Harvard University

Harvard University Biostatistics Working Paper Series

Year 2015 *Paper* 194

On Simple Relations Between Difference-in-differences and Negative Outcome Control of Unobserved Confounding

Tamar Sofer* David B. Richardson† Elena Colincino‡

Joel Schwartz** Eric J. Tchetgen Tchetgen††

*University of Washington, tsofer@uw.edu

†University of North Carolina at Chapel Hill, david.richardson@unc.edu

‡Harvard T.H. Chan School of Public Health, ecolicin@hsph.harvard.edu

**Harvard T.H. Chan School of Public Health, jschwrtz@hsph.harvard.edu

††Harvard T.H. Chan School of Public Health, etchetge@hsph.harvard.edu

(2015), **0**, 0, *pp.* 1–29
doi:awesomedoi/20150801˙TSofer˙DID

On simple relations between difference-in-differences and negative outcome control of unobserved confounding

TAMAR SOFER[1], DAVID RICHARDSON[2], ELENA COLICINO[3], JOEL SCHWARTZ[3], ERIC J. TCHETGEN TCHETGEN[4*]

[1]*Department of Biostatistics, University of Washington, UW Tower, 15th Floor, 4333 Brooklyn Ave. NE, Seattle, WA 98105, USA*
[2]*Department of Epidemiology, Gilling School of Global Public Health, 2102b Mcgavran-Greenberg 135 Dauer Drive, Chapel Hill, NC, 27599, USA*
[3]*Department of Environmental Health, Harvard T.H. Chan School of Public Health, 677 Huntington Ave, Boston, MA 02115, USA*

[4]*Department of Biostatistics, Harvard T.H. Chan School of Public Health, 677 Huntington Ave, Boston, MA 02115, USA*

etchetge@hsph.harvard.edu

Summary

The *difference-in-differences* (DID) approach is a well known strategy for estimating the effect of an exposure in the presence of unobserved confounding. The approach is most commonly used when pre- and post-exposure outcome measurements are available, and one can assume that the association of the unobserved confounder with the outcome is equal in the two exposure groups, and constant over time. Then, one recovers the treatment effect by regressing the change in outcome over time on the exposure. In this paper, we interpret the difference-in-differences as a negative outcome control (NOC) approach. We show that the pre-exposure outcome is a negative control

*To whom correspondence should be addressed.

outcome, as it cannot be influenced by the subsequent exposure, and it is affected by both observed and unobserved confounders of the exposure-outcome association of interest. The relation between DID and NOC provides simple conditions under which negative control outcomes can be used to detect and correct for confounding bias. However, for general negative control outcomes, the DID-like assumption may be overly restrictive and rarely credible, because it requires that both the outcome of interest and the control outcome are measured on the same scale. Thus, we present a scale-invariant generalization of the DID that may be used in broader NOC contexts. The proposed approach is demonstrated on a Normative Aging Study data set, in which Body Mass Index is used for NOC of the relationship between air pollution and inflammatory outcomes.

Key words: Location-scale models; Quantile-quantile transformation; Air pollution; Inflammation.

1. Introduction

Unmeasured confounding can seldom be ruled out in nonexperimental studies. Over the years, a number of analytic techniques were developed in epidemiology and the social sciences to detect and ideally, adjust for, bias due to unobserved confounding. One common approach is so-called "difference-in-differences" (DID) estimation (Meyer, 1995; Angrist and Krueger, 1999; Blundell and MaCurdy, 2000), which is typically used when

(i) One has observed the outcome pre- and post-exposure for each person, and

(ii) the association of the unobserved confounder with the outcome is assumed equal across exposure groups and constant over time.

Then, the approach entails estimating the effect of exposure by taking a difference between exposure groups of the average change in outcome over time.

Another approach for evaluating the presence of confounding bias, sometimes used in epidemiologic practice, consists of estimating an association between the exposure and a so-called negative control outcome. That is, an observed outcome not causally related to the treatment, and influenced by unmeasured confounders of the exposure-outcome relationship of primary interest (Lipsitch *and others*, 2010; Tchetgen Tchetgen, 2014; Flanders *and others*, 2011). Thus, evidence of an association between the exposure and the negative control outcome conditional on observed confounders is indicative of confounding by unmeasured factors. It is of interest to identify conditions under which the exposure-negative control outcome association gives a valid estimate of unmeasured confounding bias that can simply be removed (e.g. subtracted) from the estimated exposure-outcome association to give a valid causal effect estimate.

In this paper, we interpret the DID as a negative outcome control (NOC) approach to adjust for unobserved confounding. The equivalence follows from noting that the pre-exposure outcome in DID is an ideal negative control outcome, since it cannot be influenced by the subsequent exposure, and it is likely affected by both measured and unobserved risk factors for the post-exposure outcome. We then show that assumption (ii) is equivalent to an "additive equi-confounding" assumption that the magnitude of confounding bias for the primary outcome is equal on the additive scale to the confounding bias for the negative control outcome. Assumptions (i) and (ii) are equivalent to conditions under which one can use negative controls to detect − and also sometimes to correct for − confounding bias. However, the additive equi-confounding assumption may be overly restrictive outside of the context of pre- and post-outcome measurements, because it requires that both the primary and negative control outcomes are measured on the same scale. As a remedy, we consider a generalization of DID via a scale invariant approach largely motivated by the Change-in-changes approach of Athey and Imbens (2006), that may be more broadly applicable. Our approach however goes beyond Athey and Imbens (2006) in that we give weaker identification conditions and develop

a flexible framework for estimation and inference using a familiar location-scale model specification which allows one to easily incorporate a possibly large number of observed confounders. Both the scale-invariance property of the more general approach and its ability to incorporate covariates make our methods particularly well-suited for NOC. Importantly, while Athey and Imbens (2006) briefly consider covariate adjustment, they rely on an assumption that the unobserved confounder is independent of observed covariates conditional on the exposure. However, due to collider bias stratification (Pearl, 2009; Hernán *and others*, 2004), this latter assumption cannot hold if both observed and unobserved covariates either cause or share a common cause with the exposure, thus invalidating their proposed covariate adjustment approach when the observed covariates are confounders. Our proposed approach also offers an alternative to the control outcome calibration approach (COCA) of Tchetgen Tchetgen (2014) by avoiding the rank preservation assumption it relies on, and replacing it with milder assumptions regarding a negative control outcome.

The paper is organized as follows. In Section 2 we present the NOC framework and relate it to the DID. In Section 3 we show how negative outcomes potentially can be used in broader settings than the classical DID, and develop a general NOC approach to indirectly account for unobserved confounding, together with a framework for inference under a location-scale model. In Section 4 we provide a simulation study of the proposed methods, and in Section 5 we illustrate the method by estimating the short term effect of air pollution on blood inflammation markers, with Body Mass Index (BMI) used as a negative outcome.

2. Notation, definitions and additive equi-confounding

Let A denote the exposure received by an individual, let Y denote a post-exposure outcome, and let C denote a set of observed confounding variables of the effect of A on Y. Let U denote unmeasured confounders of the effect of A on Y. Let N denote a negative control outcome variable. The

relationships between these variable may be depicted by the causal diagram in Figure 1.

As shown in the figure, N is a negative control outcome because it is not directly influenced by exposure, but it is influenced by the unobserved confounders of the exposure-outcome association (Lipsitch *and others*, 2010). To provide identifiability conditions for the causal effect of A on Y, we now consider counterfactuals or potential outcomes under possible interventions. Let Y_a denote an individual's outcome if exposure A were set, possibly contrary to fact, to a. Similarly, let N_a denote an individual's counterfactual value for N if A were set to a. The following assumptions state that the negative outcome is not affected by the exposure, and that the observed exposure outcome corresponds to the counterfactual outcome for the observed exposure value (i.e. the so-called consistency assumption).

Assumption 1. *$N_a = N$, $a = 0, 1$, and $Y_a = Y$ if $A = a$.*

The assumption that (C, U) suffice to adjust for confounding for the effect of A on Y implies that:

$$E\{Y_0|A = 1, c, u\} - E\{Y_0|A = 0, c, u\} = 0 \tag{2.1}$$

for all (c, u), however C alone may not completely account for exposure-outcome confounding, that is

$$E\{Y_0|A = 1, c\} - E\{Y_0|A = 0, c\} \neq 0 \tag{2.2}$$

for some c, and likewise for N_0 replacing Y_0.

2.1 *Difference-in-differences as an additive negative outcome control approach*

Next, consider the longitudinal study represented in Figure 2 in which the outcome process $Z(t)$ is measured at 2 occasions, $t = 0, 1$, with $Z(0)$ and $Z(1)$ pre- and post-exposure variables, respectively. According to this graph, although A is a cause of $Z(1)$, it does not cause $Z(0)$ (although the reverse

may hold), and the unobserved confounder of the effect of A on $Z(1)$, U, is also a cause of $Z(0)$. This causal diagram represents a typical situation under which difference-in-differences may potentially be used to account for unobserved confounding by U. However an additional assumption about the underlying structure of confounding is required to justify the standard DID approach, and is described below. The similarity of the causal structure encoded in both Figures 1 and 2 is quite striking, as Figure 1 can be obtained from Figure 2 by relabeling $Z(0)$ as N and $Z(1)$ as Y, thus establishing a direct correspondence between the NOC causal framework and the DID framework. As noted above, identification of the effect of A on Y using DID, relies on further elaboration of the data generating mechanism under Figure 1. A simple causal model supposes that $Z(t)$ follows the simple linear model (where individual observations are suppressed in the notation)

$$E\left\{Z(t)|U,A,C\right\} = b(U) + m\left(t\right) + \beta tA + \gamma\left(t\right)^T C \tag{2.3}$$

such that $m\left(t\right)$ indexes a time specific intercept, $\gamma\left(t\right)$ indexes a time specific association between C and $Z(t)$, $b(U)$ indexes the effect of U on $Z(t)$ which is assumed independent of t, A and C, and β encodes the causal effect of A on $Z(1)$. Let $Z_a(t)$ denote the counterfactual outcome at t under exposure a, and note that the key assumption encoded in equation (2.3) is that

$$E\left\{Z(1) - Z(0)|U, A = a, C\right\} = E\left\{Z(1) - Z(0)|A = a, C\right\}, \ a = 0, 1 \tag{2.4}$$

which implies that C suffices to adjust for confounding between A and $Z(1) - Z(0)$, and thus

$$E\left\{Z_0(1) - Z_0(0)|A = 1, C\right\} = E\left\{Z_0(1) - Z_0(0)|A = 0, C\right\}. \tag{2.5}$$

Since treatment is assumed to start only after time 0, so that $E[Z_1(0)|A = 1, C] = E[Z_0(0)|A = 1, C]$, and using equation 2.5, we obtain the following equality:

$$E\left\{Z_1(1) - Z_1(0)|A = 1, C\right\} - E\left\{Z_0(1) - Z_0(0)|A = 0, C\right\} = \beta \tag{2.6}$$

$$= E(Z_1(1) - Z_0(1)|A = 1, C) \tag{2.7}$$

The effect identified in (2.6) defines the DID estimand under equation (2.3), and therefore under assumption (2.4) is equal to the causal effect $E\{Z_1(1) - Z_0(1)|A = 1\}$, the so-called causal effect of treatment on the treated. Interestingly, rather than assuming equation (2.3) , one may take equation (2.5) as a primitive condition, which may hold without necessarily assuming the model given by equation (2.3) holds exactly. Only assuming that (2.5) holds has previously been shown to suffice for nonparametric identification of the marginal exposure effect on the exposed even when the linear model (2.3) does not necessarily hold (Abadie, 2005). Thus, assuming no heterogeneity in the effect of A across strata of C and U as encoded in model (2.3) is not strictly necessary to estimate the causal effect of treatment on the treated.

2.2 *Additive equi-confounding bias*

Here, we are particularly interested in the following, alternative, formulation of (2.5):

$$E\{Z_0(1)|A = 1, C\} - E\{Z_0(1)\ |A = 0, C\} = E\{Z(0)|A = 1, C\} - E\{Z(0)|A = 0, C\}$$

which, upon substituting Y_0 for $Z_0(1)$ and N for $Z(0)$, is equivalently expressed:

$$E\{Y_0|A = 1, C\} - E\{Y_0|A = 0, C\} = E\{N|A = 1, C\} - E\{N\ |A = 0, C\}, \tag{2.8}$$

where the left hand-side of (2.8) encodes the degree of confounding bias (2.2) for the effect of A on Y, and the right hand-side of (2.8) likewise represents confounding bias for the (null) effect of A on N. Equation (2.8) provides the "additive equi-confounding" assumption, which connects identification in the DID approach to identification in the NOC framework.

The additive equi-confounding assumption 2.8 thus states that the magnitude of confounding bias for estimating the effect of A on Y and that of A on N are exactly equal. Thus, we may conclude that under additive equi-confounding, a DID type approach may be used to estimate, in the presence of unobserved confounding and if one has access to a negative outcome control

variable N (which may differ from a pre-exposure realization of the outcome), the conditional effect of treatment on the treated:

$$\alpha(C) = E\{Y_1 - Y_0 | A = 1, C\}$$

or the marginal average effect of treatment on the treated:

$$\alpha = E\{Y_1 - Y_0 | A = 1\}$$

Therefore, the additive equi-confounding assumption formalizes the relation between DID and NOC, making connection to a fairly rich literature on DID for inference under a NOC framework. The DID literature includes several variants of the parametric strategy described above, as well as more flexible semiparametric methods (see Abadie, 2005, and references therein). However, the additive equi-confounding assumption may only be credible in settings where the primary and the negative control outcomes are measured on the same scale, say as distinct realizations of the same underlying process as in the difference-in-differences context. This restriction is well illustrated by the linear model (2.3) in which the invariance of $b(U)$ with respect to time encodes the equivalent assumption for a negative outcome control, that the association between U and the primary outcome is the same as that between U and the negative control outcome. Such an assumption may be inappropriate even if one has available a valid negative control outcome which satisfies Figure 1. In the next section, we consider a weaker form of equi-confounding which may be more useful in practice for NOC.

3. Distributional equi-confounding and indirect NOC confounding adjustment

In this section, we consider a more general framework for NOC adjustment of unobserved confounding under assumptions considerably less restrictive than additive equi-confounding.

3.1 *General NOC identification conditions*

We relax the previous structure of unobserved confounding for Y and N, by allowing the unobserved confounder for the effect of A on Y denoted by U, to be distinct from the unobserved confounder of the effect of A on N, denoted W. This may be because Y and N are measured on different scales, or because the magnitude of the effect of the unobserved confounder on N differs from the magnitude of the effect of that confounder on Y.

Assumption 2. *$A \amalg Y_0 | C, U$, however $A \not\amalg Y_0 | C$, and*

$A \amalg N | C, W$, however $A \not\amalg N | C$.

This more general framework is depicted in Figure 3. In addition to this causal diagram, in order to appropriately account for possible non-linearity and scale differences between the outcome and the negative control outcome, we introduce a more general nonparametric structural equations model:

Assumption 3. *Y_0 and N are related to U, W and C according to*

$$Y_0 = h_y(U, C) \tag{3.9a}$$

$$N = h_n(W, C) \tag{3.9b}$$

where

$$h_y(u, c) \text{ is monotone increasing in } u \text{ for each } c, \tag{3.10a}$$

and

$$h_n(w, c) \text{ is monotone increasing in } w \text{ for each } c. \tag{3.10b}$$

This set of equations encodes the fact that consistent with Figure 3, U and C are parents of Y_0, and therefore are parents of Y, and likewise that W and C are parents of N. The direction of monotonicity in equations (3.10a,3.10b) can be changed without any real consequence.

We now consider quantile-quantile and distributional equi-confounding as a less restrictive identifying assumptions for NOC than additive equi-confounding. To proceed, we introduce the quantile-quantile transformation, as a measure of association between two variables, which we will use to encode confounding bias. Specifically, the quantile-quantile association between U and A conditional on C :

$$q_0(v|c) = F_{U|A=0,C=c} \circ F^{-1}_{U|A=1,C=c}(v),$$

$v \in [0,1]$, where $F_{U|A,C}$ denotes the cumulative distribution function of U conditional on A, C, $F^{-1}_{U|A,C}$ is its inverse map, and $f \circ g(x) = f(g(x))$ denotes composition of functions f and g. Under independence of U and A given C (i.e. no confounding bias), we have that $q_0(v|c) = v$, while any departure from the identity function encodes unobserved confounding, i.e. $q_0(v|c) - v \neq 0$ for some value c. Likewise let

$$q_1(v|c) = F_{W|A=0,C=c} \circ F^{-1}_{W|A=1,C=c}(v).$$

The quantile-quantile equi-confounding bias is given below.

Assumption 4. *Quantile-quantile equi-confounding.*

$$q_0(v|c) = q_1(v|c), \qquad v \in [0,1]. \tag{3.11}$$

This assumption implies that the association (on the quantile-quantile scale) between U and A is the same as between W and A conditional on C. Quantile-quantile equi-confounding is implied by the following somewhat stronger distributional equi-confounding bias assumption, although the latter is still considerably weaker than additive equi-confounding:

Assumption 5. *Distributional equi-confounding.*

$$U|A, C \sim W|A, C. \tag{3.12}$$

The assumption states that the conditional distribution of the unobserved confounder for Y is the same as that for the unobserved confounder for N given A and C. Note that both assumption

(3.11) and (3.12) are trivially satisfied, if as previously assumed, the unobserved confounder of Y and N is the same, i.e. $U = W$. Note also that both assumptions are considerably weaker than the previous additive equi-confounding assumption (2.8) because they place no restriction on the relationship between U and Y_0, and likewise for the relationship between W and N. Crucially, they are both invariant in a monotone transformation of the outcome, and therefore, do not suffer from the scale restriction of additive equi-confounding.

The following Theorem 1 establishes nonparametric identification of the marginal effect of treatment on the treated α under quantile-quantile equi-confounding, and therefore also under distributional equi-confounding. It requires an additional regularity condition:

Assumption 6. *Positivity.*

$$\textit{If } \; 0 < f(N^*|A=1,C), \; \textit{then } 0 < F_{N|A=0,C}(N^*) < 1 \tag{3.13}$$

This condition ensures that values of the negative outcome in the exposed are in the support of the distribution of the negative outcome in the unexposed, and the probability $F_{N|A=C}(N^*)$ will not be identically 1 or 0 for some set of plausible values of $N^* \sim N|A=1,C$.

Theorem 1 : *Let $N^* \sim N|A=1,C$. Under assumptions 1-4 and 6, we have that*

$$\begin{aligned}\alpha &= E\left\{Y_1 - Y_0|A=1\right\} \\ &= E\left\{Y|A=1\right\} - E\left\{\widetilde{Y}\right\}\end{aligned}$$

where

$$\widetilde{Y} \sim F^{-1}_{Y|A=0,C} \circ F_{N|A=0,C}\left(N^*\right)$$

By consistency, we have that $E(Y_1|A=1) = E(Y|A=1)$; therefore the main result of the theorem is to establish identification of the conditional counterfactual mean $E(Y_0|A=1)$ under our assumptions about the nature of confounding and the availability of a negative control outcome.

The theorem is a negative control analog of a similar identification result in the change-in-changes approach of Athey and Imbens (2006), which they obtain under a more stringent assumption analogous to distributional equi-confounding. Whereas Athey and Imbens (2006)'s primary goal was to account for possible non-linearity in a DID context, our primary concern has been to account for possible differential scaling in a NOC context, and to demonstrate the close relationship between these contexts as established by the above result. The isomorphism between the two frameworks further provides a principled framework for NOC of unobserved confounding, possibly using a post-exposure outcome to achieve such control. The result also offers a useful alternative to COCA of Tchetgen Tchetgen (2014) which requires rank preservation of the primary outcome, i.e. that the rank of Y_a is preserved under treatment versus control conditions.

3.2 *Indirect NOC adjustment in the location-scale model*

For inference, we discuss indirect adjustment under a location-scale semiparametric model. Specifically, suppose that both Y and N follow a location-scale model conditional on C in the unexposed, with $A = 0$. Let

$$E(N|A=0,C) = \mu_n(C)$$
$$Var(N|A=0,C) = s_n^2(C),$$
$$\text{and } \varepsilon_N = \frac{N - \mu_n(C)}{s_n(C)},$$

and likewise, let

$$E(Y|A=0,C) = \mu_y(C)$$
$$Var(Y|A=0,C) = s_y^2(C)$$
$$\text{and } \varepsilon_Y = \frac{Y - \mu_y(C)}{s_y(C)}.$$

Then the location-scale models for Y and N states that

$$\varepsilon_N|A=0,C\sim f_N(\varepsilon_N),\qquad \varepsilon_Y|A=0,C\sim f_Y(\varepsilon_Y), \tag{3.14}$$

where $f_Y(\cdot)$ and $f_N(\cdot)$ are unrestricted baseline densities. The following Corollary is obtained:

Corollary 1 *Under the assumptions stated in Theorem 1 and the location-scale model (3.14) , we have that*

$$\widetilde{Y}=s_Y\left(C\right)\left\{F_{\epsilon_Y}^{-1}\circ F_{\epsilon_N}\left(\frac{N^*-\mu_n\left(C\right)}{s_n\left(C\right)}\right)\right\}+\mu_y\left(C\right), \tag{3.15}$$

and in the special case where $F_{\epsilon_N}(\cdot)=F_{\epsilon_Y}(\cdot)$, then

$$\widetilde{Y}=s_y\left(C\right)\left\{\frac{N^*-\mu_n\left(C\right)}{s_n\left(C\right)}\right\}+\mu_y\left(C\right), \tag{3.16}$$

where N^* and $\tilde{Y}$ are given in Theorem 1.

Note also that if $F_{\epsilon_N}(\cdot)=F_{\epsilon_Y}(\cdot)$, then the regularity condition 6 is not strictly required. Next, we describe a simple practical implementation of the NOC adjustment given in Corollary 1, first assuming a location-scale family allowing $F_N(\cdot)$ and $F_Y(\cdot)$ to be different, and then further assuming $F_N(\cdot)=F_Y(\cdot)$.

Let $\widehat{\mu}_n(\cdot),\widehat{\mu}_y(\cdot)$ be estimators of the mean functions for the negative and primary outcomes under no exposure, and let $\widehat{s}_n(\cdot),\widehat{s}_y(\cdot)$ denote estimators of the standard deviations of N and Y. These can be obtained using standard models for mean and variance regression, e.g. one may take $\widehat{\mu}_n\left(C\right)=\widehat{\pi}_0+\widehat{\pi}_1'C$ the ordinary least squares estimator of $E(N|A=0,C)$ using the subsample with $A=0$, and likewise one may take $\widehat{s}_n^2\left(C\right)=\exp\left(\widehat{\omega}_0+\widehat{\omega}_1'C\right)$ a standard log-linear regression of the squared $N-\mu_n(C)$ in the unexposed subsample, and similarly for $\widehat{\mu}_y\left(\cdot\right)$ and $\widehat{s}_y\left(\cdot\right)$. Further, let $\widehat{F}_N(\cdot)$ and $\widehat{F}_Y(\cdot)$ be non-parametric estimators of the cumulative distribution functions of N and Y in the unexposed, estimated as follows: based on $\widehat{\mu}_y(\cdot),\widehat{s}_y(\cdot)$, scaled residuals are obtained as $\widehat{\epsilon}_i=\{Y_i-\widehat{\mu}_y(C_i)\}/s_y(C_i), i=1,\ldots,n_0$ for the outcome Y, in the n_0 unexposed. These residuals

are used to obtain the empirical estimate of $F_Y(\cdot)$ non-parametrically,

$$\widehat{F}_Y(u) = \frac{1}{n_0}\sum_{i=1}^{n_0} \mathbb{I}\{\epsilon_i \leqslant u\},$$

where $\mathbb{I}(\epsilon_i \leqslant u)$ is the indicator function. Similarly, $\widehat{F}_N(\cdot)$ is obtained for the negative outcome N.

1. Following Theorem 1 and Corollary 1, an estimator of α is obtained by substitution, i.e.

$$\widehat{\alpha}_1 = \frac{\sum_{i:A_i=1} Y_i - \widehat{s}_y\left(C_i\right)\left\{\widehat{F}^{-1}_{\epsilon_Y} \circ \widehat{F}_{\epsilon_N}\left(\frac{N^* - \widehat{\mu}_n(C_i)}{\widehat{s}_n(C_i)}\right)\right\} - \widehat{\mu}_y\left(C_i\right)}{\sum_{i:A_i=1} 1}. \tag{3.17}$$

2. Assuming $F_N(\cdot) = F_Y(\cdot)$, (3.17) simplifies to:

$$\widehat{\alpha}_2 = \frac{\sum_{i:A_i=1} Y_i - \widehat{s}_y\left(C_i\right)\left(\frac{N^* - \widehat{\mu}_n(C_i)}{\widehat{s}_n(C_i)}\right) - \widehat{\mu}_y\left(C_i\right)}{\sum_{i:A_i=1} 1}. \tag{3.18}$$

3. Under homoscedasticity, i.e. $\widehat{s}_y(C_i) = \widehat{s}_y$ for all C_i, and similarly for $\widehat{s}_n$, upon rearranging:

$$\begin{aligned}\widehat{\alpha}_3 &= \frac{\sum_{i:A_i=1}\left[Y_i - \widehat{\mu}_y\left(C_i\right)\right]}{\sum_{i:A_i=1} 1} - \frac{\widehat{s}_y}{\widehat{s}_n}\frac{\left[N^* - \widehat{\mu}_n\left(C_i\right)\right]}{\sum_{i:A_i=1} 1} \\ &= \widehat{\eta}_y - \frac{\widehat{s}_y}{\widehat{s}_n}\widehat{\eta}_n \end{aligned} \tag{3.19}$$

   where $\widehat{\eta}_y$ and $\widehat{\eta}_n$ are the standard estimators of the effect of treatment on the treated for Y and N respectively. This formulation provides some intuition for the proposed indirect adjustment, whereby the standard estimator of the $A-Y$ association is adjusted by subtracting an estimator of the magnitude of confounding bias given by the scaled association between N and A, with scaling factor $\widehat{s}_y/\widehat{s}_n$. The scaling factor is necessary here, to account for possible scale differences between N and Y, or between the magnitude of the effect of the unmeasured confounder on N and Y. The more complicated estimator $\widehat{\alpha}_1$ further accounts for distributional differences and possible heteroscedasticity.

In the appendix, we provide a simple expression for the large sample variance of $\widehat{\alpha}_2$ which may be used to construct confidence intervals; alternatively, we recommend using the nonparametric bootstrap for inference.

4. Simulation study

We conducted a simulation study to demonstrate the applicability of our proposed indirect NOC adjustment under a location-scale model. We generated data from the model defined by

$$Y = (U + \eta_0 + C\eta_c + A\tilde{\alpha}) \times \sigma_y$$

$$N = (W + \beta_0 + C\beta_c) \times \sigma_n$$

with U and W from the same location-scale family. We set $\sigma_y = 3, \sigma_n = 1.5$, $(\eta_0, \eta_c)^T = (1, 2)^T$, $(\beta_0, \beta_c)^T = (2, 3)^T$, and $\tilde{\alpha} = 1$, so the exposure effect on the unexposed amounted to $\alpha = \tilde{\alpha} \times \sigma_y = 3$. To simulated confounding bias between exposure groups, we determined the distribution of C, U and W by exposure status. U and W came from either a normal or a uniform distribution, with $U, W|A = 0 \sim \mathcal{N}(0, 1.5)$, and $U, W|A = 1 \sim \mathcal{N}(2, 1.5)$, or $U, W|A = 0 \sim$ uniform$(1, 9)$ and $U, W|A = 1 \sim$ uniform$(3, 13)$. The observed confounder had $C|A = 0 \sim \mathcal{N}(0, 1), C|A = 1 \sim \mathcal{N}(0.5, 1)$.

Note that a naïve analysis ignoring the possibility of unmeasured confounding between exposure groups would attribute the difference in means

$$E[Y|A = 1, C] - E[Y|A = 0, C] = \alpha + (E[U|A = 1] - E[U|A = 0]) \times \sigma_y$$

solely to the effect of treatment, when the term $(E[U|A = 1] - E[U|A = 0]) \times \sigma_y$ is in fact the bias, and is equal to 6 when U and W are normally distributed, and 9 when they are uniformly distributed. Also note that under the uniform distribution scenario, the positivity assumption 6 does not hold, and therefore the estimator α_1 from Section 3.2 may be biased. However the estimator α_2 that assumes $F_N(\cdot) = F_Y(\cdot)$ will not be biased, since in this case the positivity assumption is not required.

We generated data with $n = 100,500$ observations, with $n/2$ observations in each exposure group. We compared the accuracy of the estimators proposed in Section 3.2 over 1000 simulations.

Note that although both outcomes Y and N are generated under homoscedastic errors, with U and W following a common distribution given A and C, nonetheless, we consider inferences about the effect of treatment on the treated using the methods developed in previous sections both with and without imposing these assumptions. In addition, we compare the estimator of α using NOC to the naïve regression estimator that regresses Y on A and C.

Table 1 provides the absolute bias and MSE of the estimator of treatment effect on the treated for each of the various scenarios and assumptions described above. Using N for negative outcome control assuming a location-scale model yields very good results. The data were simulated with homoscedastic errors and a common location-scale family for Y and N, so that the qq-transformation between the standardized Y and N in the unexposed group is the identity. Accordingly, when homoscedasticity and identity qq-transformations were assumed, the estimated effects are unbiased and the MSE is smallest compared to other scenarios. Relaxing the homoscedasticity assumption and modeling the variance via a log-linear model resulted in only slightly larger MSEs. However, modeling the qq-transformation between the standardized Y and N in the unexposed group had mixed effects. Under normal distribution of the unobserved confounders, modeling the qq-transformation had little effect on the bias and efficiency of the estimators. However, under uniform distribution of the unmeasured confounders, modeling this transformation resulted in substantially larger MSEs and biased estimators. This is because the positivity condition did not hold. The naïve estimator that regresses Y against A and C showed had the expected bias.

5. Data analysis

We implemented the proposed NOC indirect adjustment to account for confounding in studying the effect of short term (4 weeks) exposure to black carbon (BC, an air pollution component) on fibrinogen, a blood inflammation marker. We selected BMI as the negative control outcome, since

BMI is likely not affected by short term exposure to air pollution, while it likely shares unmeasured confounders with inflammation markers. In prior work by Zeka *and others* (2006), fibrinogen levels were shown to be associated with 4 weeks exposure to BC in the Normative Aging Study (NAS) cohort. The investigators took 4 weeks moving average of BC, measured at an areal sensor, just prior to a clinic visit as the exposure, and adjusted to multiple confounders, including BMI. We now reuse this data set.

The NAS is a longitudinal study following a cohort of US veterans. They report to the clinic every 3-4 years. We consider a data set of complete cases (i.e. includes exposure, adjusting variables, and outcome values) from visits between November 14, 2000 and December 31, 2004, as in Zeka *and others* (2006). We use BC values measured either at the areal sensor in Boston (as in Zeka *and others* (2006)), or geospatial model-predicted values at participants' home addresses (Gryparis *and others*, 2007). The covariates were age and weather-related variables: season, mean barometric pressure, relative humidity, and temperature in the 24 hours preceding the clinic visit. Table 3 provides the cohort characteristics. BC is dichotomized and set to 0 if BC is less than the median observed in the data ("low exposure"), and 1 otherwise ("high exposure") .We implemented the four models compared in the simulations, i.e. the more robust models allowing for heteroscedasticity, and/or different location-scale family, and the model that assumes homoscedasticity and same location-scale family. In addition to these models of indirect adjustment, we also compared the analysis to the a naïve regression analysis that regresses log-fibrinogen on the BC measure of interest, covariates, and BMI, and to the "DID" analysis that assumes that the negative control outcome log-BMI is measured on the same scale as the primary outcome log-fibrinogen. Note that this is not in fact the standard DID, as log-BMI is obviously not the baseline measure of log-fibrinogen.

We evaluated the assumptions on the distributions of Y and N. To evaluate the assumption of a common location-scale family of Y and N, we considered the histograms of scaled residuals

of BMI, fibrinogens, and their log-transformation in the low-exposure group, after regressing on covariates. These histograms are provided in Figure 7. One can see that after log-transformation both the primary and control outcomes have symmetric distributions, and it is reasonable to assume that they are sampled from the same location-scale family. We also observed that log-fibrinogen and log-BMI are measured on different scales. Next, we assessed the homoscedasticity assumption on the residuals of Y and N in the low exposure group. We used a 5-fold cross validation of the restricted data set, where in each "fold" we took a fourth of the participants, and generated mean and variance models to predict the outcomes (log-fibrinogen) of the held-out fifth of the participants. We calculated the mean squared errors for these predictions as $\sum_{i=1}^{n_k}((y_i - C\widehat{\boldsymbol{\beta}}_y)^2 - \exp(C\widehat{\boldsymbol{\omega}}_y))^2$, where n_k is the number of observation in the $k = 1, \ldots, 5$ set of observations, $\widehat{\boldsymbol{\beta}}_y$ is the vector of regression coefficients of the outcome y, and $\boldsymbol{\omega}$ is the vector of regression coefficients in the log-linear models of the residuals. The cross-validated prediction score is the mean of these 5 scores. Table 2 provides the results of these cross validations, and they suggest that modeling the variances of both Y and N conditional on covariates is beneficial.

Figure 5 provides effect estimates using the various models described above, and their 95% bootstrap confidence intervals from 1000 bootstrap samples. One can see that when using more robust models (that make fewer assumptions), the confidence intervals are wider, in agreement with the simulations studies, in which MSEs were larger for more general models. Consider the dichotomized (high vs low) BC exposure measured at an areal sensor. For this model, based on the histograms in Figure 4 and the results from assesing heteroscedasticity in Table 2, the most appropriate estimates assumes that Y and N come from the same location-scale family ("same LS" in the figure) and, the variance varies within levels of covariates ("var(C)" in the figure). Interestingly, in this case the effect estimates of BC are larger than the standard regression estimate.

The "DID" analysis had hardly any impact on the results compared to the ordinary regression

analysis, since log-BMI is measured on a different scale than log-fibrinogen, and more accurately - in values much closer to zero. This demonstrates the importance of accounting for the outcome's scale in DID-type analysis. More generally, even if the negative outcome is the same as the primary outcome, there may be difference in variances across groups that are important to account for.

Interestingly, when using the predicted BC measures at the participants' home addresses, BC effect estimates are closer to null. This is likely due to measurement error from the geospatial model used to predict the BC measurements. Such models were shown to often lead to biases towards the null in estimating air pollution effects (Zeger *and others*, 2000).

In contrast with standard regression, estimates based on NOC approaches under different location-scale families found no significant exposure effect; however, confidence intervals from all models contained the point estimate obtained using standard regression, suggesting that BMI does not provide any significant evidence of unobserved confounding bias.

6. Software

Software in the form of R code, together with a sample input data set and complete documentation is available on request from the corresponding author (etchetge@hsph.harvard.edu).

7. Supplementary Material

Acknowledgments

The authors thank... *Conflict of Interest*: None declared.

References

Abadie, Alberto. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies* **72**(1), 1–19.

Angrist, Joshua D and Krueger, Alan B. (1999). Empirical strategies in labor economics. *Handbook of labor economics* **3**, 1277–1366.

Athey, Susan and Imbens, Guido W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica* **74**(2), 431–497.

Blundell, Richard and MaCurdy, Thomas. (2000). Labor supply. In: Ashenfelter, O. and D. Card, eds. (editors), *Handbook of Labor Economics*. North Holland: Elsevier, pp. 1559–1695.

Flanders, W Dana, Klein, Mitchel, Darrow, Lyndsey A, Strickland, Matthew J, Sarnat, Stefanie E, Sarnat, Jeremy A, Waller, Lance A, Winquist, Andrea and Tolbert, Paige E. (2011). A method for detection of residual confounding in time-series and other observational studies. *Epidemiology (Cambridge, Mass.)* **22**(1), 59.

Gryparis, Alexandros, Coull, Brent A, Schwartz, Joel and Suh, Helen H. (2007). Semiparametric latent variable regression models for spatiotemporal modelling of mobile source particles in the greater boston area. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **56**(2), 183–209.

Hernán, Miguel A, Hernández-Díaz, Sonia and Robins, James M. (2004). A structural approach to selection bias. *Epidemiology* **15**(5), 615–625.

Lipsitch, Marc, Tchetgen Tchetgen, Eric and Cohen, Ted. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology* **21**(3), 383–388.

Meyer, Breed D. (1995). Natural and quasi-experiments in economics. *Journal of business & economic statistics* **13**(2), 151–161.

Pearl, Judea. (2009). *Causality*. Cambridge university press.

Tchetgen Tchetgen, Eric J. (2014). The control outcome calibration approach for causal inference with unobserved confounding. *American journal of epidemiology* **179**(5), 633–640.

Zeger, Scott L, Thomas, Duncan, Dominici, Francesca, Samet, Jonathan M, Schwartz, Joel, Dockery, Douglas and Cohen, Aaron. (2000). Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environmental health perspectives* **108**(5), 419.

Zeka, Ariana, Sullivan, James R, Vokonas, Pantel S, Sparrow, David and Schwartz, Joel. (2006). Inflammatory markers and particulate air pollution: characterizing the pathway to disease. *International Journal of Epidemiology* **35**(5), 1347–1354.

APPENDIX

A. Mathematical derivations

Proof of Theorem 1: *Let* $S_{N|A,C}(n) = P\{N \geqslant n|A,C\}$ *and* $F_{N|A,C}(n) = P\{N < n|A,C\}$.

First we establish that assumption (3.11) *is equivalent to:*

$$F_{Y_0|A=0,C} \circ F^{-1}_{Y_0|A=1,C}(v) = F_{N|A=0,C} \circ F^{-1}_{N|A=1,C}(v)$$

since

$$\begin{aligned} & F_{Y_0|A=0,C} \circ F^{-1}_{Y_0|A=1,C}(v) \\ &= \Pr\left\{Y_0 \leqslant F^{-1}_{Y_0|A=1,C}(v)\,|C, A=0\right\} \\ &= \Pr\left\{h_y(U,C) \leqslant F^{-1}_{Y_0|A=1,C}(v)\,|C, A=0\right\} \end{aligned}$$

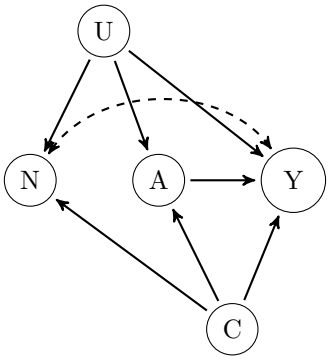

Fig. 1. Directed acyclic graph depicting the causal association between the treatment A, primary outcome Y, negative control outcome N, measured pre-exposure confounders C, and unmeasured confounders U.

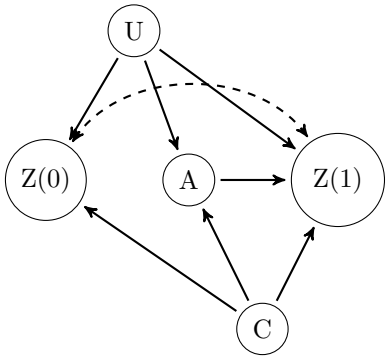

Fig. 2. Directed acyclic graph depicting the causal association between the treatment A, pre-exposure outcome $Z(0)$, post-exposure outcome $Z(1)$, measured pre-exposure confounders C, and unmeasured confounders U.

Table 1. *Finite sample bias and MSE (in parenthesis) averaged over 1000 simulations, for estimating the effect of treatment on the treated ($\alpha = 3$) via indirect adjustment, under the location-scale model. The unmeasured confounders were sampled from either the normal or the uniform family. The Naïve model is the regression estimator of Y on C and A. Other estimators either assume homoscedasticity, or model the variance as a function of covariates C, and either assume that the qq-transformation between the standardized primary outcome Y and the negative control outcome N in the unexposed group is the identity ("QQ identity"), or model this transformation nonparametrically.*

Family	n	Naïve	Assuming homoscedasticity		Modeling the variance	
		regression	QQ identity	QQ modeled	QQ identity	QQ modeled
Normal	100	5.99(36.83)	0.05(02.72)	0.47(02.54)	0.13(02.99)	0.52(02.65)
Normal	500	5.99(36.06)	0.01(00.53)	0.12(00.57)	0.03(00.59)	0.12(00.61)
Uniform	100	9.09(85.15)	0.03(05.65)	2.61(10.06)	0.03(05.98)	2.59(10.03)
Uniform	500	8.97(81.02)	0.01(01.23)	2.34(06.22)	0.03(01.27)	2.31(06.10)

and note that

$$\begin{aligned} F_{Y_0|A=1,C}(y) &= \Pr\{Y_0 \leqslant y|C, A=1\} \\ &= \Pr\{h_y(U,C) \leqslant y|C, A=1\} \\ &= \Pr\left\{U \leqslant h_y^{-1}(y,C)\,|C, A=1\right\} \\ &= F_{U|C,A=1}\left(h_y^{-1}(y,C)\right) \end{aligned}$$

so that

$$F^{-1}_{Y_0|A=1,C}(v) = h_y\left(F^{-1}_{U|C,A=1}(v), C\right)$$

and therefore we may conclude that

$$\begin{aligned} &F_{Y_0|A=0,C} \circ F^{-1}_{Y_0|A=1,C}(v) \\ &= \Pr\left\{h_y(U,C) \leqslant F^{-1}_{Y_0|A=1,C}(v)\,|C, A=0\right\} \\ &= \Pr\left\{h_y(U,C) \leqslant h_y\left(F^{-1}_{U|C,A=1}(v), C\right)|C, A=0\right\} \\ &= \Pr\left\{U \leqslant F^{-1}_{U|C,A=1}(v)\,|C, A=0\right\} \\ &= F_{U|A=0,C} \circ F^{-1}_{U|A=1,C}(v) \end{aligned}$$

Likewise,

$$F_{N|A=0,C} \circ F^{-1}_{N|A=1,C}(v) = F_{W|A=0,C} \circ F^{-1}_{W|A=1,C}(v)$$

proving

$$\begin{aligned} F_{U|A=0,C} \circ F^{-1}_{U|A=1,C}(v) &= F_{W|A=0,C} \circ F^{-1}_{W|A=1,C}(v) \\ &\iff F_{Y_0|A=0,C} \circ F^{-1}_{Y_0|A=1,C}(v) = F_{N|A=0,C} \circ F^{-1}_{N|A=1,C}(v) \end{aligned} \tag{A.1}$$

the result. □

Proof of Corollary 1: From Theorem 1

$$\widetilde{Y} \sim F^{-1}_{Y|A=0,C} \circ F_{N|A=0,C}\left(N^*\right).$$

First, note that

$$\begin{aligned} F_{N|A=0,C}(v) &= p(N < v|A=0,C) = p\left(\frac{N-\mu_n(C)}{s_n(C)} < \frac{v-\mu_n(C)}{s_n(C)}\middle| A=0,C\right) \\ &= F_{\epsilon_N}\left(\frac{v-\mu_n(C)}{s_n(C)}\right). \end{aligned}$$

Second, let $F^{-1}_{Y|A=0,C}(u) = v$, for $0 < u < 1$. The inverse probability function can be defined as $F^{-1}_{Y|A=0,C}(u) = \min\{v : p(Y < v|A=0,C) = u\}$. Then:

$$\begin{aligned} u = F_{Y|A=0,C}(v) &= p(Y < v|A=0,C) = p\left(\frac{Y-\mu_y(C)}{s_y(C)} < \frac{v-\mu_y(C)}{s_y(C)}\middle| A=0,C\right) \\ &= F_{\epsilon_Y}\left(\frac{v-\mu_y(C)}{s_y(C)}\right) \end{aligned}$$

Thus,

$$v = s_y(C)F^{-1}_{\epsilon_Y}(u) + \mu_y(C) = F^{-1}_{Y|A=0,C}(u).$$

Combining the two results, we get:

$$\widetilde{Y} = s_y(C)F^{-1}_{\epsilon_Y} \circ F_{\epsilon_N}\left(\frac{N^*-\mu_n(C)}{s_n(C)}\right) + \mu_y(C).$$

Now, if $F_{\epsilon_Y}(\cdot) = F_{\epsilon_N}(\cdot)$, trivially

$$\tilde{Y} = s_y(C)\left(\frac{N^*-\mu_n(C)}{s_n(C)}\right) + \mu_y(C). \quad \square$$

B. Asymptotic variance of the location-scale NOC estimate

Assume that $s_y(C_i) = s_y,\ s_n(C_i) = s_n$. The estimating equation $U(\boldsymbol{\theta})$ for $\boldsymbol{\theta} = (\boldsymbol{\beta}_y, \boldsymbol{\beta}_n, s_y, s_n, \alpha)$ is given by

$$U(\boldsymbol{\theta}) = \begin{pmatrix} U(\boldsymbol{\beta}_y) \\ U(\boldsymbol{\beta}_n) \\ U(s_y^2; \boldsymbol{\beta}_y) \\ U(s_n^2; \boldsymbol{\beta}_n) \\ U(\alpha) \end{pmatrix}$$

with influence function

$$E\left[\frac{\partial}{\partial\boldsymbol{\theta}}U(\boldsymbol{\theta})\right]^{-1}U(\boldsymbol{\theta})$$

with:

$$\begin{aligned}
U(\boldsymbol{\beta}_y) &= \frac{1}{\sum_{i:A_i=0}1}\sum_{i:A_i=0}C_i\left(Y_i - C_i^T\boldsymbol{\beta}_y\right)\\
U(\boldsymbol{\beta}_n) &= \frac{1}{\sum_{i:A_i=0}1}\sum_{i:A_i=0}C_i\left(N_i - C_i^T\boldsymbol{\beta}_n\right)\\
U(s_y) &= \frac{1}{\sum_{i:A_i=0}1}\sum_{i:A_i=0}\left(Y_i - C_i^T\boldsymbol{\beta}_y\right)^2 - s_y^2\\
U(s_n) &= \frac{1}{\sum_{i:A_i=0}1}\sum_{i:A_i=0}\left(N_i - C_i^T\boldsymbol{\beta}_n\right)^2 - s_n^2\\
U(\alpha) &= \frac{1}{\sum_{i:A_i=1}1}\sum_{i:A_i=1}\left[Y_i - s_y\left\{\frac{N^* - C_i^T\boldsymbol{\beta}_n}{s_n}\right\} - C_i^T\boldsymbol{\beta}_y\right] - \alpha.
\end{aligned}$$

The matrix $\frac{\partial}{\partial\boldsymbol{\theta}}U(\boldsymbol{\theta})$ is given by:

$$\begin{pmatrix}
\frac{\partial}{\partial\boldsymbol{\beta}_y}U(\boldsymbol{\beta}_y) & 0 & 0 & 0 & 0\\
0 & \frac{\partial}{\partial\boldsymbol{\beta}_n}U(\boldsymbol{\beta}_n) & 0 & 0 & 0\\
\frac{\partial}{\partial\boldsymbol{\beta}_y}U(s_y) & 0 & \frac{\partial}{\partial s_y}U(s_y) & 0 & 0\\
0 & \frac{\partial}{\partial\boldsymbol{\beta}_n}U(s_n) & 0 & \frac{\partial}{\partial s_n}U(s_n) & 0\\
\frac{\partial}{\partial\boldsymbol{\beta}_y}U(\alpha) & \frac{\partial}{\partial\boldsymbol{\beta}_n}U(\alpha) & \frac{\partial}{\partial s_y}U(\alpha) & \frac{\partial}{\partial s_n}U(\alpha) & \frac{\partial}{\partial\alpha}U(\alpha)
\end{pmatrix},$$

with:

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\beta}_y}U(\boldsymbol{\beta}_y) &= -\frac{1}{\sum_{i:A_i=0}1}\sum_{i:A_i=0}C_iC_i^T\\
\frac{\partial}{\partial\boldsymbol{\beta}_n}U(\boldsymbol{\beta}_n) &= -\frac{1}{\sum_{i:A_i=0}1}\sum_{i:A_i=0}C_iC_i^T\\
\frac{\partial}{\partial\boldsymbol{\beta}_y}U(s_y) &= -\frac{2}{\sum_{i:A_i=0}1}\sum_{i:A_i=0}C_i\left(Y_i - C_i^T\boldsymbol{\beta}_y\right)\\
\frac{\partial}{\partial s_y}U(s_y) &= -2s_y\\
\frac{\partial}{\partial\boldsymbol{\beta}_n}U(s_n) &= -\frac{2}{\sum_{i:A_i=0}1}\sum_{i:A_i=0}C_i\left(N_i - C_i^T\boldsymbol{\beta}_n\right)
\end{aligned}$$

$$\frac{\partial}{\partial s_n}U(s_y) = -2s_n$$
$$\frac{\partial}{\partial \boldsymbol{\beta}_y}U(\alpha) = -\frac{1}{\sum_{i:A_i=1} 1}\sum_{i:A_i=1} C_i^T$$
$$\frac{\partial}{\partial \boldsymbol{\beta}_n}U(\alpha) = \frac{1}{\sum_{i:A_i=1} 1}\sum_{i:A_i=1} \frac{s_y}{s_n} C_i^T$$
$$\frac{\partial}{\partial s_y}U(\alpha) = -\frac{1}{\sum_{i:A_i=1} 1}\sum_{i:A_i=1} \left\{\frac{N^* - C_i^T\boldsymbol{\beta}_n}{s_n}\right\}$$
$$\frac{\partial}{\partial s_n}U(\alpha) = \frac{1}{\sum_{i:A_i=1} 1}\sum_{i:A_i=1} s_y\left\{\frac{N^* - C_i^T\boldsymbol{\beta}_n}{s_n^2}\right\}$$
$$\frac{\partial}{\partial \alpha}U(\alpha) = -1.$$

Finally, the covariance matrix of the estimators is given by

$$\left[\frac{\partial}{\partial \boldsymbol{\theta}}U(\boldsymbol{\theta})\right]^{-1} \mathbb{P}_n\left[U_i(\boldsymbol{\theta})U_i^T(\boldsymbol{\theta})\right]\left[\frac{\partial}{\partial \boldsymbol{\theta}}U(\boldsymbol{\theta})\right]^{-1},$$

where U_i is an individual equation for subject i, and $\mathbb{P}_n[x_i] = 1/n\sum_{i=1}^n x_i$.

[*Received August 1, 2010; revised October 1, 2010; accepted for publication November 1, 2010*]

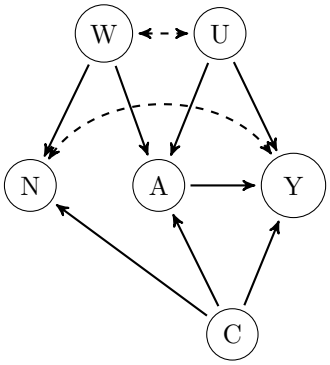

Fig. 3. Directed acyclic graph depicting the causal association between the treatment A, primary outcome Y, negative control outcome N, measured pre-exposure confounders C, and unmeasured confounders U and W of the primary and secondary outcomes, respectively.

Table 2. *5-fold cross-validated prediction scores comparing two models for the variances. The 'homoscedasticity' option assumes homoscedasticity across all levels of the confounding variables, and 'model variance' assumes that the covariates affect the error variance via a log-linear model.*

Outcome	homoscedasticity	model variance
log-fibrinogen	0.032	0.007
log-BMI	0.032	0.001

Table 3. *NAS cohort characteristics, for participants observed between November 2000 and December 2004. Measures are given in medians and ranges are in parentheses.*

Characteristic	value
Number of participants	616
Number of visits	703
Age	74 (58, 92)
BMI	27.6 (17.9, 46)
Fibrinogen	328 (109, 741)
Black carbon concentration (Areal)	1.18 (0.32, 2.02)
Black carbon concentration (Address)	0.75 (0.42, 1.17)

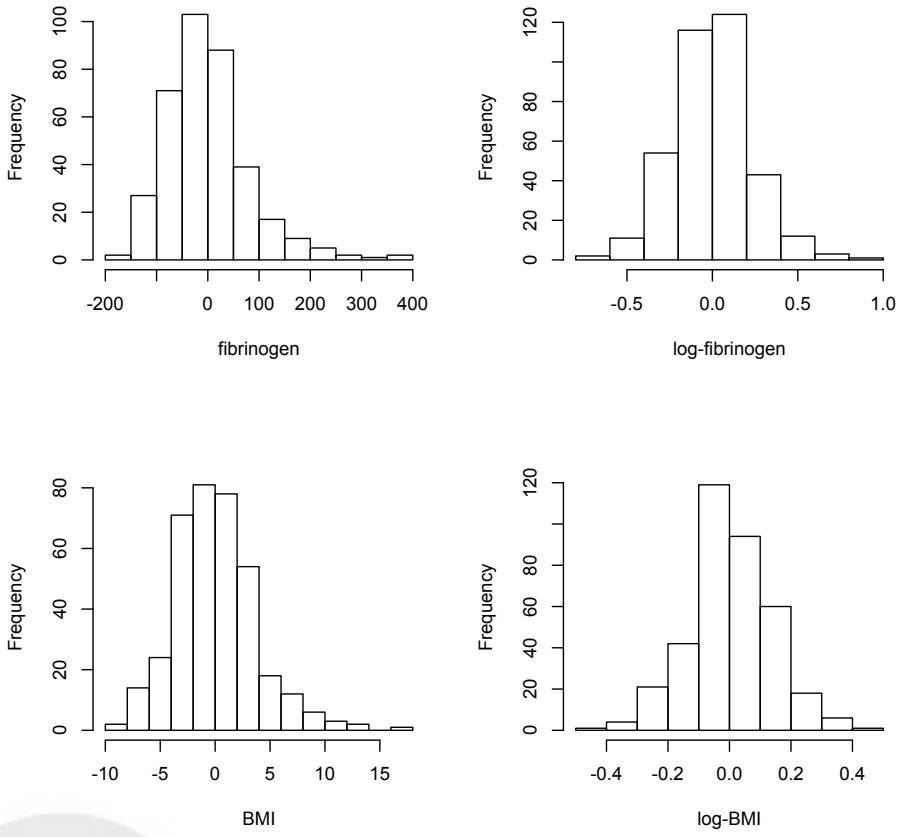

Fig. 4. Histograms of the residuals of the primary outcome (fibrinogen) and negative control outcome (BMI), and their log transformations, after regressing on the covariates in the low-exposure group.

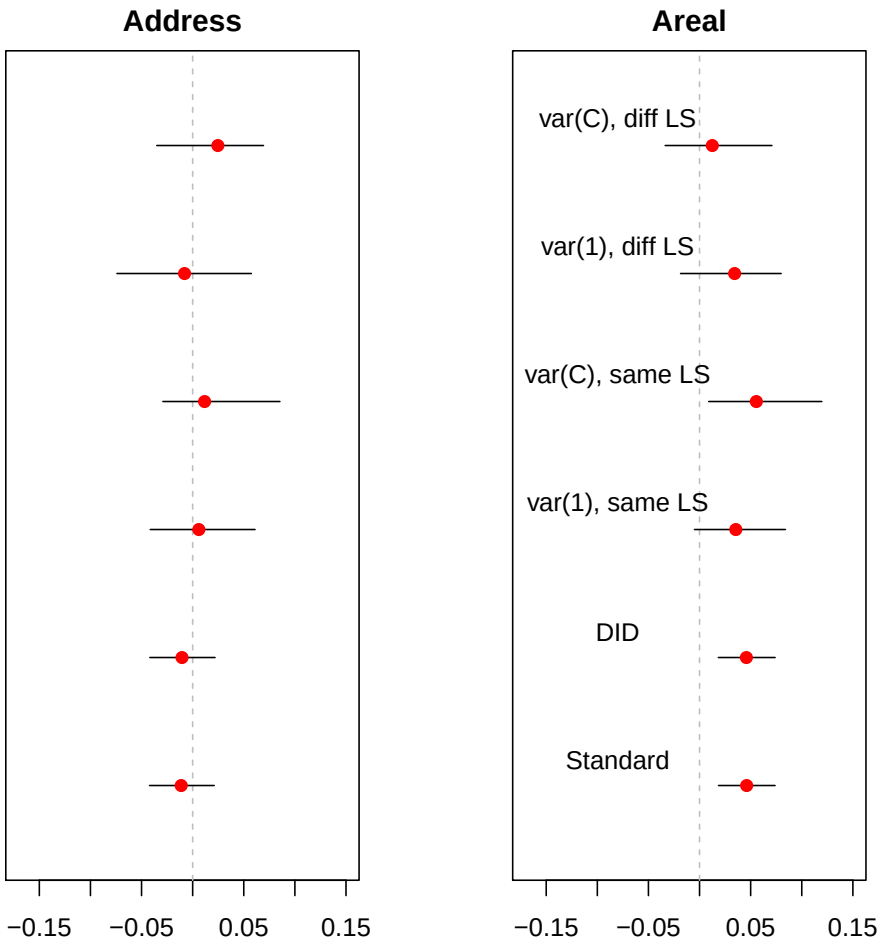

Fig. 5. Estimates of the effect of exposures to BC on log-fibrinogen as a binary variable, with values either predicted at participants' home addresses (left), or measured at an areal sensor at Boston (right), and 95% bootstrap confidence intervals. Effects were estimated using the indirect adjustment method with log-BMI as the negative control outcome, and compared to standard regression adjusted to BMI, and to the naïve DID method that assumes that the negative control outcome log-BMI is measured at the same scale as the primary outcome. var(1) and var(C) refer to modeling the variance using a log-linear model assuming dependence on covariates (var(C)), or homoscedasticity (var(1)). The location-scale family (LS) of log-fibrinogen and log-BMI were assumed identical ("same") or different ("diff").