the disease is not rare in the target population but the sampling fraction for cases and control is known, a straightforward application of inverse-selection-probability-weighting COCA estimation may be used to recover the correct population inference.

Sensitivity analysis for an imperfect negative control

Heretofore, we have assumed a perfect negative control outcome is available, such that assumption 1 holds exactly in the observed data. We now propose to relax this assumption, in order to allow for the possibility that \mathcal{Y}_A may not fully account for unobserved confounding between A and Z . This could happen, say, if there was an unobserved common cause of A and Z that does not also confound the relation between A and Y . If this were the case, COCA as developed in previous sections would fail to unbiasedly estimate the causal effect of A on Y , even if all fitted models are correctly specified. To address this potential issue, a sensitivity analysis approach is proposed, which may be used to assess the extent to which inference about the causal effect of A on Y may be altered by a violation of assumption 1.

To describe the sensitivity analysis technique, suppose that Y , A , and Z are continuous, and to simplify the exposition, suppose that there are no covariates (i.e., C is the empty set). Furthermore, we shall suppose that the following linear models generated the observed data:

$$\begin{aligned} A &= \alpha_0 + \alpha_1 Y_0 + \Delta \\ Y &= Y_0 + \psi_0 A \\ Z &= \beta_0 + \beta_1 Y_0 + \kappa, \end{aligned}$$

where Δ and κ are mean 0 error terms, uncorrelated with Y_0 . Then, if assumption 1 holds, we have that κ and Δ are independent, and therefore $E(\kappa\Delta) = 0$. To encode a violation of assumption 1, we set

$$\kappa = \rho\Delta + \chi,$$

where χ is an independent error term, and ρ is a sensitivity parameter that encodes the magnitude of unobserved confounding for the association between A and Z upon adjustment for Y_0 . To implement the sensitivity analysis requires an estimate of $\Delta = \{A - E(A|Y_0)\}$. For fixed ψ , let $\hat{\Delta}(\psi)$ denote the OLS residual from regressing A on $Y(\psi)$ using a simple linear regression. Define $\hat{\psi}(\rho)$ as the midpoint of the 95% confidence interval corresponding to values of ψ such that the null hypothesis of $\beta_2(\rho, \psi) = 0$ fails to reject at the 0.05 α level in the following regression model:

$$Z = \beta_0 + \beta_1 Y(\psi) + \beta_2 A + \rho\hat{\Delta} + \chi \quad (1)$$

with (ρ, ψ) fixed and $(\beta_0(\rho, \psi), \beta_1(\rho, \psi), \beta_2(\rho, \psi))$ estimated by OLS of Z on A with an offset equal to $\rho\hat{\Delta}$. A sensitivity analysis is then obtained by repeating the above steps for different values of ρ on an interval containing $\rho = 0$ (which recovers the analysis obtained under assumption 1). Although we have motivated the sensitivity analysis technique assuming continuous A , the approach equally applies for binary A , upon replacing linear regression with binary regression, such as, say, logistic regression, to fit $E(A|Y(\psi))$ and to construct $\Delta(\psi) = A - E(A|Y(\psi))$. We also note that the parametric models used above were specified primarily to simplify the exposition, and it is possible to more formally motivate the parametrization for the linear regression (1) using nonparametric arguments along the lines of Robins et al. (7), allowing for a more general functional form for the models and also incorporating covariates.