Distribution Agreement

Signature:

_______________________________ _______________

Xinyue Chen Date

Using Negative Exposures to Partially Control for Unmeasured Confounders in Time-Series Analysis of Air Pollution and Health

By

Xinyue Chen

Master of Public Health

Department of Biostatistics and Bioinformatics

____

Howard Chang, PhD

(Thesis Advisor)

____

Stefanie Ebelt, PhD

(Reader)

Using Negative Exposures to Partially Control for Unmeasured Confounders in Time-Series Analysis of Air Pollution and Health

By

Xinyue Chen

B.S., Wuhan University, 2020

Thesis Committee Chair: Howard Chang, PhD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of Master of Public Health

in Department of Biostatistics and Bioinformatics

2023

Abstract

Abstract

Using Negative Exposures to Partially Control for Unmeasured Confounders in Time-Series Analysis of Air Pollution and Health

By Xinyue Chen

Background: Observational studies face the challenge of confounding. Residual confounding may persist if known confounders are not properly measured or if unknown confounders are present. A regression-based method was proposed to directly reduce residual confounding in observational studies based on negative control exposures.

Objectives: The objective of this study is to investigate additional forms of negative exposure controls, including those that are lagged and the use of multiple negative exposures, as methods for reducing residual confounding by unmeasured or mis-measured confounders in time-series studies. These methods were evaluated in a time-series analysis of daily air pollution and asthma emergency department visits, as well as in simulation studies.

Methods: The study employed a log-linear model to estimate short-term effects of air pollution and observed counts of asthma ED visits using Poisson regression with overdispersion based on observed ozone level data during the study period. Model misspecifications were intentionally created by omitting one of the measured confounders and reducing the degree of freedom of nature cubic spline term for calendar time. We also fitted the model by adding the negative exposure, by adding each of the four negative exposures one-at-a-time, or combining a lagged negative exposure with the future negative exposure. Bias, root mean square error (RMSE), and coverage of the 95% confidence interval were calculated.

Results: Simulation studies showed that adding the negative control exposure can reduce bias in some scenarios. The standard error associated with estimated log relative risks for ozone is negligibly different when comparing the inclusion of negative control exposure versus excluding it. In some scenarios, the negative control exposure slightly increased coverage.

Conclusions: In simulation studies, we have shown that in estimating short-term health effects of air pollution with Poisson log-linear time-series model, the bias resulting with unmeasured confounders can be smaller when negative control exposure is included the model. In some cases, we also find that the use of multiple negative controls can further reduce bias. But this reduction of bias is not guaranteed.

Using Negative Exposures to Partially Control for Unmeasured Confounders in Time-Series Analysis of Air Pollution and Health

By

Xinyue Chen

B.S., Wuhan University, 2020

Thesis Committee Chair: Howard Chang, PhD

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of Master of Public Health

in Department of Biostatistics and Bioinformatics

2023

Table of Contents

1. Introduction

Confounding poses a crucial challenge to the validity of observational studies as it involves the influence of external variables on the effect of a primary exposure of interest, leading to a distortion in the observed exposure-outcome association of interest (Schriger, 2008). When a causal graph accurately depicts the causal relationships, the existence of particular causal patterns (confounding paths) can be discerned (Greenland et al.). Adequate stratification or covariate effect modeling can be employed to block the confounding path and control for confounding analytically. Nonetheless, confounding that persists even after these measures are taken is referred to as residual confounding. Residual confounding may arise if known confounders are not properly measured or if their form is mis-specified. It can also occur if unknown confounders are present (Fewell et al.). To assess the potential impact of residual confounding, sensitivity analyses can be important (Goodman et al.; Greenland, 2005).

Time-series studies examining the short-term health effects of environmental exposures are concerned about confounding by temporal factors, and frequently control for confounding by incorporating covariates (e.g., day of the week, temperature, and humidity) into the model. Parametric splines or other time-related factors are also included to adjust for unmeasured factors that vary over time. Previously, Flanders et al. demonstrated how to test for residual confounding or other model misspecifications by considering temporality and causal relationships of the time-series study design. The test involves incorporating a residual-confounding indicator variable into the health model that satisfies two properties: 1) the indicator must be independent of the disease in a correctly specified model, meaning that it neither causes the disease nor is caused by it (after controlling for the effects of controlled factors); 2) the indicator should be linked to the exposure of interest and, similar to the exposure, to unmeasured confounders (Flanders et al., 2011a, 2011b; Flanders et al., 2009). A regression-based method was also proposed by Flanders et al. to directly reduce residual confounding in observational studies based on negative control exposures. The main result shows that adding a negative control exposure to the model is expected to reduce residual confounding, under certain assumptions. The method can be used as the basis for sensitivity analyses to partially correct for bias (Flanders et al., 2017). As a commentary, Miao and Tchetgen proposed two more methods to further reduce the residual confounding (Miao & Tchetgen Tchetgen, 2017).

The objective of this study is to investigate additional forms of negative exposure controls, including those that are lagged and the use of multiple negative exposures, as methods for reducing residual confounding by unmeasured or mismeasured confounders in time-series studies. These methods were evaluated in this in a time-series analysis of daily air pollution and asthma emergency department visits, as well as simulation studies.

2. Methods

2.1 Data

2.1.1 Asthma emergency department data

Patient level daily emergency department (ED) visit data for asthma and wheeze were collected from individual hospitals for the years 1998 to 2009, and from the Georgia Hospital Association for the years 2010-2020. International Classification of Diseases (ICD) 9th revision (ICD-9) diagnosis codes were used for ED visits prior to October 1, 2015, followed by the use of ICD 10th revision (ICD-10) codes. We identified ED visits using primary diagnosis code for asthma (ICD-9: 493, ICD-10: J45, J46). We aggregated ED visits across the 20-county Atlanta metropolitan area based on patient residual ZIP codes.

2.1.2 Ozone and meteorological data

Daily meteorological data were obtained from the National Centers for Environmental Information. We used measurements from the automated surface observing station located at the Atlanta Hartsfield International Airport, which included daily average temperature and dew-point temperature from 1998 to 2011.

Daily Ozone exposure was derived from monitors in the US Environmental Protection Agency's Air Quality System and from the Aerosol Research Inhalation Epidemiology Study. A total of 14 ozone monitors were included in the Atlanta study area and hourly measurements were used to calculate daily averaged 8-hour maximum for the study period. Daily ozone exposure for the Atlanta study region was generated by a population-weighted average based on ZIP code population data from the US Census.

2.2 Statistical Analysis

2.2.1 Time-Series Analysis of Ozone and Emergency Department Visit

We first describe the log-linear model typically employed in estimation short-term effects of air pollution. We modeled the observed counts asthma ED visits using Poisson regression with overdispersion:

$$\log(E[Y_t]) = \beta_0 + \beta_1 x_t + \sum_{k=1}^{p} \beta_k C_{kt} \qquad (1)$$

where X_t is the 2-day lagged ozone concentration, and C_{kt} is the set of confounders including, indicator of date of week, 3-day (lags 0, 1, and 2) moving averages of daily average temperature and daily average dewpoint-temperature and indicators for federal holiday. A natural cubic spline for calendar dates with monthly knots were also used in the model to control for the long-term time trend and seasonality. Non-linear effects of meteorology were modeled with natural cubic splines with 6 degrees of freedom. We chose 2-day lagged ozone concentration as the primary exposure of interest based on preliminary analyses. The 2-day lagged exposure had the strongest association with asthma ED visits compared to other lags.

2.2.2 Proposed Negative Exposure Indicators for Unmeasured Confounder Adjustment

The directed acyclic diagram (Greenland et al.) below is used to illustrate assumed casual relationships.

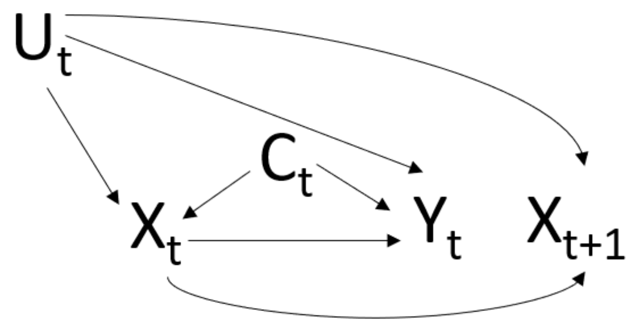


Figure 1. Directed acyclic graph summarizing assumed causal relationships.

The ED visits on day t is the health event of interest (Y_t) and ozone level on day t is the exposure of interest (X_t). C_t represents the set of measured confounders on day t (e.g., meteorological factors). U_t is defined as the set of unmeasured, sometimes unrecognized, confounders associated with X_t and is a cause of Y_t; U_t is the component we are concerned the most. We assumed that unmeasured confounder is not only related to the exposure on the same day t (X_t) but also associated with exposure on the next day (X_{t+1}). We further assume that the impact of measurement error and functional form misspecification is negligible throughout the analysis. Therefore, any potential model "misspecification" is primarily caused by the omission of confounders U_t. We refer to X_{t+1} as a negative exposure indicator because it does not cause Y_t . Any observed association between Y_t and X_{t+1} is due to omission of U_t. Previous work (Flanders et al., 2017) has shown that including negative exposure in the health model can partially correct for bias in the association between Y_t and X_t.

In this work, we further consider 3 additional negative exposures based on the historical ozone level: (1) ozone level at 7-day lag, (2) ozone level at 14-day lag, (3) the moving average of ozone levels from lag 7 to lag 14. Furthermore, the combinations of future ozone level and historical ozone level: (4) the combination of future 1-day ozone level and ozone level at 7-day lag, (5) the combination of future 1-day ozone level and ozone level at 14-day lag, and (6) the combination of future 1-day ozone level and the moving average from lag 7 to lag 14. The use of these lagged ozone levels assumes that there is no ozone effect beyond the 6^{th} day lag.

3. Simulation Study

Simulation studies were used to evaluate the performance of using negative exposure to partially correct for unmeasured confounding. To make the simulations more realistic, we generated simulated counts of daily asthma ED visits from a Poisson distribution with means given by Equation (1) with the regression parameters estimated from the Atlanta data. We set the true ozone log relative risk (β_1) to be 0.0236.

We then fitted models on the simulated data omitting one of the measured confounders to intentionally create model misspecification. Another type of model misspecification we generated is by reducing the degree of freedom of nature cubic spline term for calendar time. Since the study period is from 1998-2009, the true model has 132 knots. We replaced 132 with 12, 24, 36, 48, 60, 72, 84, 96, 108, and 120 as misspecification. We also fitted the model by adding the negative exposure, by adding each of the four negative exposures one-at-a-time, or combining a lagged negative exposure with the future negative exposure. Each model mis-specification scenario was run with 200 simulated datasets. Bias was calculated by taking the mean of the difference between β_1 and the estimated value. Root mean square error (RMSE) and coverage of the 95% confidence interval were also calculated.

4. Results

4.1 Descriptive analyses

Table 1 presents a summary of the daily ED visits for asthma, ozone exposure levels and meteorological factors observed during the study period. The daily 8-hour maximum population-average ozone exposure had an average concentration of 42.90 parts per billion (ppb). The dew-point temperature and average temperature had an average magnitude of 50.10 Fahrenheit and 62.42 Fahrenheit, respectively. The average number of asthma ED visits was 68.26 per day during the study period.

Figure 2 shows the daily asthma ED visits between 1993 and 2020. Asthma ED visits show clear seasonality with peaks during the winter months. There was an increasing trend in ED visits from 1993 to 2006, followed by a plateau from 2008 to 2020. A decrease in asthma ED visits in 2020 was observed, which could be attributed to the COVID-19 pandemic. Figure 3 shows the daily ozone levels during the study period. Ozone levels were higher during the summer season.

4.2 Results of Real data analysis

Figure 4 shows that there is a robust positive association between daily ozone and ED visits for asthma in Atlanta [RR = 1.024 (95% CI: 1.010, 1.038) per 26.97 per ppb increase]. The use of future exposure, 2-week lag exposure or moving average of 1-week to 2-week lag exposure resulted in a slight attenuation of the relative risk [RR = 1.023 (95% CI: 1.009, 1.037), RR = 1.024 (95% CI: 1.010, 1.0371), RR = 1.023 (1.009, 1.037)]. The use of 1-week lag exposure resulted in a slight increasing of the relative risk [RR = 1.024 (95% CI: 1.011, 1.038)].

4.3 Results of Simulation studies

The biases under different scenarios of missing a specific confounders or model mis-specification of the true temporal trends are shown in Table 2. A small bias in the estimated β_1 was introduced by dropping weekday indicators, average temperature, dew-point temperature and the holiday indicator. And the bias caused by wrongly specified knots of natural cubic spline term of calendar time is more severe when smaller degrees of free was used compared to the true model.

The future exposure reduced bias when indicator of date of week, nature cubic spline term of dew-point temperature or indicator of federal holiday was omitted from the model or the knots of nature cubic spline term of calendar dates are 0.

The group of lagged exposure (1-week lag, 2-week lag and the moving average of 1-week to 2-week lag) reduced bias when the natural cubic spline term of average temperature and the knots of the natural cubic spline term are 0, 36, 72 and 108. The group of lagged exposure reduced bias in the scenarios that the future spline didn't reduce bias. And both lagged exposures and future exposure reduced bias when the knots of nature cubic spline term of calendar dates are 0.

The combination of future exposure and 1-week lag exposure reduced bias when date of week indicator, nature cubic spline term of average temperature, nature cubic spline term of dew-point temperature or federal holiday indicator was omitted and the knots of the nature cubic term of calendar dates are 0, 36, 72 and 108, respectively. The reduction happened in three more scenarios that the nature cubic spline term of average temperature was omitted and the knots of nature cubic spline of calendar dates are 36 and 72 compared with only adding the future exposure. When the knots of nature spline term for calendar dates are 0, the further reduction of bias happened.

The combination of future exposure and 2-week lag exposure reduced bias when date of week indicator or federal holiday indicator was omitted and the knots of the nature cubic term of calendar dates are 0 and 36, respectively. When the knots of nature spline term for calendar dates are 0, the further reduction of bias also happened.

The combination of future exposure and moving average of 1-week to 2-week lag exposure reduced bias when date of week indicator or federal holiday indicator was omitted and the knots of the nature cubic term of calendar dates are 0, 36 and 72, respectively. When the knots of nature spline term for calendar dates are 0, the further reduction of bias also happened.

Table 3 shows the standard error associated with the estimated log relative risks for ozone in various scenarios, and there is negligible difference in the standard error when comparing the inclusion of negative control exposure versus excluding it. This suggests that the negative control exposure does not have an adverse impact on the precision of the results. Table 4 displays coverage under different scenarios. There are no significant differences between adding negative control exposure and not adding negative control exposure when the nature cubic spline term for average temperature or federal holiday indicator was omitted and knots of nature cubic spline term for calendar dates is 0. When weekday indicators were omitted, the coverage was slightly increase when the negative control exposure of next-day ozone level was included into the regression model. When the degrees of freedom for temporal trend was much smaller than the true model, there was a slight increase in coverage when the negative control exposure of ozone level at 7-day lag (X_{t-7}), ozone level at 14-day lag (X_{t-14}), or the moving average of ozone levels from lag 7 to lag 14.

5. Discussion

In simulation studies, we have shown that in estimating short-term health effects of air pollution with Poisson log-linear time-series model, the bias resulting with unmeasured confounders can be smaller when negative control exposure is included the model. In some cases, we also find that the use of multiple negative controls can further reduce bias. But this reduction of bias is not guaranteed. This may be due to the violation of two important conditions: (1) on average, when U_t increases, X_t and X_{t+1} tend to change in the same direction and (2) X_t is positively associated with X_{t+1} that mentioned in previous work (Flanders et al., 2017).

There are additional methods for incorporating negative exposures that warrant exploration in the future. Specifically, Miao and E (2017) proposed two additional techniques to minimize bias in time-series studies by using a more intricate Directed Acyclic Graph (DAG) (Miao & Tchetgen Tchetgen, 2017). The first method involves employing a forward-in-time regression model to establish the relationship between present confounding factors and future exposures. Although this approach can still decrease bias when there is a positive correlation between confounders and future exposures, it is not guaranteed to do so when the correlation is negative. These findings can be applied to models that use a log link for the outcome and a location-shift model for the error distribution of the confounder. The second technique involves a nonparametric method that utilizes both negative control outcomes and exposures to identify the causal effect without the need for additional parametric assumptions (Miao et al., 2018; Miao et al., 2020).

The previous studies did not provide specific information about the degree of reduction in bias that would be considered meaningful, nor did they discuss the criteria for defining a negative control exposure as "good". Therefore, further research and exploration are needed to fully understand and address these issues in this area.

Table 1. Summary Statistics for daily asthma emergency department visits, ozone exposure, and meteorology in Atlanta, 1993 to 2021

Variable	Mean	SD	Median	Min	Max	IQR
Criteria Gas						
8-hour Max Ozone (ppb)	42.90	18.85	40.41	2.53	117.86	26.97
Meteorology						
Dew-point temperature (Fahrenheit)	50.10	16.27	53.50	-1.50	74.40	26.50
Mean temperature (Fahrenheit)	62.42	14.33	64.35	17.30	91.00	23.50
Asthma ED Visits						
Asthma, primary diagnosis, daily visit counts	68.26	28.64	63.00	15.00	205.00	40.00

Figure 2. Emergency Department Visits for Asthma in Atlanta, 1993 to 2021

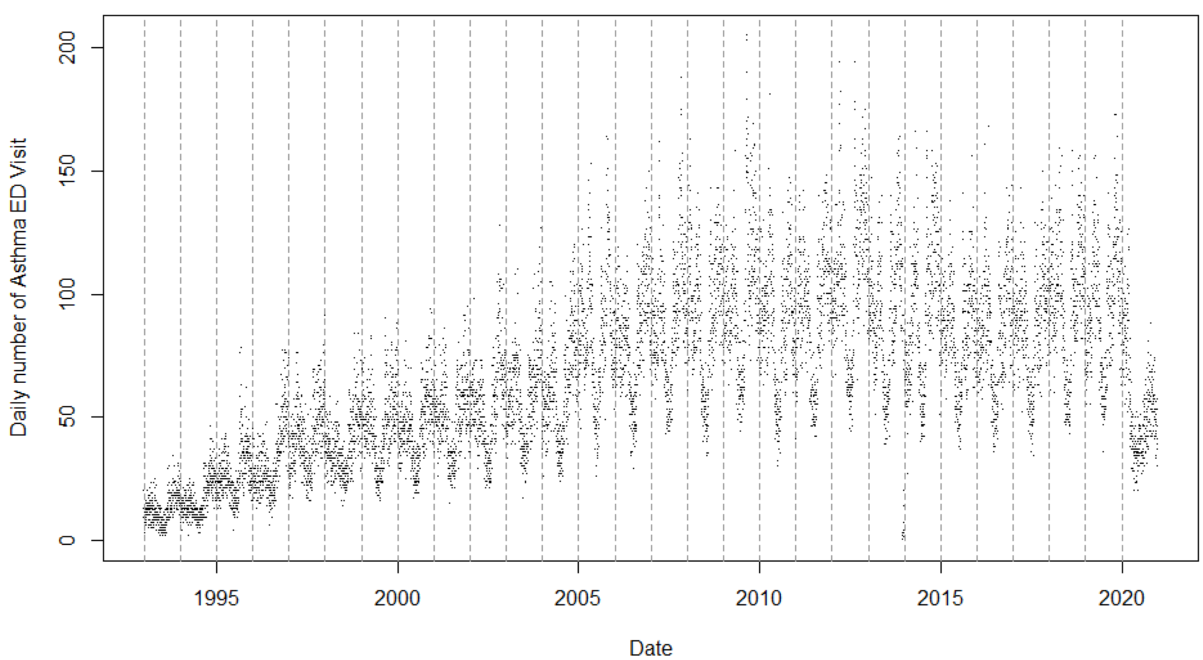


Figure 3. Daily Eight-hour Maximum Population-Averaged Ozone Level in Atlanta, 1998 to 2011

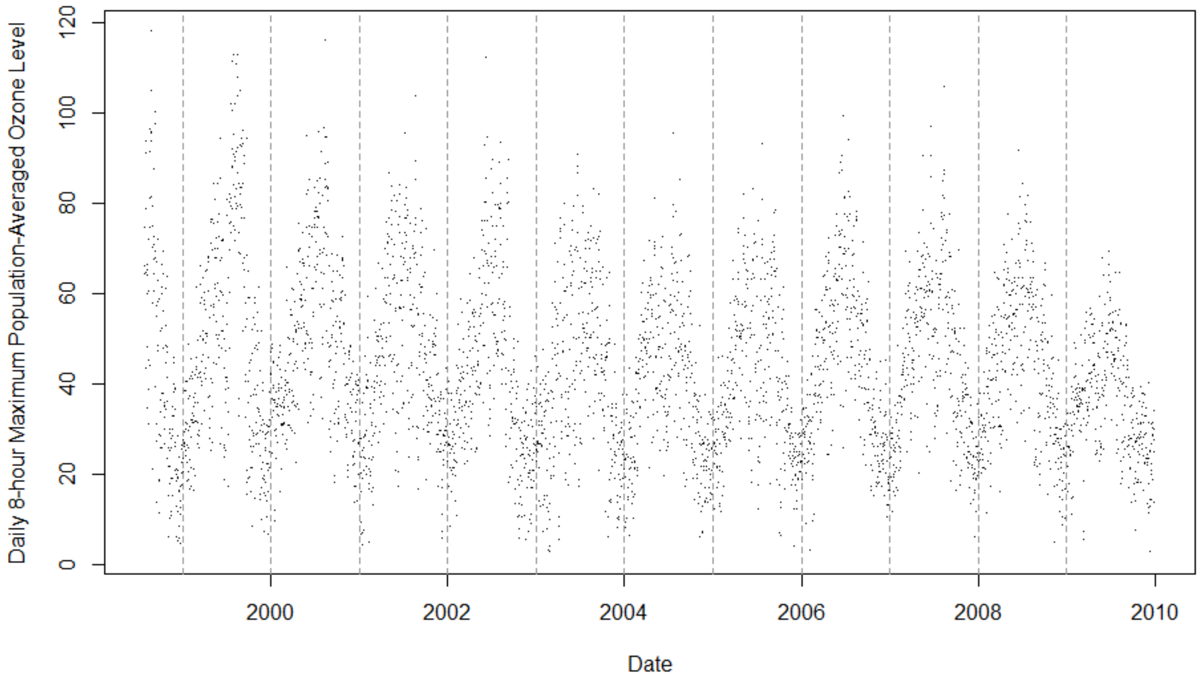


Table 2. Comparing Bias of Ozone Effect on Asthma ED Visits with and without Negative Control Exposure, under Varying Patterns of Misspecification

		No NCE	With negative exposures: future exposure (F), 1-week lag (LW1), 2-week lag (LW2), and moving average of 1-week to 2-week lag (MA)						
Scenario			F	LW1	LW2	MA	F+LW1	F+LW2	F + MA
1	True model	-0.057	-0.050	-0.057	-0.057	-0.051	-0.050	-0.050	-0.043
2	Omit weekday indicator	0.629	0.580	0.633	0.649	0.646	0.585	0.599	0.593
3	Omit average temperature	0.722	0.727	0.715	0.721	0.714	0.721	0.727	0.723
4	Omit dew point temperature	0.400	0.391	0.400	0.409	0.409	0.393	0.401	0.407
5	Omit holiday indicator	-0.068	-0.063	-0.067	-0.069	-0.059	-0.062	-0.064	-0.051
	Omit time spline								
6	Knots=0	-21.853	-19.252	-19.711	-20.400	-18.238	-17.989	-18.313	-16.971
7	Knots=36	2.729	2.802	2.473	2.559	2.263	2.542	2.628	2.324
8	Knots=72	1.371	1.445	1.291	1.301	1.265	1.363	1.371	1.337
9	Knots=108	0.355	0.391	0.334	0.323	0.337	0.372	0.356	0.375

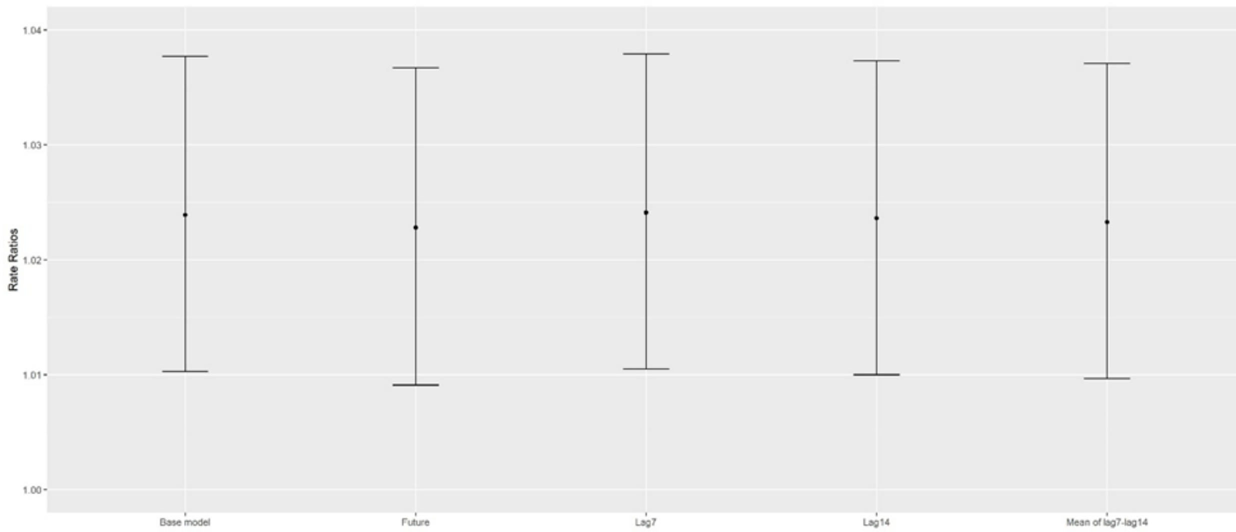
Table 3. Comparing Standard Error of Ozone Effect on Asthma ED Visits with and without Negative Control Exposure, under Varying Patterns of Misspecification

		No NCE	With negative exposures: future exposure (F), 1-week lag (LW1), 2-week lag (LW2), and moving average of 1-week to 2-week lag (MA)						
Scenario			F	LW1	LW2	MA	F+LW1	F+LW2	F + MA
1	Omit Nothing	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037
2	Omit weekday indicator	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037
3	Omit average temperature	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030
4	Omit dew point temperature	0.032	0.032	0.032	0.032	0.032	0.032	0.032	0.032
5	Omit holiday indicator	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037
	Omit time spline								
6	Knots=0	0.029	0.030	0.030	0.029	0.030	0.030	0.030	0.030
7	Knots=36	0.033	0.033	0.034	0.033	0.033	0.034	0.033	0.033
8	Knots=72	0.036	0.036	0.036	0.036	0.036	0.036	0.036	0.036
9	Knots=108	0.036	0.036	0.036	0.036	0.036	0.036	0.036	0.036

Table 4. Comparing Coverage of Ozone Effect on Asthma ED Visits with and without Negative Control Exposure, under Varying Patterns of Misspecification

		No NCE	With negative exposures: future exposure (F), 1-week lag (LW1), 2-week lag (LW2), and moving average of 1-week to 2-week lag (MA)						
Scenario			F	LW1	LW2	MA	F+LW1	F+LW2	F + MA
1	Omit Nothing	0.940	0.945	0.945	0.935	0.940	0.945	0.935	0.955
2	Omit weekday indicator	0.900	0.920	0.900	0.900	0.895	0.925	0.920	0.920
3	Omit average temperature	0.570	0.570	0.570	0.565	0.585	0.570	0.565	0.590
4	Omit dew point temperature	0.825	0.825	0.830	0.820	0.825	0.835	0.830	0.825
5	Omit holiday indicator	0.945	0.945	0.940	0.940	0.945	0.940	0.940	0.955
	Omit time spline								
6	knots=0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
7	knots=36	0.005	0.005	0.020	0.020	0.040	0.020	0.020	0.025
8	knots=72	0.425	0.380	0.510	0.490	0.530	0.435	0.440	0.465
9	knots=108	0.915	0.910	0.925	0.930	0.925	0.915	0.920	0.920

Figure 4. Estimated Associations between Daily ozone and Emergency Department Visits for Asthma in Atlanta, 1998 to 2009 with and without the use of negative exposure



Reference

Fewell, Z., Davey Smith G Fau - Sterne, J. A. C., & Sterne, J. A. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. (0002-9262 (Print)).

Flanders, W. D., Klein, M., Darrow, L. A., Strickland, M. J., Sarnat, S. E., Sarnat, J. A., Waller, L. A., Winquist, A., & Tolbert, P. E. (2011a). A method for detection of residual confounding in time-series and other observational studies. *Epidemiology*, *22*(1), 59-67. https://doi.org/10.1097/EDE.0b013e3181fdcabe

Flanders, W. D., Klein, M., Darrow, L. A., Strickland, M. J., Sarnat, S. E., Sarnat, J. A., Waller, L. A., Winquist, A., & Tolbert, P. E. (2011b). A method to detect residual confounding in spatial and other observational studies. *Epidemiology*, *22*(6), 823-826. https://doi.org/10.1097/EDE.0b013e3182305dac

Flanders, W. D., Klein, M., Strickland, M., Darrow, L., Sarnat, S., Sarnat, J., Waller, L., & Tolbert, P. (2009). A Method of Identifying Residual Confounding and Other Violations of Model Assumptions. *Epidemiology*, *20*(6). https://journals.lww.com/epidem/Fulltext/2009/11001/A_Method_of_Identifying_Residual_Confounding_and.100.aspx

Flanders, W. D., Strickland, M. J., & Klein, M. (2017). A New Method for Partial Correction of Residual Confounding in Time-Series and Other Observational Studies. *Am J Epidemiol*, *185*(10), 941-949. https://doi.org/10.1093/aje/kwx013

Goodman, M., Barraj Lm Fau - Mink, P. J., Mink Pj Fau - Britton, N. L., Britton Nl Fau - Yager, J. W., Yager Jw Fau - Flanders, W. D., Flanders Wd Fau - Kelsh, M. A., & Kelsh, M. A. Estimating uncertainty in observational studies of associations between continuous variables: example of methylmercury and neuropsychological testing in children. (1742-5573 (Electronic)).

Greenland, S. (2005). Multiple-Bias Modelling for Analysis of Observational Data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *168*(2), 267-306. https://doi.org/10.1111/j.1467-985x.2004.00349.x

Greenland, S., Pearl J Fau - Robins, J. M., & Robins, J. M. Causal diagrams for epidemiologic research. (1044-3983 (Print)).

Miao, W., Geng, Z., & Tchetgen Tchetgen, E. (2018). Identifying Causal Effects With Proxy Variables of an Unmeasured Confounder. *Biometrika*, *105*(4), 987-993. https://doi.org/10.1093/biomet/asy038

Miao, W., Shi, X., & Eric. (2020). A Confounding Bridge Approach for Double Negative Control Inference on Causal Effects. *arXiv pre-print server*. https://doi.org/None

arxiv:1808.04945

Miao, W., & Tchetgen Tchetgen, E. (2017). Invited Commentary: Bias Attenuation and Identification of Causal Effects With Multiple Negative Controls. *Am J Epidemiol*, *185*(10), 950-953. https://doi.org/10.1093/aje/kwx012

Schriger, D. L. (2008). Book and media review. *Annals of Emergency Medicine*, *52*(4), 480. https://doi.org/10.1016/j.annemergmed.2008.06.461