A Negative Control Outcome Regression Accounting for Unobserved Confounding and Lagged Causal Effects

Hongkai Li  
Shandong University

Yuanyuan Yu  
Shandong University

Lei Hou  
Shandong University

Xiaoru Sun  
Shandong University

Xinhui Liu  
Shandong University

Fan Yang  
Shandong University

Yifan Yu  
Shandong University

Qing Wang  
Shandong University

Fuzhong Xue  
Shandong University

Research Article

Keywords: Negative Control Outcome Regression, Unobserved Confounding, Weighted Linear Regression Model, Lagged Causal Effects

DOI: https://doi.org/10.21203/rs.3.rs-696980/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
A Negative Control Outcome Regression Accounting for Unobserved Confounding and Lagged Causal Effects

Hongkai Li \(^1,^2\‡\), Yuanyuan Yu \(^1,^2\‡\), Lei Hou \(^1,^2\), Xiaoru Sun \(^1,^2\), Xinhui Liu \(^1,^2\), Fan Yang \(^1,^2\), Yifan Yu \(^1,^2\), Qing Wang \(^2\), Fuzhong Xue \(^1,^2\)

1. Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan, People’s Republic of China, 250000

2. Institute for Medical Dataology, Cheeloo College of Medicine, Shandong University, Jinan, People’s Republic of China, 250000

Email addresses:

HKL: lihongkaiyouxiang@163.com

YYY: yu_yy_1993@163.com

LH: 201815762@mail.sdu.edu.cn

XRS: sunny150139@126.com

XHL: lxinhui_1995@163.com

FY: fanyang.sdu@gmail.com

YFY: emailyuyifan@163.com

QW: 201999000066@sdu.edu.cn

FZX: xuefzh@sdu.edu.cn

‡ Equal contributors

* Corresponding Author: Dr. Fuzhong Xue. Email addresses: xuefzh@sdu.edu.cn
Abstract

**Background:** Epidemiologists are increasingly interested in using negative controls to eliminate unobserved confounding. Particularly, difference-in-differences method, which uses pre-exposure outcomes as negative control outcomes, is widely used. However, it obtains biased estimations when pre-exposure outcome has lagged causal effect on post-exposure outcome.

**Methods:** Taking advantage of pre-exposure outcomes as negative control outcomes, Negative Control Outcome Regression (NCOR) is proposed to eliminate unobserved confounding. The intercept term of NCOR provides an unbiased causal effect estimate of exposure on post-exposure outcome, and the slope minus 1 denotes the lagged causal effect estimation of pre-exposure outcome on post-exposure outcome. We then illustrate the potential of NCOR in a challenging application to estimate the causal association of PM$_{2.5}$ on all-cause mortality rates (AMR) and lagged causal effect of pre AMR on post AMR.

**Results:** Both theoretical justifications and simulation studies validate that the causal effect of exposure on outcome, along with the lagged causal effect of outcomes are identifiable and can be estimated by proposed NCOR model. The application results demonstrate that the previously estimated association between PM$_{2.5}$ and AMR can be attributed to the unobserved confounding. Furthermore, the NCOR model reveal that pre AMR has no causal association with post AMR.

**Conclusion:** The proposed NCOR model can obtain unbiased and robust causal effect estimation of exposure on outcome, and the lagged causal effect of outcomes. The proposed NCOR is implemented as an R package, called NCOR, and is freely available on GitHub.

**Key words:** Negative Control Outcome Regression; Unobserved Confounding; Weighted Linear Regression Model; Lagged Causal Effects
Background

Unobserved confounding is a well-known threat to identify and estimate the causal effects, and it is rarely avoided with certainty in observational studies [1, 2]. The use of negative controls to detect or eliminate unobserved confounding has gained increasing acceptance and popularity [3-5]. In general, a variable that is related to the unobserved confounding factors but not causally affected by the observed exposure is called negative control outcome [6-8]. Similarly, a negative control exposure is associated with the unobserved confounding factors, but it does not causally affect the interested outcome [6-8]. Remarkably, a proper negative control exposure or outcome needs to share common confounding mechanisms with the exposure or outcome variables. Difference-in-differences (DID) is typically quasi-experimental design to estimate the causal effect by comparing the changes in outcomes over time between two different groups [9-12]. As a matter of fact, DID method uses the pre-exposure outcomes as the negative control outcomes [9]. 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 [10-12]. However, it obtains biased estimations when pre-exposure outcome has lagged causal effect on post-exposure outcome.

Studies that used negative controls to detect potential confounding can be traced back to Rosenbaum who detected the potential confounders using an auxiliary outcome [13]. In epidemiological studies, Lipsitch et al. tested the unmeasured confounding, selection and measurement biases using negative control variables [4]. Recently, the use of negative control methods for correcting the confounding bias has been popularized instead of merely detecting the unmeasured confounding. Tchetgen et al proposed the outcome calibration method that regarded the counterfactual primary outcome as the proxy variable of unmeasured confounders [5]. In addition, Schuemie et al. developed an empirical null distribution of treatment effect via
a collection of negative controls to correct the p-value [14]. Sofer et al. interpreted the DID method as a negative outcome control approach through a monotonicity hypothesis of the unmeasured confounding effects [9]. However, it needs a strict assumption that pre-exposure outcome has no lagged causal effect on post-exposure outcome [9]. Using both negative control exposure and outcome variables under certain completeness conditions, Miao et al. proposed practical inference methods for identifying and inferencing the bridge function and the average causal effect [6]. Moreover, Shi et al. identified the average treatment effect through nonparametric estimation under weaker conditions when the unmeasured confounding and negative control variables were categorical [15]. They also proposed multiply robust and locally efficient estimators when nonparametric estimation may not be feasible [15]. Recently, Shi et al. reviewed the existing negative control methods including the detection, reduction, and correction for confounding bias, and further provided practical guidance on the use of negative control methods in epidemiology [16]. In the time-series studies, Flanders et al. partially corrected the residual confounding by using a future exposure as a negative control exposure [17]. However, future exposure worked as a surrogate for unobserved confounders, hence, it cannot obtain unbiased causal effect estimation [17]. The authors further proposed a bias reduction method for linear and log-linear time-series models through a negative control exposure, but this method required prior knowledge of the association between the confounder and negative control exposure [18]. Using both past and future exposures as multiple negative control exposures, Miao et al. extend Flanders’ method to identify and estimate the causal effect in the time-series setting [19]. Recently, Yu et al proposed a negative control exposure based on a time-series study model to effectively eliminate unobserved confounders using post-outcome exposure as a negative control exposure 20.

In this study, following the idea of meta-regression, we propose a negative control outcome regression (NCOR) approach to accurately estimate the causal effect of exposure on the future
outcome using only a pre-exposure outcome. Both theoretical proofs and simulation studies are
performed to validate the effectiveness of NCOR. Furthermore, we illustrate the potential of
the NCOR to estimate the causal association of PM$_{2.5}$ with the all-cause mortality rates (AMRs).
In addition, we also provide an R package that can be used by the research community to
implement the NCOR model, which is freely available on GitHub (https://github.com/yuyy-
shandong/NCOR).

Methods

Difference-in-difference model

DID model is most commonly used when pre- and post-exposure outcome can be observed.
The causal effect can be estimated by comparing the changes in outcomes over time between
two different groups. Figure 1(a) illustrates a visual representation of DID model. The outcome
$Y(t)$ is measured at 2 occasions, $t = 0, 1$ with $Y(0)$ and $Y(1)$ pre- and post-exposure
outcome, respectively. $X$ represents the exposure. In addition, we assume that confounders
(represented by a single variable $U$) are unknown.

Assumption 1(a). (Linear additive structural model)

$$E[Y(1)|U, X] = \alpha_0 + c_0 X + c_3 U;$$

Assumption 2(a). (No reverse causality)

$$Y(0) \perp X | U;$$

Assumption 3(a). (Parallel trend assumption)

$$c_1 = c_3,$$

where $c_1$ is the effect of unobserved confounder ($U$) on pre-exposure outcome ($Y(0)$) and
$c_3$ is the effect of unobserved confounder ($U$) on post-exposure outcome ($Y(1)$).

For Assumption 1(a), a simple causal model supposes that $Y(t)$ follows the simple linear
model, which is depend on exposure ($X$) and unmeasured confounders ($U$). Assumption 2(a)
represents that \( Y(0) \) does not affect future exposure \( X \), and the future exposure \( X \) does not affect pre-exposure outcome \( Y(0) \), reflected by the absence of arrow between \( Y(0) \) and \( X \).

For assumption 3(a), the association of the unobserved confounder with the outcome is assumed equal across exposure groups and constant over time.

Under the assumption as above[9, 11], the formula \( \frac{\partial E(Y(1) \mid X)}{\partial X} - \frac{\partial E(Y(0) \mid X)}{\partial X} \) can be used to estimate the causal effect of exposure on post-exposure outcomes.

However, the biased causal effect estimation can be obtained when the pre-exposure outcome has effect on the post-exposure outcome. In order to account for unobserved confounding and lagged causal effects, we proposed a novel method, NCOR, which can accurately estimate the causal effect of exposure on the future outcome, and the lagged causal effect of pre-exposure outcome on the post-exposure outcome.

**Negative Control Outcome Regression Model**

Figure 1(b) illustrates a visual representation of the NCOR model. Let \( Y(0) \) and \( Y(1) \) denote observed pre- and post- exposure outcome at time \( t=0 \) and \( t=1 \), respectively. \( X \) is an exposure and \( U \) denotes a set of confounders that considers all confounding effects between \( Y \) and \( X \). Let \( U(0), U(1/2), U(1) \) denote the unmeasured confounding factors at times \( t = 0, 1/2, 1 \), respectively. We are interested in estimating the causal effect of \( X \) on \( Y(1) \) and the lagged causal effect of \( Y(0) \) on \( Y(1) \). Using the \( do(x) \) operator of Pearl and linear additive structural model [21-24], the causal effect of \( X \) on \( Y(1) \) can be expressed as \( c_0 \) and the lagged causal effect of \( Y(0) \) on \( Y(1) \) can be expressed as \( \delta \).

We consider that \( J \) independent studies from different areas of spatial or research centers. For each participant \( i \) in the \( j \)-th \( (j=1,L,J) \) study, a pre-exposure outcome \( Y(0)_{i,j} \), an exposure \( (X_{i,j}) \), and a post-exposure outcome \( (Y(1)_{i,j}) \) can be observed.
Model Assumptions

Assumption 1(b). (Linear additive structural model),

For each study $j$, $E(Y_{1,j} | U_{1,j}, X_{j}, Y_{0,j}) = \alpha_{0,j} + \alpha_{1}X_{j} + \alpha_{2}U_{1,j} + \delta Y_{0,j}$;

Assumption 2(b). (No reverse causality assumption)

For each study $j$, $Y_{0,j} \perp X_{j} | U_{0,j}$;

Assumption 3(b). (Parallel trend assumption)

For each study $j$, $c_{1,j} = c_{2,j}$ and $U_{0,j}, U_{1/2}, U_{1,j}$ satisfy the stationary time series with fixed variance;

Assumption 4. (Fixed causal effect assumption)

For multiple studies, the causal effect ($c_{0}$) of $X$ on $Y(1)$ and the lagged causal effect ($\delta$) of $Y(0)$ on $Y(1)$ are fixed;

Assumption 5. (Random effect assumption)

For each study $j$, $c_{1,j} \xrightarrow{i.i.d.} F_{1}(c_{1}), r_{1,j} \xrightarrow{i.i.d.} F_{2}(r_{1}), c_{2,j} \xrightarrow{i.i.d.} F_{3}(c_{2})$.

Comparing with Assumption 1(a), Assumption 1(b) allows for the lagged causal effect of $Y(0)$ on $Y(1)$. Similarly with Assumption 2(a), Assumption 2(b) means that $Y(0)$ does not affect future exposure $X_{j}$, and the future exposure $X_{j}$ does not affect pre-exposure outcome $Y(0)$ for each study $j$. Assumptions 3(b) implies that the effects of confounders $U$ on the outcome $Y$ do not change over time for each study $j$. For example, it is understandable that the effect of socioeconomic status on the mortality in 2014 is same as that in 2015. The unmeasured confounding factors in DID model are single variables and do not change over time, Assumptions 3(b) assumes that the unmeasured confounders ($U$) is a stationary autoregressive process with a fixed auto-correlation coefficient (i.e. the effect of $U(0)$ on $U(1/2)$ is the same as the effect of $U(1/2)$ on $U(1)$, $r_{1,j} = r_{2,j}$) and fixed variance.
(\(\sigma_{U(0)}^2 = \sigma_{U(1)}^2 = \sigma_{U(2)}^2\)) for each study \(j\). In order to satisfy assumptions 3(b), we prefer that the time points of \(Y(0)\) and \(Y(1)\) are symmetrical with respect to those of \(X\).

**Theorem.** Under Assumptions 1(b)-3(b), 4, 5, the causal effect of \(X\) on \(Y(1)\) and the lagged causal effect of \(Y(0)\) on \(Y(1)\) can be identified and estimated by the following weighted regression based on \(J\) independent studies.

\[
\beta_1 = \kappa_0 + \kappa_1 \beta_o + \varepsilon \quad \varepsilon \sim N\left(0, \Sigma\right)
\]

\[
\Sigma = \text{diag}(se(\beta_{1,2}^2), se(\beta_{1,1}^2), ..., se(\beta_{1,J}^2))
\]

where \(\beta_1 = (\beta_{1,1}, \beta_{1,2}^T, ..., \beta_{1,J}^T)\), \(\beta_0 = (\beta_{0,1}, \beta_{0,2}^T, ..., \beta_{0,J}^T)\). For each study \(j\), \(\beta_{1,j}\) can be estimated by the linear regression \(Y(1)_j = \alpha_{1,j} + \beta_{1,j}X_j + \varepsilon_{1,j} \sim N(0, \sigma_{1,j}^2)\). And \(\beta_{0,j}\) can be estimated by the linear regression \(Y(0)_j = \alpha_{0,j} + \beta_{0,j}X_j + \varepsilon_{2,j} \sim N(0, \sigma_{2,j}^2)\). This model performs NCOR, a special case of the general method of meta-regression [25, 26]. Then, the causal effect of \(X\) on \(Y(1)\) can be identified and estimated by the intercept term \(\hat{\kappa}_0\). The lagged causal effect from \(Y(0)\) to \(Y(1)\) can be estimated by \(\hat{\kappa}_1 - 1\).

Based on a weighted linear regression model with the residual term following a normal distribution \(\Sigma\) [27-30], the estimator for \(\kappa_0\) is

\[
\hat{\kappa}_0 = \left\{ I_j - \beta_0 \left( \beta_o^T \Sigma^{-1} \beta_0 \right)^{-1} \left( \beta_o^T \Sigma^{-1} I_j \right) \right\}^T \Sigma^{-1} \beta_1 \left\{ I_j - \beta_0 \left( \beta_o^T \Sigma^{-1} \beta_0 \right)^{-1} \left( \beta_o^T \Sigma^{-1} I_j \right) \right\}^T \Sigma^{-1} I_j
\]

and the estimator for \(\kappa_1\) is

\[
\hat{\kappa}_1 = \left\{ \beta_0 - I_j \left( I_j^T \Sigma^{-1} I_j \right)^{-1} I_j^T \Sigma^{-1} \beta_0 \right\}^T \Sigma^{-1} \beta_1 \left\{ \beta_0 - I_j \left( I_j^T \Sigma^{-1} I_j \right)^{-1} I_j^T \Sigma^{-1} \beta_0 \right\}^T \Sigma^{-1} \beta_0
\]

The proof for NCOR is shown in Web Appendix 1 in Supplementary Material. More details
about the sensitivity analysis for Assumptions violated are provided in Web Appendix 2 in Supplementary Material.

Sample size calculation

To implement the NCOR method, we provide a formula to calculate the number of studies \((J)\) and sample size of each study \((n_j)\). The number of studies \((J)\) required for level \(\alpha\) and power \(\pi\) is

\[
J = \max(n_{\kappa_0}, n_{\kappa_1}), \quad \text{where } n_{\kappa_0} = \frac{[\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(\pi)]^2 \sigma(n_{\kappa_0})^2}{\kappa_0^2} + 1
\]

and

\[
n_{\kappa_1} = \frac{[\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(\pi)]^2 \sigma(n_{\kappa_1})^2}{(\kappa_1 - 1)^2} + 1.
\]

The sample size of each study \((n_j)\) required for level \(\alpha\) and power \(\pi\) is

\[
n_j = \max(n_{j,1}, n_{j,2}), \quad \text{where } n_{j,1} = \frac{[\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(\pi)]^2 \sigma(n_{j,1})^2}{\beta_{j,1}^2} + 1
\]

and

\[
n_{j,2} = \frac{[\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(\pi)]^2 \sigma(n_{j,2})^2}{\beta_{j,2}^2} + 1.
\]

More details about the sample size calculation are provided in Web Appendix 3 in Supplementary Material.

Simulation

To further investigate the statistical potential of the NCOR model, we perform a series of simulation studies with the artificial data of \(J\) studies under realistic conditions. For each study, the simulated data is generated as follows:

\[
U(0,1/2,1) \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & r & r^2 \\ r & 1 & r \\ r^2 & r & 1 \end{bmatrix}\right);
\]

\[
Y(0) = c_1 U(0) + \varepsilon_1, \quad \varepsilon_1 \sim N(0, \sigma^2_{\varepsilon_1});
\]

\[
X = c_2 U(1/2) + \varepsilon_2, \quad \varepsilon_2 \sim N(0, \sigma^2_{\varepsilon_2});
\]

\[
Y(1) = c_3 X + c_4 U(1) + \delta Y(0) + \varepsilon_3, \quad \varepsilon_3 \sim N(0, \sigma^2_{\varepsilon_3});
\]

\[
c_2 \sim N(p_2, \sigma^2_{c_2}); \quad c_3 \sim N(p_3, \sigma^2_{c_3}).
\]
To evaluate the performances of NCOR model, five methods including the linear regression model conditional on the surrogate of unmeasured confounders (i.e., the pre-exposure outcome) (LR-SU model), fixed-effect meta-analysis based on the LR-SU model (LR-SU-Meta model), DID model, fixed-effect meta-analysis based on the DID model (DID-Meta model) and NCOR model are used to estimate the causal effect of $X$ on $Y(1)$. The details of these models are as follows:

i) LR-SU model: $y(1) = \lambda_0 + \lambda_1 x + \lambda_2 y(0) + \epsilon_\lambda$ for all data, where $\lambda_1$ stands for the association between $X$ and $Y(1)$ after partial correction for unmeasured confounders;

ii) LR-SU-Meta model: For each study $j$, we have $y(1)_j = \lambda_{0,j} + \lambda_1 x_j + \lambda_2 y(0)_j + \epsilon_{\lambda,j}$. A fixed-effect meta-analysis is used to estimate the association between $X$ and $Y(1)$;

iii) DID model: $y(1) - y(0) = \eta_0 + \eta_1 x + \epsilon_\eta$ for all data, where $\eta_1$ is the causal effect estimation of $X$ on $Y(1)$;

iv) DID-Meta model: For each study $j$, we have $y(1) - y(0)_j = \eta_{0,j} + \eta_1 x_j + \epsilon_{\eta,j}$. A fixed-effect meta-analysis is used to estimate the causal effect of $X$ on $Y(1)$;

v) NCOR model: $\hat{\lambda}_0 = \frac{\left(I_J - \beta_0 \left(\beta_0^T \Sigma^{-1} \beta_0\right)^{-1} \left(\beta_0^T \Sigma^{-1} I_J\right)\right)^T \Sigma^{-1} \beta_1}{\left(I_J - \beta_0 \left(\beta_0^T \Sigma^{-1} \beta_0\right)^{-1} \left(\beta_0^T \Sigma^{-1} I_J\right)\right)^T \Sigma^{-1} I_J}$, where $\beta_1 = (\beta_{1,1}, \beta_{1,2}, \ldots, \beta_{1,J})^T$, $\beta_0 = (\beta_{0,1}, \beta_{0,2}, \ldots, \beta_{0,J})^T$.

In addition, a linear regression model after adjusting for the observed confounders (LR-AC model), fixed effect meta-analysis based on the LR-AC model (Meta-LR-AC model), NCOR model are used to estimate the lagged causal effect of $Y(0)$ on $Y(1)$.

i) LR-AC model: $y(1) = \lambda_0 + \lambda_1 x + \lambda_2 y(0) + \epsilon_\lambda$ for all data, where $\lambda_2$ stands for the association after adjusting for the observed confounders;

ii) Meta-LR-AC model: For any study or time series fragment $j$, we have
\[ y(1)_j = \hat{\lambda}_{0,j} + \hat{\lambda}_{1,j} x_j + \hat{\lambda}_{2,j} y(0)_j + \varepsilon_{1,j}. \]

Then, the fixed-effect model is used to estimate the effects of \( Y(0) \) on \( Y(1) \):

\[
\hat{\kappa}_1 = -1 = \frac{\beta_0 - I_j (I_j^T \Sigma_j^{-1} I_j)^{-1} I_j^T \Sigma_j^{-1} \beta_0}{\Sigma_j \beta_0}.
\]

iii) NCOR model:

In the first simulation study, we compare the biases, standard errors (SEs) and mean square errors (MSEs) of these models under different sample sizes, effect parameters and number of studies with 1000 simulation repetitions. In the second simulation study, the type I error rates and statistical power are used to evaluate the stability and power of the NCOR model when varying across the number of studies. The type I error rate occurs when the test rejects a null hypothesis that is actually true. The statistical power is defined as the probability that the test correctly rejects the null hypothesis when the alternative hypothesis is true. In the third simulation, we evaluate the sensitivity of the NCOR estimator by relaxing Assumption 2(b) and 3(b), respectively. For relaxed Assumption 2(b) (Web Figure 1 in Supplementary Material), we compare biases, SEs and MSEs of these models when varying across the effect of \( Y(0) \) on \( X \) with other parameters fixed. For Assumption 3(b), we do three sensitivity analyses. i) we change the size of \( \omega = c_3 - c_1 \) from 0 to 1 to observe the biases, SEs, and MSEs of these models. ii) we consider another model structures for the unmeasured confounder to observe the performance of the NCOR model (Web Figure 2 in Supplementary Material).

\[
U(1/2) = \gamma_1 U(0) + \varepsilon_{a2} \quad \varepsilon_{a2} \sim N(0, \sigma_a^2 - \gamma_1^2)
\]

\[
U(1) = \gamma_2 U(1/2) + \gamma_1 U(0) + \varepsilon_{a2} \quad \varepsilon_{a2} \sim N(0, \sigma_a^2 - \text{var}(\gamma_2 U(1/2) + \gamma_1 U(0)))
\]

iii) we change the size of \( \eta = r_2 - \hat{r}_1 \) to observe the difference of \( \eta \) on the causal effect estimation of \( X \) on \( Y(1) \) and the lagged causal effect estimation of \( Y(0) \) on \( Y(1) \) via the proposed NCOR model. The R language is used to conduct the statistical simulations.

Application
To illustrate the NCOR model in epidemiology studies, we assess the potential causal effect of fine particulate matter (PM$_{2.5}$) on the AMR and the lagged causal effect of AMR. Data for county-level PM$_{2.5}$ are derived from the Health Effects Institute 2019, State of Global Air 2019 (https://www.stateofglobalair.org/data/#/air/table). PM$_{2.5}$ is defined as fine particles with an aerodynamic diameter of less than or equal to 2.5μm. The adult mortality rate (probability of dying between 15 and 60 years per 1000 population) for each country is extracted from World Health Organization (https://www.who.int/data/gho).

Evidence from cohort studies indicate an association between increased AMR and long-term exposure to fine PM$_{2.5}$ pollution at reasonably low concentrations (<30 μg/m$^3$) [31-33]. However, ruling out confounding factors remains a great challenge. Web Figure 3 in Supplementary Material shows the realistic causal diagram of PM$_{2.5}$ and AMR. PM$_{2.5}$ represents the exposure in year (from 2010 to 2015). AMR$_{pre}$ and AMR$_{post}$ are all-cause adult mortality rates in year $-1$ and year $+1$, respectively. $U(0), U(1/2), U(1)$ represent unmeasured or unobserved confounders. Potential confounders in time series studies are time-varying variables such as complex climatological, atmospheric processes, volatile organic compounds, baseline and NOx emissions are known to vary spatially and have strong associations with PM$_{2.5}$ and AMR, however, they are often unmeasured [34-36]. We divide 183 countries into six studies according to the continent. To demonstrate the potential of the NCOR model, LR-SU, LR-SU-Meta, DID, DID-Meta and NCOR model are used to study the causal effect of PM$_{2.5}$ on AMR in the population worldwide. LR-AC, LR-AC-Meta and NCOR model are used to estimate the lagged causal effect of AMR$_{pre}$ on AMR$_{post}$.

Results

Simulation results

As for the causal effect of $X$ on $Y(l)$, our simulation results (Figure 2 and Web Figure 4-8 in Supplementary Material) demonstrate that in comparison to the LR-SU, LR-SU-Meta, DID...
and DID-Meta model, which show substantial biases, the NCOR model can get almost unbiased causal effect estimates. Furthermore, the SEs and MSEs of NCOR model are generally smaller than those of other models. Figure 2a illustrates that NCOR model can obtain unbiased causal effect estimations of $X$ on $Y(1)$ as the number of studies increases. Furthermore, among the four models (LR-SU, LR-SU-Meta, DID and DID-Meta) that show significant biases, DID and DID-Meta methods have less biases than LR-SU and LR-SU-Meta models. Figure 2b illustrates that the NCOR model achieve the highest precision among all models, followed by DID, LR-SU, DID-Meta, and LR-SU-Meta model. In addition, the MSEs of the proposed NCOR model perform better than the other models (Figure 2c). Robust performances of the NCOR model are also observed when varying across the effect of $U$ on $X$, $U(0)$ on $Y(0)$, $X$ on $Y(l)$ and sample size, respectively (Web Figure 4-7 in Supplementary Material, respectively). Specifically, when varying across the effect of $Y(0)$ on $Y(l)$ ($\delta$) with other parameters fixed (Web Figure 8 in Supplementary Material), the classical DID model and DID-Meta model could obtain unbiased causal effect estimations in the case of $\delta = 0$, that is, there is no lag effect of the pre-exposure outcome on the future outcome. Moreover, the biases monotonically increase with $\delta$ increasing.

Figure 3 indicates the type I error rates and statistical power of the NCOR model with the increasing number of studies. The rejection rates of the causal null hypothesis are controlled at a nominal 5% level for the NCOR models. In addition, the statistic power monotonically increases with the number of studies increasing.

Sensitivity analysis which relaxes Assumptions 2(b), 3(b) is performed to examine the robustness of NCOR model. When Assumption 2(b) is not satisfied ($Y(0)$ has effect on $X$), the SEs of NCOR model increase with the effect of $Y(0)$ on $X$ increasing, however, biases and MSEs are lower than those of other models (Web Figure 9 in Supplementary Material). When Assumption 3(b) is violated, Web Figure 10-12 in Supplementary Material shows that the SEs
and MSEs of the NCOR model are lower than those of other models with accepted biases.

For the lagged causal effect of $Y(0)$ on $Y(1)$, the simulation results (Web Figure 13-18 in Supplementary Material) indicate that NCOR model show unbiased causal effect estimations, whereas the LR-AC and LR-AC-Meta models show conspicuous biased estimations. The SEs and MSEs of the NCOR model are generally smaller than those of the other three models. Web Figure 19 in Supplementary Material reveals that type I error rates for NCOR model remain nominal ($\alpha = 0.05$). Furthermore, the power of the NCOR model monotonically increases with increasing numbers of studies. When Assumption 2(b) is not satisfied, the biased estimation of $Y(0)$ on $Y(1)$ can be observed when it varies across the effect of $Y(0)$ on $X$ (Web Figure 20 in Supplementary Material). When Assumption 3(b) is violated, the biases of NCOR model are lower than those of other models with acceptance SEs and MSEs (Web Figure 21-23 in Supplementary Material).

**Application results**

Previous studies showed that the number of deaths is positively associated with high levels of PM$_{2.5}$ [37-41]. Table 1 shows the causal effect estimations and the 95%CI of PM$_{2.5}$ on AMR through LR-SU, LR-SU-Meta, DID, DID-Meta, and NCOR models. LR model shows that the higher PM$_{2.5}$ leads to higher mortality with statistical significances in 2010 and 2011. However, there was no association between PM$_{2.5}$ and AMR in other years. When the fixed effect meta-analysis was used to estimate the causal association between PM$_{2.5}$ and AMR, AMR is not associated with PM$_{2.5}$. Finally, the other three models (DID, DID-Meta, NCOR models) reveal PM$_{2.5}$ has no causal association with AMR. The estimated associations in association analysis are most likely caused by unobserved confounders.

Table 2 shows the lagged causal effect estimations and the 95%CI of AMR$_{Year-1}$ on AMR$_{Year+1}$ via the LR-AC, LR-AC-Meta, and NCOR models. LR and Meta-LR-AC show positive association with statistical significances between AMR$_{Year-1}$ on AMR$_{Year+1}$. 


Nevertheless, the NCOR model reveal that AMR_{Year-1} has no causal association with AMR_{Year+1}.

The results indicate that the association between AMR_{Year-1} and AMR_{Year+1} is most likely caused by spurious associations due to unobserved confounders.

DISCUSSION

A significant challenge in observational studies is to control the potential confounders between exposure and outcome [1, 2]. The negative control analysis aims to identify the presence of residual confounders and further correct the unmeasured confounders [7, 19]. The DID method can be regarded as a negative outcome control approach through a monotonicity hypothesis of unmeasured confounding effects[9]. Theoretically, the formula 
\[
\frac{\partial E(Y(1) | X)}{\partial X} - \frac{\partial E(Y(0) | X)}{\partial X}
\]
can be used to estimate the causal effect of exposure on post-exposure outcomes. The causal effect estimation is \( c_0 + c_1c_2\delta \) and the bias is \( c_1c_2\delta \) because of the existence of a lagged causal effect of pre-exposure outcome on post-exposure outcome. The proof for the bias formula of the DID model is shown in Web Appendix 4 in Supplementary Material.

In this paper, we develop a new framework for the identification and inference of causal effects of exposure on outcome using a pre-exposure outcome as a negative control outcome. For multiple studies, the intercept term of the NCOR model is the causal effect estimation of exposure on outcome. Then, the slope coefficient of the NCOR model minus 1 is the lagged causal effect estimation of pre-exposure outcome on post-exposure outcome. Under feasible assumptions, unbiased causal effect estimation of the exposure on the outcome in the NCOR model can be obtained. Furthermore, a robust estimation can still be obtained when relaxed assumptions. Simulations suggest that the NCOR model could eliminate approximately all unmeasured confounders and obtain unbiased and robust causal effect estimations. In contrast, other models, such as the LR-SU, LR-SU-Meta, LR-AC and Meta-LR-AC, DID, DID-meta models can’t remove residual confounding.
In the context of severe episodes of poor air quality in the 20th century, such as the London Fog of 1952, researchers originally recognized the correlation between particulate pollution and all-cause mortality [42]. These episodes clearly indicate that the increase in deaths is associated with high levels of particulate matter. Time-series studies can be confounded by time-varying factors such as influenza epidemics and temperature; however, statistical methods have been developed to reduce such unmeasured confounding [37-41]. In this paper, linear regression model, fixed effect meta-analysis, DID, DID-Meta and NCOR models are used to estimate the causal effect of PM$_{2.5}$ on all-cause mortality rates and the lagged causal effect of all-cause mortality rates. The results of the proposed model indicated that there was no causal effect of pre-exposure AMR on post-exposure AMR. Therefore, both the proposed and DID models could obtain an unbiased causal effect of PM$_{2.5}$ on AMR. As expected, DID, DID-Meta and NCOR models reveal PM$_{3.5}$ has no causal association with AMR, whereas the LR model showed an inconsistent conclusion. Therefore, the estimated associations in association analysis are most likely caused by unobserved confounders. In conclusion, our method can remove approximately all confounders using a pre-exposure outcome as a negative control outcome.

**Conclusions**

The proposed NCOR model can obtain unbiased and robust causal effect estimation of the exposure on the outcome, along with the lagged causal effect of outcomes. The application results demonstrate that the previously estimated association between PM$_{2.5}$ and all-cause mortality rates can be attributed to the unobserved confounding factors. Furthermore, the NCOR model reveal that pre all-cause mortality rates has no causal association with post all-cause mortality rates.

**Abbreviations List**

AMR, All-cause Mortality Rate; MSE, Mean Square Error; NCOR, Negative Control Outcome
Declarations

Ethics approval and materials
Not applicable

Consent for publication
Not applicable

Availability of data and material
Not applicable

Competing interests
The authors declare that they have no competing interests.

Funding
This work was supported by the National Natural Science Foundation of China (Grant 81773547, 82003557), Natural Science Foundation of Shandong Province (ZR2019ZD02) and Shandong Provincial Key Research and Development project (2018CXGC1210). The funding body had no role in the design, methods, data collection, analysis, data interpretation, or writing of the manuscript.

Authors' contributions
FX, YY and HL jointly conceived the main idea behind the article and designed the study. YY helped conduct the literature review, performed the simulation and prepared the first draft of the manuscript. LH, XS, XL, FY, YY and QW participated in the design of the study. FX
advised on critical revision of the manuscript for important intellectual content. All authors revising it critically for important intellectual content, read and approved the final manuscript.

Acknowledgments

We would like to thank the Health Effects Institute 2019, State of Global Air 2019 and World Health Organization to provide adult mortality rate data and county-level PM$_{2.5}$ data.

References

1. Streeter AJ, Lin NX, Crathorne L, Haasova M, Hyde C, Melzer D, Henley WE: Adjusting for unmeasured confounding in nonrandomized longitudinal studies: a methodological review. J Clin Epidemiol 2017, 87:23-34.

2. Groenwold RHH, Sterne JAC, Lawlor DA, Moons KGM, Hoes AW, Tilling K: Sensitivity analysis for the effects of multiple unmeasured confounders. Ann Epidemiol 2016, 26(9):605-611.

3. Richardson DB, Laurier D, Schubauer-Berigan MK, Tchetgen E, Cole SR: Assessment and Indirect Adjustment for Confounding by Smoking in Cohort Studies Using Relative Hazards Models. Am J Epidemiol 2014, 180(9):933-940.

4. Lipsitch M, Tchetgen TE, Cohen T: Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies (vol 21, pg 383, 2010). Epidemiology 2010, 21(4):589.

5. Tchetgen ET: The Control Outcome Calibration Approach for Causal Inference With Unobserved Confounding. Am J Epidemiol 2014, 179(5):633-640.

6. Miao W, Tchetgen ET: A Confounding Bridge Approach for Double Negative Control Inference on Causal Effects (Supplement and Sample Codes are included). 2018.
7. Richardson DB, Keil AP, Tchetgen ET, Cooper G: Negative Control Outcomes and the Analysis of Standardized Mortality Ratios. Epidemiology 2015, 26(5):727-732.

8. Richardson DB: Occupational Exposures and Lung Cancer Adjustment for Unmeasured Confounding by Smoking. Epidemiology 2010, 21(2):181-186.

9. Sofer T, Richardson DB, Colicino E, Schwartz J, Tchetgen Tchetgen EJ: On Negative Outcome Control of Unobserved Confounding as a Generalization of Difference-in-Differences. Stat Sci 2016, 31(3):348-361.

10. Li H, Graham DJ, Majumdar A: The effects of congestion charging on road traffic casualties: A causal analysis using difference-in-difference estimation. Accident Anal Prev 2012, 49(SI):366-377.

11. Lechner M: The Estimation of Causal Effects by Difference-in-Difference Methods. Foundations and Trends in Econometrics 2011, 4(3):165-224.

12. Puhani PA: The treatment effect, the cross difference, and the interaction term in nonlinear "difference-in-differences" models. Econ Lett 2012, 115(1):85-87.

13. ROSENBAUM PR: The Role of Known Effects in Observational Studies. Biometrics 1989, 45(2):557-569.

14. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D: Interpreting observational studies: why empirical calibration is needed to correct p-values. Stat Med 2014, 33(2):209-218.

15. Shi X, Miao W, Nelson JC, Tchetgen Tchetgen EJ: Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. J R Stat Soc B 2020, 82(2):521-540.
16. Shi X, Miao W, Tchetgen ET: A Selective Review of Negative Control Methods in Epidemiology. CURRENT EPIDEMIOLOGY REPORTS 2020, 7(4):190-202.

17. Flanders WD, Klein M, Darrow LA, Strickland MJ, Sarnat SE, Sarnat JA, Waller LA, Winquist A, Tolbert PE: A Method for Detection of Residual Confounding in Time-series and Other Observational Studies. Epidemiology 2011, 22(1):59-67.

18. Flanders WD, Strickland MJ, Klein M: A New Method for Partial Correction of Residual Confounding in Time-Series and Other Observational Studies. Am J Epidemiol 2017, 185(10):941-949.

19. Miao W, Tchetgen ET: Invited Commentary: Bias Attenuation and Identification of Causal Effects With Multiple Negative Controls. Am J Epidemiol 2017, 185(10):950-953.

20. Yu Y, Li H, Sun X, Liu X, Yang F, Hou L, Liu L, Yan R, Yu Y, Jing M et al: Identification and Estimation of Causal Effects Using a Negative-Control Exposure in Time-Series Studies With Applications to Environmental Epidemiology. Am J Epidemiol 2021, 190(3):468-476.

21. Pearl J: The Do-Calculus Revisited. Computer Science 2012.

22. Pearl J: Causal Diagrams for Empirical Research. Biometrika 1995, 82(4):669.

23. Pearl J: Causal inference in statistics: An overview. Statistics Surveys 2009, 3:96-146.

24. Pearl J: The Deductive Approach to Causal Inference. Journal of Causal Inference 2014, 2(2):115-129.

25. Copas JB, Malley PF: A robust P-value for treatment effect in meta-analysis with publication bias. Stat Med 2008, 27(21):4267-4278.

26. Bowden J, Smith GD, Burgess S: Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol 2015, 44(2):512-525.
27. Grant AJ, Burgess S: An efficient and robust approach to Mendelian randomization with measured pleiotropic effects in a high-dimensional setting. 2019.

28. Lee CH, Cook S, Lee JS, Han B: Comparison of Two Meta-Analysis Methods: Inverse-Variance-Weighted Average and Weighted Sum of Z-Scores. Genomics & Informatics 2016, 14(4):173.

29. Bowden J, Davey Smith G, Haycock PC, Burgess S: Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genet Epidemiol 2016, 40(4):304-314.

30. Burgess S, Bowden J, Dudbridge F, Thompson SG: Robust instrumental variable methods using multiple candidate instruments with application to Mendelian randomization. 2016.

31. Li T, Zhang Y, Wang J, Xu D, Yin Z, Chen H, Lv Y, Luo J, Zeng Y, Liu Y et al: All-cause mortality risk associated with long-term exposure to ambient PM2.5 in China: a cohort study. LANCET PUBLIC HEALTH 2018, 3(10):E470-E477.

32. Hoek G, Krishnan RM, Beelen R, Peters A, Ostro B, Brunekreef B, Kaufman JD: Long-term air pollution exposure and cardio-respiratory mortality: a review. Environ Health-Glob 2013, 12(43).

33. Cesaroni G, Badaloni C, Gariazzo C, Stafoggia M, Sozzi R, Davoli M, Forastiere F: Long-Term Exposure to Urban Air Pollution and Mortality in a Cohort of More than a Million Adults in Rome. Environ Health Persp 2013, 121(3):324-331.

34. Papadogeorgou G, Choirat C, Zigler CM: Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching. Biostatistics 2019, 20(2):256-272.

35. Greven S, Dominici F, Zeger S: An Approach to the Estimation of Chronic Air Pollution
Effects Using Spatio-Temporal Information. J Am Stat Assoc 2011, 106(494):396-406.

36. Pun VC, Kazemiparkouhi F, Manjourides J, Suh HH: Long-Term PM2.5 Exposure and Respiratory, Cancer, and Cardiovascular Mortality in Older US Adults. Am J Epidemiol 2017, 186(8):961-969.

37. Lin Y, Zou J, Yang W, Li C: A Review of Recent Advances in Research on PM2.5 in China. Int J Env Res Pub He 2018, 15(3):438.

38. Apte JS, Marshall JD, Cohen AJ, Brauer M: Addressing Global Mortality from Ambient PM2.5. Environ Sci Technol 2015, 49(13):8057-8066.

39. Sanders NJ, Barreca AI, Neidell MJ: Estimating Causal Effects of Particulate Matter Regulation on Mortality. Epidemiology 2020, 31(2):160-167.

40. Kloog I, Ridgway B, Koutrakis P, Coull BA, Schwartz JD: Long- and Short-Term Exposure to PM2.5 and Mortality. Epidemiology 2013, 24(4):555-561.

41. Cao Q, Rui G, Liang Y: Study on PM2.5 pollution and the mortality due to lung cancer in China based on geographic weighted regression model. Bmc Public Health 2018, 18(1).

42. Balmes JR: Do We Really Need Another Time-Series Study of the PM2.5-Mortality Association? New Engl J Med 2019, 381(8):774-776.

**Figure legend**

**Figure 1.** Causal diagram for Difference in Difference (DID) model (a). \( X \) is the exposure and \( Y(0) \) and \( Y(1) \) are the pre- and post-exposure outcome, respectively. \( U \) represents the unobserved confounders. Causal diagram for Negative Control Outcome Regression (NCOR) (b). \( U(0), U(1/2), U(1) \) denote the unobserved confounding factors at times \( t=0, 1/2, 1 \).
respectively. $X$ represents the exposure. $Y(0)$ and $Y(1)$ represents the outcome at time $t=0, 1$, respectively. $\{c_0, c_1, c_2, c_3, \gamma_1, \gamma_2\}$ are the relevant effect parameters.

**Figure 2** Simulation results of the bias (a), Standard Error, SE (b) and Mean Square Error, MSE (c) for causal effect of $X$ on $Y(1)$ when the number of studies change from 5 to 200. LR-SU represents linear regression model conditional on the surrogate of unmeasured confounders, the pre-exposure outcome. LR-SU-Meta denotes the fixed effect meta-analysis based on LR-SU model. DID model is Difference in Difference model and DID-Meta model is the fixed effect meta-analysis based on DID model. NCOR is Negative Control Outcome Regression model.

**Figure 3** Type I error (a) and statistic power (b) of NCOR model for causal effect estimation of $X$ on $Y(1)$ with the number of studies increasing.
Table 1. The causal effect estimations and its 95%CI of PM$_{2.5}$ on AMR

| Year | LR-SU       | LR-SU-Meta  | DID         | DID-Meta    | NCOR        |
|------|-------------|-------------|-------------|-------------|-------------|
| 2010 | 0.137(0.043, 0.231) | 0.107(-0.184, 0.397) | 0.025(-0.096, 0.146) | 0.052(-0.246, 0.350) | 0.084(-0.137, 0.304) |
| 2011 | 0.349(0.082, 0.616)  | 0.095(-0.150, 0.340)  | 0.138(-0.170, 0.446) | 0.014(-0.234, 0.262) | -0.011(-0.682, 0.661) |
| 2012 | 0.131(-0.024, 0.285) | 0.019(-0.208, 0.246)  | 0.046(-0.110, 0.202) | -0.011(-0.233, 0.212) | -0.007(-0.285, 0.272) |
| 2013 | 0.053(-0.070, 0.177) | -0.015(-0.221, 0.191) | 0.015(-0.106, 0.136) | -0.043(-0.249, 0.164) | -0.049(-0.126, 0.027) |
| 2014 | 0.064(-0.009, 0.137) | 0.016(-0.194, 0.226)  | -0.013(-0.093, 0.066) | -0.014(-0.226, 0.197) | 0.028(-0.057, 0.113)  |
| 2015 | 0.062(-0.008, 0.133) | 0.003(-0.192, 0.197)  | -0.050(-0.134, 0.033) | -0.037(-0.234, 0.161) | 0.044(-0.141, 0.229)  |

LR-SU represents linear regression model conditional on the surrogate of unmeasured confounders, the pre-exposure outcome. LR-SU-Meta denotes the fixed effect meta-analysis based on LR-SU model. DID represents Difference in difference model and DID-Meta denotes the fixed effect meta-analysis based on DID model. NCOR is Negative Control Outcome Regression model.

Table 2. The lagged causal effect estimations and its 95%CI of AMR$_{Year-1}$ on AMR$_{Year+1}$

| Year | LR-AC       | Meta-LR-AC  | NCOR       |
|------|-------------|-------------|------------|
| 2010 | 0.910(0.895, 0.926) | 0.940(0.811, 1.068) | -0.018(-0.210, 0.175) |
| 2011 | 0.820(0.777, 0.863) | 0.870(0.708, 1.033) | 0.168(-0.545, 0.881) |
| 2012 | 0.940(0.912, 0.969) | 0.968(0.853, 1.084) | 0.008(-0.223, 0.238) |
| 2013 | 0.972(0.950, 0.995) | 0.973(0.867, 1.079) | 0.005(-0.057, 0.068) |
| 2014 | 0.952(0.938, 0.965) | 0.963(0.860, 1.066) | -0.031(-0.101, 0.039) |
| 2015 | 0.925(0.910, 0.939) | 0.952(0.852, 1.052) | -0.060(-0.222, 0.103) |

LR-SU represents linear regression model conditional on the surrogate of unmeasured confounders, the pre-exposure outcome. Meta-LR-SU denotes the fixed effect meta-analysis based on LR-SU model. NCOR is Negative Control Outcome Regression model.
Figures

(a) Causal diagram for Difference in Difference (DID) model (a). X is the exposure and and are the pre- and post-exposure outcome, respectively. U represents the unobserved confounders. Causal diagram for Negative Control Outcome Regression (NCOR) (b). U(0), U(1/2), U(1) denote the unobserved confounding factors at times t=0, 1/2, 1 respectively. X represents the exposure. Y(0) and Y(1) represents the outcome at time t=0, 1, respectively. \{c_0, c_1, c_2, c_3, \gamma_1, \gamma_2\} are the relevant effect parameters.

Figure 1

Simulation results of the bias (a), Standard Error, SE (b) and Mean Square Error, MSE (c) for causal effect of X on Y(1) when the number of studies change from 5 to 200. LR-SU represents linear regression model conditional on the surrogate of unmeasured confounders, the pre-exposure outcome. LR-SU-Meta denotes the fixed effect meta-analysis based on LR-SU model. DID model is Difference in Difference model and DID-Meta model is the fixed effect meta-analysis based on DID model. NCOR is Negative Control Outcome Regression model.

Figure 2
Figure 3

Type I error (a) and statistic power (b) of NCOR model for causal effect estimation of X on Y(1) with the number of studies increasing.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx