Statistical Science
2016, Vol. 31, No. 3, 348–361
DOI: 10.1214/16-STS558

On Negative Outcome Control of Unobserved Confounding as a Generalization of Difference-in-Differences

Tamar Sofer, David B. Richardson, Elena Colicino, Joel Schwartz and Eric J. Tchetgen Tchetgen

Abstract. 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 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 in simulations and 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 and phrases: 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)

Tamar Sofer is Research Scientist at the Department of Biostatistics, University of Washington, UW Tower, 15th Floor, 4333 Brooklyn Ave. NE, Seattle, Washington 98105, USA (e-mail: tsofer@uw.edu). David B. Richardson is an Associate Professor of Epidemiology at the UNC Gillings School of Global Public Health, 2102b Mcgavran-Greenberg 135 Dauer Drive, Chapel Hill, North Carolina 27599, USA (e-mail: david.richardson@unc.edu). Elena Colicino is a Research Scientist at the Department of Environmental Health Sciences, Columbia University, 722 West 168th St. New York, New York 10032, USA (e-mail: ec3168@cumc.columbia.edu). Joel Schwartz is a Professor of Environmental Epidemiology, Harvard T.H. Chan School of Public Health, 665 Huntington Avenue, Landmark Center Room 415, Boston, Massachusetts 02115, USA (e-mail: joel@hsph.harvard.edu). Eric J. Tchetgen Tchetgen is a Professor of Biostatistics and Epidemiologic Methods, Harvard T.H. Chan School of Public Health, 677 Huntington Ave, Boston, Massachusetts 02115, USA (e-mail: etchetge@hsph.harvard.edu).

estimation (Card, 1990; Angrist and Krueger, 1999; Meyer, Kip Viscusi and Durbin, 1995), 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, Tchetgen Tchetgen and Cohen (2010) and Flanders et al. (2011) discussed using negative control outcomes to detect confounding by unmeasured factors, as indicated by evidence of an association between the exposure and the negative control outcome conditional on observed confounders. Tchetgen Tchetgen (2014) proposed a calibration approach to correct causal effect estimates for bias due to unobserved confounding. However, the identification conditions of Tchetgen Tchetgen (2014) require that the ranks of the outcome of interest be preserved under exposure and no exposure conditions (also known as rank preservation). It is of interest to identify more general 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 that 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, their model-based approach relies 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, Hernández-Díaz and Robins, 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 Y_a denote an individual's outcome if exposure A were set, possibly contrary to fact, to a. In this work, we are interested in estimating the so-called marginal average effect of treatment on the treated (ETT), defined as

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

Let N denote a negative control outcome variable. The relationships between these variables 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, Tchetgen Tchetgen and Cohen, 2010). Note that N and Y can be (but do not have to be) associated independently of their common causes U and C, as in the simple DID scenario in which the negative control outcome is the pre-exposure value of the outcome of interest.

To provide identifiability conditions for the causal effect of A on Y, 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 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

$$E\{Y_0 | A = 1, c, u\} - E\{Y_0 | A = 0, c, u\} = 0 \quad (1)$$

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

$$E\{Y_0 | A = 1, c\} - E\{Y_0 | A = 0, c\} \neq 0 \quad (2)$$

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 $Y(t)$ is measured at 2 occasions, $t = 0, 1$, with $Y(0)$ and $Y(1)$ pre- and post-exposure variables, respectively. According to this graph, although A is a cause of $Y(1)$, it does not cause $Y(0)$ (although the reverse may hold), and the unobserved confounder of the effect of A on $Y(1)$, U, is also a cause of $Y(0)$. 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 $Y(0)$ as N and $Y(1)$ 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 $Y(t)$ follows the simple linear model (where individual observations are suppressed in the notation)

$$\begin{aligned} &E\{Y(t) | U, A, C\} \\ &\quad = b(U) + m(t) + \beta t A + \gamma(t)^T C \end{aligned} \quad (3)$$

such that $m(t)$ indexes a time-specific intercept, $\gamma(t)$ indexes a time-specific association between C and

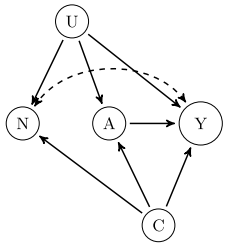

FIG. 1. *Directed acyclic graph depicting the causal association between the treatment A, primary outcome Y, negative control outcome N, measured pre-exposure confounders C and unmeasured confounders U.*

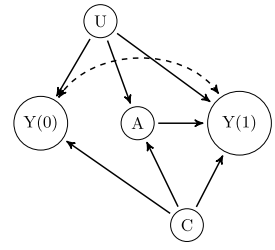

FIG. 2. *Directed acyclic graph depicting the causal association between the treatment A, pre-exposure outcome $Y(0)$, post-exposure outcome $Y(1)$, measured pre-exposure confounders C and unmeasured confounders U.*

$Y(t)$, $b(U)$ indexes the effect of U on $Y(t)$ which is assumed independent of t, A and C, and β encodes the causal effect of A on $Y(1)$. Let $Y_a(t)$ denote the counterfactual outcome at t under exposure a, and note that the key assumption encoded in equation (3) is that

$$E\{Y(1) - Y(0)|U, A = a, C\} = E\{Y(1) - Y(0)|A = a, C\}, \quad a = 0, 1 \tag{4}$$

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

$$E\{Y_0(1) - Y_0(0)|A = 1, C\} = E\{Y_0(1) - Y_0(0)|A = 0, C\}. \tag{5}$$

Since treatment is assumed to start only after time 0, so that $E[Y_1(0)|A = 1, C] = E[Y_0(0)|A = 1, C]$, and using equation (5), we obtain the following equality:

$$E\{Y_1(1) - Y_1(0)|A = 1, C\} - E\{Y_0(1) - Y_0(0)|A = 0, C\} \tag{6}$$

$$= \beta \tag{7}$$

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

The effect identified in (6) defines the DID estimand under equation (3) and, therefore, under assumption (4) is equal to $E\{Y_1(1) - Y_0(1)|A = 1\}$, the marginal causal effect of treatment on the treated. Interestingly, rather than assuming equation (3), one may take equation (5) as a primitive condition, which may hold without necessarily assuming the model given by equation (3) holds exactly. Only assuming that (5) holds has previously been shown to suffice for nonparametric identification of the marginal ETT even when the linear model (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 (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 (5):

$$E\{Y_0(1)|A = 1, C\} - E\{Y_0(1)|A = 0, C\} = E\{Y(0)|A = 1, C\} - E\{Y(0)|A = 0, C\}$$

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

$$E\{Y_0|A = 1, C\} - E\{Y_0|A = 0, C\} = E\{N|A = 1, C\} - E\{N|A = 0, C\}, \tag{8}$$

where the left-hand side of (8) encodes the degree of confounding bias (2) for the effect of A on Y, and the right-hand side of (8) likewise represents confounding bias for the (null) effect of A on N. Equation (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 (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 the marginal ETT α in the presence of unobserved confounding and likewise if one has access to a negative outcome control variable N (which may differ from a pre-exposure realization of the outcome).

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 Angrist and Pischke, 2008; 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 (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

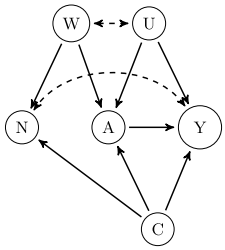

FIG. 3. *Directed acyclic graph depicting the causal association between the treatment A, primary outcome Y, negative control outcome N, measured pre-exposure confounders C and unmeasured confounders U and W of the primary and secondary outcomes, respectively.*

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.

ASSUMPTION 2. *$A \perp\!\!\!\perp Y_0 | C, U$, however $A \not\perp\!\!\!\perp Y_0 | C$; and $A \perp\!\!\!\perp N | C, W$, however $A \not\perp\!\!\!\perp N | C$.*

This more general framework is depicted in Figure 3. In addition to this causal diagram, in order to appropriately account for possible nonlinearity 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*

$$Y_0 = h_y(U, C), \tag{9a}$$

$$N = h_n(W, C), \tag{9b}$$

where

$$h_y(u, c) \text{ is monotone increasing in } u \text{ for each } c, \tag{10a}$$

and

$$h_n(w, c) \text{ is monotone increasing in } w \text{ for each } c. \tag{10b}$$

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 (10a), (10b) can be changed without any real consequence. This assumption might be most compelling if one has available specific knowledge about what common cause of the treatment and the outcome, although unobserved, might satisfy the monotonicity assumption, even if just approximately. Such knowledge would strengthen credibility in the monotonicity assumption, violation of which is likely to invalidate the proposed approach without an alternative assumption.

We now consider quantile–quantile and distributional equi-confounding as 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. Let $F_{X|Z}(\cdot)$ denote the cumulative distribution function of a X given Z, let $f_{X|Z}$ be the corresponding density function, $F^{-1}_{X|Z}(\cdot)$ its inverse map, and let $f \circ g(x) = f(g(x))$ denotes composition of functions f and g. Define the quantile–quantile (qq) association between U and A conditional on C:

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

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, that is, $q_0(v|c) - v \neq 0$ for some value c. Likewise let

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

Figure 4 provides an example of an estimated qq-transformation function between two distributions estimated from the data set discussed in Section 5. The

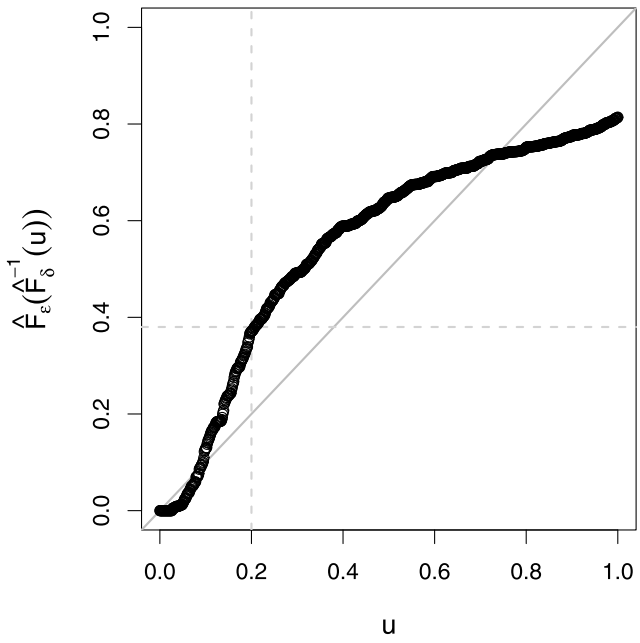

FIG. 4. *The qq-transformation between $\widehat{F}_\delta$ to $\widehat{F}_\varepsilon$, defined as $\widehat{F}_\varepsilon \circ \widehat{F}_\delta^{-1}(u)$ for $u \in [0, 1]$. This qq-transformation was estimated from the NAS data set, where u is the empirical cumulative probability of the scaled residuals of log-BMI and the qq-transformation maps each such value to the empirical cumulative probability of the scaled residuals of log-fibrinogen.*

function maps the probability distribution of the scaled residuals of log-BMI to a probability distribution of the scaled residuals of log-fibrinogen. The diagonal curve corresponds to the hypothesis that both sets of residuals follow a common distribution such that the 20th percentile under one matches that under the other. However, in Figure 4 the value of the 20th percentile of $\widehat{F}_\delta$ corresponds to the 38th percentile under $\widehat{F}_\varepsilon$. The quantile–quantile equi-confounding assumption is given below.

ASSUMPTION 4. *Quantile–quantile equi-confounding*.

$$q_0(v|c) = q_1(v|c), \quad v \in [0,1]. \tag{11}$$

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. Note that both q_0 and q_1 while being equal under the assumption, will generally not be equal to the identity map in the presence of unobserved confounding. 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. Let $X \sim Z$ denote that X and Z follow a common distribution.

ASSUMPTION 5. *Distributional equi-confounding*.

$$U|A, C \sim W|A, C. \tag{12}$$

The assumption states that the conditional distribution of the unobserved confounder for Y is the same as that for N given A and C. Note that both Assumptions 4 and 5 are trivially satisfied, if as previously assumed, the unobserved confounder of Y and N is the same, that is, $U = W$. Note also that both assumptions are considerably weaker than the previous additive equi-confounding assumption (8) because they place no restriction beyond monotonicity 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. Define $N^* \sim N|A = 1, C$ to be a random variable distributed as the negative outcome in the exposed group. The main identification result requires the additional regularity condition.

ASSUMPTION 6. *Positivity.*

$$\begin{aligned} &\textit{If } 0 < f_{N|A=1,C}(N^*), \\ &\quad \textit{then } 0 < F_{N|A=0,C}(N^*) < 1. \end{aligned} \tag{13}$$

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 $F_{N|A=0,C}(N^*)$ will not be identically 1 or 0 for some set of plausible values of N^*.

THEOREM 1. *Under Assumptions 1–4 and 6, we have that*

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

where

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

All proofs can be found in Appendix A. It is helpful to contrast the estimand obtained in Theorem 1 under qq equi-confounding, to that obtained under additive equi-confounding. Recall that under the latter condition equation (6) states that the ETT is given by

$$\begin{aligned} \beta = {} & E\{Y|A = 1, C\} - E\{Y|A = 0, C\} \\ & + E\{N|A = 0, C\} - E\{N|A = 1, C\}. \end{aligned}$$

Under qq equi-confounding bias $E\{F^{-1}_{Y|A=0,C} \circ F_{N|A=0,C}(N^*)\}$ is substituted for $E\{Y|A = 0, C\} + E\{N|A = 0, C\} - E\{N|A = 1, C\}$ as the negative control-adjusted identifying expression for the conditional counterfactual mean $E(Y_0|A = 1)$.

Theorem 1 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's (2006) primary goal was to account for possible nonlinearity 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.

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

$$E(N|A=0,C) = \mu_n(C),$$
$$\mathrm{Var}(N|A=0,C) = s_n^2(C),$$

and

$$\delta = \frac{N-\mu_n(C)}{s_n(C)},$$

and likewise, let

$$E(Y|A=0,C) = \mu_y(C),$$
$$\mathrm{Var}(Y|A=0,C) = s_y^2(C)$$

and

$$\varepsilon = \frac{Y-\mu_y(C)}{s_y(C)}.$$

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

$$\varepsilon|A=0,C \sim f_\varepsilon(\varepsilon), \qquad \delta|A=0,C \sim f_\delta(\delta), \tag{14}$$

where $f_\varepsilon(\cdot)$ and $f_\delta(\cdot)$ are unrestricted baseline densities with cumulative distribution functions $F_\varepsilon(\cdot)$ and $F_\delta(\cdot)$.

COROLLARY 1. *Under the assumptions stated in Theorem 1 and the location-scale model (14), we have that*

$$\widetilde{Y} = s_Y(C)\left\{F_\varepsilon^{-1} \circ F_\delta\left(\frac{N^*-\mu_n(C)}{s_n(C)}\right)\right\} + \mu_y(C), \tag{15}$$

and in the special case where $F_\varepsilon(\cdot) = F_\delta(\cdot)$, then

$$\widetilde{Y} = s_y(C)\left\{\frac{N^*-\mu_n(C)}{s_n(C)}\right\} + \mu_y(C). \tag{16}$$

Note also that if $F_\delta(\cdot) = F_\varepsilon(\cdot)$, that is, if the distribution of scaled-residuals ε and δ coincide 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 $F_\delta(\cdot)$ and $F_\varepsilon(\cdot)$ to be different, and then further assuming $F_\delta(\cdot) = F_\varepsilon(\cdot)$.

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, for example, one may take $\widehat{\mu}_n(C) = \widehat{\pi}_0 + \widehat{\pi}_1' C$ the ordinary least squares estimator of $E(N|A=0,C)$ using the subsample with $A = 0$, and likewise one may take $\widehat{s}_n^2(C) = \exp(\widehat{\omega}_0 + \widehat{\omega}_1' C)$ a standard log-linear regression of the squared $N - \mu_n(C)$ in the unexposed subsample, and similarly for $\widehat{\mu}_y(\cdot)$ and $\widehat{s}_y(\cdot)$. Further, let $\widehat{F}_\delta(\cdot)$ and $\widehat{F}_\varepsilon(\cdot)$ denote the empirical cumulative distribution functions of $\widehat{\varepsilon}$ and $\widehat{\delta}$ where $\widehat{\varepsilon} = \{Y - \widehat{\mu}_y(C_i)\}/\widehat{s}_y(C_i)$ and $\delta = \{N - \widehat{\mu}_n(C)\}/\widehat{s}_n(C)$ among the unexposed, that is, when $A = 0$. Specifically,

$$\widehat{F}_\varepsilon(u) = \frac{1}{n_0}\sum_{i=1}^{n_0} \mathbb{I}\{\widehat{\epsilon}_i \le u\},$$

where $\mathbb{I}(\cdot)$ is the indicator of an event.

Let n_1 denote the number of exposed persons.

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

$$\widehat{\alpha}_1 = \Bigg(\sum_{i:A_i=1} Y_i - \widehat{s}_y(C_i)\left\{\widehat{F}_\varepsilon^{-1} \circ \widehat{F}_\delta\left(\frac{N_i^*-\widehat{\mu}_n(C_i)}{\widehat{s}_n(C_i)}\right)\right\} - \widehat{\mu}_y(C_i)\Bigg) \Big/ n_1. \tag{17}$$

2. Under homoscedasticity, that is, $\widehat{s}_y(C_i) = \widehat{s}_y$ obtained in a intercept-only regression, and similarly for $\widehat{s}_n$, we get

$$\widehat{\alpha}_3 = \Bigg(\sum_{i:A_i=1} Y_i - \widehat{s}_y\left\{\widehat{F}_\varepsilon^{-1} \circ \widehat{F}_\delta\left(\frac{N_i^*-\widehat{\mu}_n(C_i)}{\widehat{s}_n}\right)\right\} - \widehat{\mu}_y(C_i)\Bigg) \Big/ n_1. \tag{18}$$

3. Assuming $F_\delta(\cdot) = F_\varepsilon(\cdot)$, (17) simplifies to

$$\widehat{\alpha}_3 = \Bigg(\sum_{i:A_i=1} Y_i - \widehat{s}_y(C_i)\left(\frac{N_i^*-\widehat{\mu}_n(C_i)}{\widehat{s}_n(C_i)}\right) - \widehat{\mu}_y(C_i)\Bigg) \Big/ n_1. \tag{19}$$

4. And finally, under both homoscedasticity and $F_\delta(\cdot) = F_\varepsilon(\cdot)$, we get

$$\widehat{\alpha}_4 = \frac{\sum_{i:A_i=1}[Y_i - \widehat{\mu}_y(C_i)]}{n_1} - \frac{\widehat{s}_y}{\widehat{s}_n}\frac{[N_i^* - \widehat{\mu}_n(C_i)]}{n_1} = \widehat{\eta}_y - \frac{\widehat{s}_y}{\widehat{s}_n}\widehat{\eta}_n, \tag{20}$$

where $\widehat{\eta}_y$ and $\widehat{\eta}_n$ are regression-based 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, obtained from a linear regression of Y on A and C, 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.

These four estimators are all regular and asymptotically linear under standard regularity conditions. In the Appendix, we provide a simple expression for the large sample variances of $\widehat{\alpha}_3$ and $\widehat{\alpha}_4$ 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 simulate 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 was generated under $C|A = 0 \sim \mathcal{N}(0, 1), C|A = 1 \sim \mathcal{N}(0.5, 1)$.

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

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

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.

Briefly consider the assumptions that our estimators are based on in light of the generating models for the simulations. First, it is clear that U and C (W and C) are associated with both Y (N) and A, so that C and U, and C and W comprise of all the confounders of the A–Y and A–N associations respectively, satisfying Assumption 2. Further, because U and W have the same distributions, the distributional equi-confounding bias Assumption 5 (and, therefore, the weaker Assumption 4) is satisfied. Assumption 3 is clearly satisfied. Finally, note that under the uniform distribution scenario, the positivity Assumption 6 does not hold and, therefore, the estimators $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ from Section 3.2 may be biased. However, the estimators $\widehat{\alpha}_3$ and $\widehat{\alpha}_4$ that assume $F_N(\cdot) = F_Y(\cdot)$ should not be biased, since in this case the positivity assumption is not required.

We generated data with $n = 100, 500$ observations, and $n/2$ 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 are generated under homoscedastic errors, with U and W following a common distribution given A and C, nonetheless, we consider inferences about the marginal ETT α using the NOC 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-based estimator that simply 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 (estimator $\widehat{\alpha}_4$), 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 (estimator $\widehat{\alpha}_1$) resulted in only slightly larger MSEs. However, nonparametric estimation of the qq-transformation had mixed effects. Under normal distribution of the unobserved confounders, estimating the qq-transformation (estimators $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$) had little effect on the bias and efficiency of the estimators. However, under uniform distribution of the unmeasured confounders, estimating this transformation

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. "Standard" is the naïve regression estimator of Y on C and A. Other estimators either model the qq-transfomation between the standardized primary outcome Y and the negative control outcome N in the unexposed group nonparametrically (α_1, α_2, "Nonparametric qq"), or assume that this it the identity (α_3, α_4, "Identity qq"), and either model the variance as a function of covariates C (α_1, α_3), or assume homoscedasticity (α_2, α_4)

			Nonparametric qq		Identity qq	
Family	***n***	**Standard**	α_1	α_2	α_3	α_4
Normal	100	5.99 (36.83)	0.52 (02.65)	0.47 (02.54)	0.13 (02.99)	0.05 (02.72)
Normal	500	5.99 (36.06)	0.12 (00.61)	0.12 (00.57)	0.03 (00.59)	0.01 (00.53)
Uniform	100	9.09 (85.15)	2.59 (10.03)	2.61 (10.06)	0.03 (05.98)	0.03 (05.65)
Uniform	500	8.97 (81.02)	2.31 (06.10)	2.34 (06.22)	0.03 (01.27)	0.01 (01.23)

resulted in substantially larger MSEs and biased estimators. This may be because the positivity condition did not hold in this setting. The naïve estimator that regresses Y against A and C 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 et al. (2006), fibrinogen levels were shown to be associated with 4 weeks of exposure to BC in the Normative Aging Study (NAS) cohort. The investigators took 4 weeks moving averages of BC, measured at an areal sensor, just prior to a clinic visit as the exposure, and adjusted for multiple confounders, including BMI. We now reanalyze 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 1727 complete cases (i.e., with observed exposures, measured covariates and outcome values) from visits between November 14, 2000, and December 31, 2004, as in Zeka et al. (2006). We use BC values measured either at the areal sensor in Boston (as in Zeka et al., 2006), or geospatial model-predicted values at participants' home addresses (Gryparis et al., 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 2 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 estimators compared in the simulations, that is, the more robust models allowing for heteroscedasticity, and/or different location-scale family and the model that assumes homoscedasticity and a common location-scale family. In addition to these estimators, we also compared the analysis to the naïve analysis that regresses log-fibrinogen on the BC measure of interest, covariates and BMI, as well as to a NOC approach under additive equi-confounding, which amounts to a standard DID-type analysis that assumes that the negative control outcome log-BMI is measured on the same scale as the primary outcome log-fibrinogen.

TABLE 2

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)

In order to evaluate the assumption of a common location-scale family, 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 4. One can see that after log-transformation both the primary and control outcomes have symmetric distributions, and it may be reasonable to assume that they are from a common location-scale family. We also observed that log-fibrinogen and log-BMI are measured on different scales. We also considered the empirical qq-transformation $\widehat{F}_Y^{-1} \circ \widehat{F}_N(u)$ in Figure 4. Upon inspection of the figure, this empirical curve clearly departs from the identity function, although such informal inspection does not appropriately account for uncertainty. This suggests that assuming a common location-scale family for the primary and negative control outcomes may not be appropriate.

Next, we assessed the homoscedasticity assumption. We used a 5-fold cross-validation of the restricted data set, where in each "fold" we took four-fifth of the participants to form a training set in which we estimated mean and variance models used to predict the outcomes (log-fibrinogen) of the held-out validation data set. We calculated the mean squared errors for these predictions as $\sum_{i=1}^{n_k}((y_i - C_i\widehat{\boldsymbol{\beta}}_y)^2 - \exp(C_i\widehat{\boldsymbol{\omega}}_y))^2$, 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 3 provides cross validation results, suggesting 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, $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$), the confidence intervals are wider, in agreement with the simulations studies. Next, consider our second set of analyses in which the dichotomized (high vs. low) BC exposure was measured at an areal sensor. For this model, based on the histograms in Figure 6 and the results from assessing heteroscedasticity in Table 3, the most appropriate estimators assumes that Y and N come from the same location-scale family ($\widehat{\alpha}_3$ and $\widehat{\alpha}_4$) and with heteroscedastic error ($\widehat{\alpha}_1$ and $\widehat{\alpha}_3$). Interestingly, in this case the effect estimates of BC are larger than the standard regression estimate.

TABLE 3

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

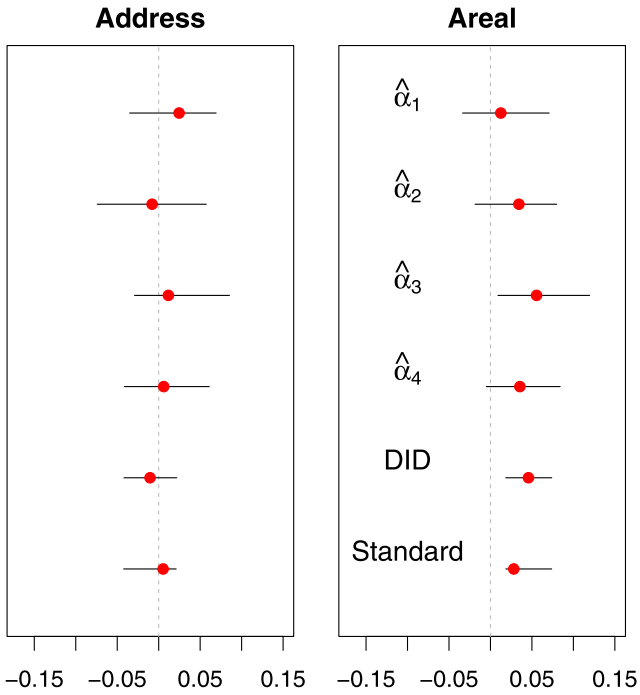

FIG. 5. *Estimates of the effect of exposures to BC on log-fibrinogen as a binary variable, with values either predicted at participants' home addresses (left), or measured at an areal sensor at Boston (right), and 95% bootstrap confidence intervals. Effects were estimated using the four estimators $\alpha_1, \ldots, \alpha_4$, 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.*

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 and the restrictive nature of the additive equi-confounding assumption in this application. More generally, even if the negative outcome is the pre-exposure value of the primary outcome, there may be important differences in variances across groups.

Interestingly, when using the predicted BC measures at the participants' home addresses, BC effect estimates are closer to null. This may be due to measurement error from the geospatial model used to predict

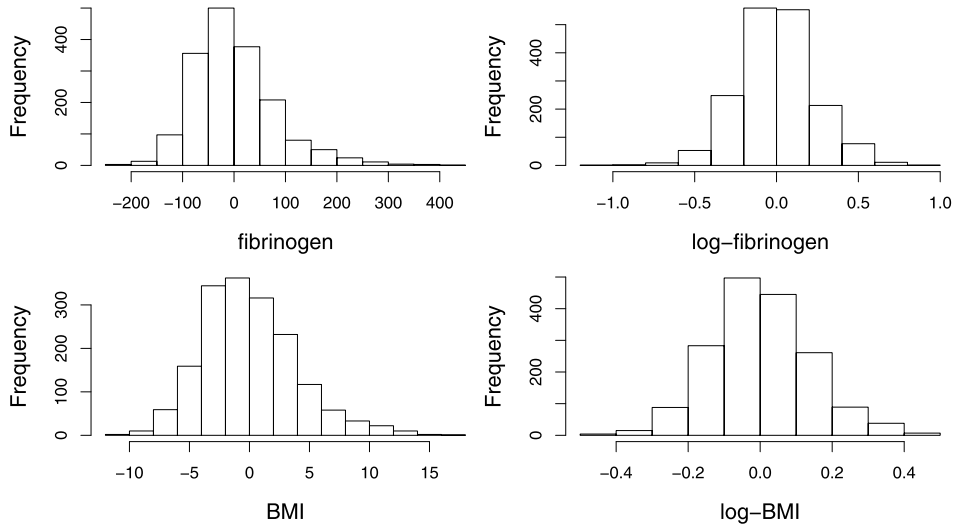

FIG. 6. *Histograms of the residuals of the primary outcome (fibrinogen) and negative control outcome (BMI), and their log transformations, after regressing on the covariates in the low-exposure group.*

the BC measurements. Such models were shown to often lead to biases toward the null in estimating air pollution effects (Zeger et al., 2000).

In contrast with standard regression, estimates based on NOC approaches allowing for 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. DISCUSSION

In this paper, we propose identification conditions paired with a principled approach for negative outcome control of unmeasured confounding to make inferences about ETT. Our approach draws on simple relations between DID and NOC to obtain simple identifying conditions for NOC. Our work also further generalizes such identifying assumptions by leveraging previous work by Athey and Imbens (2006). Another important contribution of this paper has been in addition to drawing parallels between seemingly unrelated literature, to propose a fairly flexible mode of inference for practical NOC application under a general location-scale formulation. Our simulation studies demonstrate that our proposed estimators perform well when the assumptions we posit are met; however, the approach appears to be particularly sensitive to violation of the positivity assumption, which is violated when the support of the scaled residuals of N in the exposed is not entirely contained in the support of the scaled residuals of N in the unexposed.

The location-scale model we use for estimation is an example of a so-called "transformation model". The main assumption of the location-model is that the association between the covariates and the outcome is only on the mean and variance scale, so that once centered and standardized, the outcome (i.e., scaled residual) is independent of covariates. A more familiar formulation of the model is $Y = \mu(X) + \sigma(X)\varepsilon$ where ε is an independent mean zero error with unit variance, and $\sigma^2(X)$ is the variance of $Y|X$. This formulation highlights the connection to standard regression analysis with heteroscedastic error. It is customary to assumed that both $\mu(X)$ and $\sigma(X)$ follow simple parametric models as posited in the paper. The semiparametric efficiency bound of this model is given in Bickel et al. (1993) where they show that the efficient score of regression and variance parameters depend on the density of ε. Another known example of a transformation model is the accelerated failure time model (Robins and Tsiatis, 1992; Cox and Oakes, 1984).

An outstanding question not directly addressed in this paper is how to select a good negative control variable in a given application. In general, a useful negative control outcome is easy to come about if the exposure has a specific target, for example, a vaccination for a specific disease. Then the selection of an outcome known to have no causal relation to the exposure in view can be better informed. Prior information on the possible source of unmeasured confound-

ing might also help identify a compelling control outcome. For instance, in a recent paper, Richardson et al. (2014) were interested in assessing possible confounding by (unmeasured) smoking behavior in evaluating the causal link between radon exposure and lung cancer. Both specificity of the exposure-outcome relation and the hypothesized confounder led the authors to selecting COPD as a credible negative control outcome. In our application, it is less clear. Many variables that likely share the same confounders as fibrinogen, share its biological pathways and that a causal link between BC exposures and such potential variables cannot a priory be ruled out with certainty. Therefore, our criterion for selecting a negative control outcome was sharpened by incorporating a restriction on the hypothesized time frame required to affect the negative control outcome to rule out the possibility of such a causal link. While many inflammation-related markers can be modified in short time frames such as 4 weeks, BMI will typically remain unchanged in this short time frame.

The instrumental variable approach is a well-known common approach to address unmeasured confounding in the social sciences and epidemiology. Although both IV and NOC address the challenging issue of unmeasured confounding, their assumptions have important differences. A valid IV must be directly related to the exposure, only affect the outcome through the exposure and must be independent of unmeasured confounders. Therefore, while a good negative control outcome is in essence as closely related to the unmeasured confounder as possible, the opposite is desirable of a valid IV. Despite this important distinction, both approaches can be viewed as a way to estimate the degree of selection bias due to unobserved confounding (Tchetgen Tchetgen and Vansteelandt, 2013).

SOFTWARE

An R function implementing the proposed estimators and an example simulation code can be found at: https://github.com/tamartsi/NOC_adjustment_to_estimate_ETT.

APPENDIX A: MATHEMATICAL DERIVATIONS

Define the inverse of the cumulative distribution function as

$$F^{-1}_{Y|A=0,C}(u) = \min\{v : p(Y < v|A = 0, C) = u\}.$$

PROOF OF THEOREM 1. Let $S_{N|A,C}(n) = P\{N \geq n|A, C\}$ and $F_{N|A,C}(n) = P\{N < n|A, C\}$.

First, we establish that Assumption 4 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

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

and also

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

where the third equality holds due to the monotonicity Assumption 3, taking the inverse and using Assumption 3 we get

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

and, therefore, by plugging-in (A.1) into the expression for $F_{Y_0|A=0,C} \circ F^{-1}_{Y_0|A=1,C}(v)$, we may conclude that

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

where we used Assumption 3 in the third identity and the definition of the cumulative distribution function in the last. Likewise,

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

Therefore, by the quantile–quantile equi-confounding Assumption 4,

$$\begin{aligned} &F_{U|A=0,C} \circ F^{-1}_{U|A=1,C}(v) \\ &\quad = F_{W|A=0,C} \circ F^{-1}_{W|A=1,C}(v) \\ &\quad \iff F_{Y_0|A=0,C} \circ F^{-1}_{Y_0|A=1,C}(v) \\ &\qquad = F_{N|A=0,C} \circ F^{-1}_{N|A=1,C}(v). \end{aligned} \tag{A.2}$$

Finally, we get

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

since $N^* \sim N|A=1,C$, or equivalently, $N^* \sim F^{-1}_{N|A=1,C}(V)$, where V is a uniformly distributed random variable, and using the positivity Assumption 6, and the consistency Assumption 1, we get that

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

proving the result. In the last equation, note that if $F^{-1}_{Y_0|A=0,C} \circ F_{N|A=0,C}(u) = u$, the positivity assumption is not required, since the $F_{N|A=0,C}$ is not applied to N^*. □

PROOF OF COROLLARY 1. From Theorem 1,

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

First, note that

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

Second, let $F^{-1}_{Y|A=0,C}(u) = v$, for $0 < u < 1$. Then

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

Thus,

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

Combining the two results, we get

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

Now, if $F_\epsilon(\cdot) = F_\delta(\cdot)$, trivially

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

APPENDIX B: ASYMPTOTIC VARIANCE OF THE LOCATION-SCALE NOC ESTIMATE

Assume that $s_y(C_i) = s_y$, $s_n(C_i) = s_n$. Let $\boldsymbol{\beta}_y$, $\boldsymbol{\beta}_n$ be the covariates effects on the outcomes Y and N, respectively, in a model where $\mu_y(C) = C^T\boldsymbol{\beta}_y$, $\mu_n(C) = C^T\boldsymbol{\beta}_n$. An 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:

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

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

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

$$\frac{\partial}{\partial s_y}U(s_y) = -2s_y,$$

$$\frac{\partial}{\partial \boldsymbol{\beta}_n}U(s_n) = -\frac{2}{\sum_{i:A_i=0}1}\sum_{i:A_i=0} C_i\big(N_i - C_i^T\boldsymbol{\beta}_n\big),$$

$$\frac{\partial}{\partial s_n}U(s_y) = -2s_n,$$

$$\frac{\partial}{\partial \boldsymbol{\beta}_y}U(\alpha) = -\frac{1}{\sum_{i:A_i=1}1}\sum_{i:A_i=1} C_i^T,$$

$$\frac{\partial}{\partial \boldsymbol{\beta}_n}U(\alpha) = \frac{1}{\sum_{i:A_i=1}1}\sum_{i:A_i=1}\frac{s_y}{s_n}C_i^T,$$

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

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

$$\frac{\partial}{\partial \alpha}U(\alpha) = -1.$$

Finally, the covariance matrix of the estimators is given by

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

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

REFERENCES

ABADIE, A. (2005). Semiparametric difference-in-differences estimators. *Rev. Econ. Stud.* **72** 1–19. MR2116973

ANGRIST, J. D. and KRUEGER, A. B. (1999). Empirical strategies in labor economics. In *Handbook of Labor Economics* **3A** (O. Ashenfelter and D. Card, eds.) 1277–1366. Elsevier, Amsterdam.

ANGRIST, J. D. and PISCHKE, J.-S. (2008). *Mostly Harmless Econometrics*: *An Empiricist's Companion*. Princeton Univ. Press, Princeton, NJ.

ATHEY, S. and IMBENS, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica* **74** 431–497. MR2207397

BICKEL, P. J., KLAASSEN, C. A., BICKEL, P. J., RITOV, Y., KLAASSEN, J., WELLNER, J. A. and RITOV, Y. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press, Baltimore, MD.

BLUNDELL, R. and MACURDY, T. (2000). Labor supply. In *Handbook of Labor Economics* **3A** (O. Ashenfelter and D. Card, eds.) 1559–1695. Elsevier, Amsterdam.

CARD, D. (1990). The impact of the Mariel boatlift on the Miami labor market. *Industrial & Labor Relations Review* **43** 245–257.

COX, D. R. and OAKES, D. (1984). *Analysis of Survival Data* **21**. CRC Press, Boca Raton. FL.

FLANDERS, W. D., KLEIN, M., DARROW, L. A., STRICKLAND, M. J., SARNAT, S. E., SARNAT, J. A., WALLER, L. A., WINQUIST, A. and TOLBERT, P. E. (2011). A method for detection of residual confounding in time-series and other observational studies. *Epidemiology* **22** 59.

GRYPARIS, A., COULL, B. A., SCHWARTZ, J. and SUH, H. H. (2007). Semiparametric latent variable regression models for spatiotemporal modelling of mobile source particles in the greater Boston area. *J. Roy. Statist. Soc. Ser. C* **56** 183–209. MR2359241

HERNÁN, M. A., HERNÁNDEZ-DÍAZ, S. and ROBINS, J. M. (2004). A structural approach to selection bias. *Epidemiology* **15** 615–625.

LIPSITCH, M., TCHETGEN TCHETGEN, E. and COHEN, T. (2010). Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology* **21** 383–388.

MEYER, B. D. (1995). Natural and quasi-experiments in economics. *J. Bus. Econom. Statist.* **13** 151–161.

MEYER, B. D., KIP VISCUSI, W. and DURBIN, D. L. (1995). Workers' compensation and injury duration: Evidence from a natural experiment. *The American Economic Review* **85** 322–340.

PEARL, J. (2009). *Causality*: *Models*, *Reasoning*, *and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166

RICHARDSON, D. B., LAURIER, D., SCHUBAUER-BERIGAN, M. K., TCHETGEN, E. T. and COLE, S. R. (2014). Assessment and indirect adjustment for confounding by smoking in cohort studies using relative hazards models. *Am. J. Epidemiol.* **180** 933–940.

ROBINS, J. and TSIATIS, A. A. (1992). Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika* **79** 311–319. MR1185133

TCHETGEN TCHETGEN, E. J. (2014). The control outcome calibration approach for causal inference with unobserved confounding. *Am. J. Epidemiol.* **179** 633–640.

TCHETGEN TCHETGEN, E. J. and VANSTEELANDT, S. (2013). Alternative identification and inference for the effect of treatment on the treated with an instrumental variable. Technical report, Harvard Univ. Biostatistics Working Paper Series.

ZEGER, S. L., THOMAS, D., DOMINICI, F., SAMET, J. M., SCHWARTZ, J., DOCKERY, D. and COHEN, A. (2000). Exposure measurement error in time-series studies of air pollution: Concepts and consequences *Environ. Health Perspect.* **108** 419.

ZEKA, A., SULLIVAN, J. R., VOKONAS, P. S., SPARROW, D. and SCHWARTZ, J. (2006). Inflammatory markers and particulate air pollution: Characterizing the pathway to disease. *Int. J. Epidemiol.* **35** 1347–1354.