Assessing Seasonal Confounding and Model Selection Bias in Air Pollution Epidemiology Using Positive and Negative Control Analyses

Thomas Lumley Lianne Sheppard

NRCSE

Technical Report Series

NRCSE-TRS No. 030

Assessing seasonal confounding and model selection bias in air pollution epidemiology using positive and negative control analyses

Thomas Lumley Lianne Sheppard

Department of Biostatistics, and

National Research Center for Statistics and the Environment

University of Washington

Seattle

July 15, 1999

Abstract

Much of the evidence for health effects of particulate air pollution has come from ecologic time series studies that regress mortality or morbidity event counts on pollutant data routinely collected for other purposes. The modelling approach typically involves selecting both a lag at which the effect of particulates should be evaluated and a level of filtering to remove long term associations confounded with seasonal variations and secular trend. In this paper we investigate the bias introduced by model selection and residual confounding using simulations based on a previous analysis of data from King County, Washington. We find that the bias is small in absolute terms but of the same order as the estimated health impacts.

1 Introduction

The 1990 Clean Air Act mandates that EPA set and review standards for ambient air pollutants to protect the public from adverse effects. Much of the evidence for air pollution health effects has come from epidemiologic studies of short-term associations between pollutant levels and mortality or morbidity events in a defined geographic area. These are typically ecologic time series studies that regress event counts on pollutants from data obtained from routine sources. Mortality or morbidity counts are tallied from death certificate or hospital admissions databases. Often compilation of the the exposure data is fairly ad hoc: researchers will assemble all available pollutant measurements over a specific geographic area and then average the measurements by date for each pollutant. The usual approach to analysis involves fitting an overdispersed Poisson regression model with application of generalized additive models for control of time-varying covariates with smooth functions. The relative risk estimates obtained from these studies tend to be quite small (e.g. 1.05 for an interquartile range change). While the ubiquity of the exposure makes the public health importance of any true health effect undeniable, such small relative risk estimates from observational studies should naturally lead us to carefully examine the role of potential biases.

In this paper we will assess potential biases in an existing dataset. We will limit our attention to particulate matter (PM) as the air pollutant of interest. PM is subject to regulation as one of EPA's criteria pollutants (along with carbon monoxide, ozone, sulfur dioxide, and nitrogen dioxide). Unlike the gases, PM is a generic term for a broad class of particles originating from a variety of sources, existing in a range of sizes, and having chemically diverse properties. Given ambient air pollutants tend to coexist in the atmosphere, a single air pollution time series cannot provide good data for distinguishing the effects of different pollutants. Pollutants emissions from a common source, such as car exhaust or wood smoke, are highly correlated over time. Even when ambient levels of these pollutants are not highly correlated, the variation in correlation over time may be largely

due to varying meteorological conditions which are in themselves an important confounding factor. For these reasons, we will focus on PM while regarding it as a marker for a source or set of sources that produce a mix of pollutants whose relative health impact must be established from other data. We thus refer to the 'air pollution hypothesis' to describe increased risk of health outcomes due to increases in ambient air pollutants indexed by PM. Our perspective is consistent with others such as a simulation study confirming in the air pollution setting the well-known theoretical problem of distentangling the effects of highly correlated predictors (Chen et al., 1999), a critical review of the epidemiologic evidence for the PM hypothesis (Moolgavkar & Lubeck, 1996), and an examination of the uncertainties in attributing health effects to specific pollutants (Lipfert & Wyzga, 1995).

One important difficulty in evaluating even the weaker air pollution hypothesis is that there are seasonal variations in mortality and morbidity and in air pollution that are partially confounded. Another difficulty is that there is no *a priori* reason to specify any particular induction period for the effects of air pollution, so a search must be conducted over a number of possible models. The potential biases that may result from removal of seasonal confounding by filtering or smoothing techniques together with the effects of searching over a handful of possible lags would usually be regarded by statisticians as relatively minor problems since they are expected to only cause very small biases. When studying the health effects of air pollution, however, the observed excess risks are typically a few percent per interquartile range of a pollutant. In this case biases that are ordinarily negligible can be of great importance. The relative risks are similar to those seen in very large clinical trials of treatment for myocardial infarction (eg The GUSTO investigators., 1993; The GUSTO III Investigators, 1997) rather than in typical observational studies, and the examination of possible bias in design and analysis must be correspondingly acute.

In this paper we will assess the magnitude of some potential biases. Rather than attempt to remove these biases by statistical manipulation we decide to perform a controlled experiment by using the same model selection and adjustment procedures in a situation

where the true association is known to be zero.

Sheppard et al. (1999) presented a study of the associations between hospital admissions for asthma in non-elderly residents of greater Seattle and various components of air pollution. They introduced the idea of a separate control analysis by also studying associations between air pollution and admissions for appendicitis over the same period. Using the same model selection procedure for asthma admissions and for appendicitis admissions, where no association was expected, they could conclude that the associations were larger than could be explained by the multiple testing involved in model selection. The main limitation of their control analysis is that asthma and appendicitis do not have the same pattern of variation with season and temperature. The two analyses thus do not have the same pattern of potential confounding by season.

In this paper we extend the control analysis performed by Sheppard et al. (1999) to include the effects of confounding as well as model selection. We use the observed asthma admissions and meterological data together with simulated air pollution time series. With this approach we preserve the potential for confounding by modelling the relationship between temperature at various lags, calendar date and particulate air pollution and using this model to induce an association between season and the simulated air pollution data. The simulation does not include any association between particulate air pollution and the outcome, so any apparent association will be due entirely to confounding and model selection biases. For comparison, we also conduct a positive control analysis in which a specified non-zero excess risk is added to the simulation, to examine the extent of bias when a true association is present.

We also perform two simpler control analyses, regressing asthma admissions on $PM_{2.5}$ from a different year or a different location. Again, there can be no causal assocation, so any apparent health impact is the result of residual confounding or statistical artifact.

2 Methods

2·1 *Data sources*

Data on hospital admissions for asthma in people under 65 in King County, Washington, were obtained from the Comprehensive Hospital Abstracts Reporting System database, which records every hospital discharge in Washington State. The air pollution exposure measurements were averaged over the three monitors in the Seattle area. Meteorological information was measured at the Seattle–Tacoma International Airport. The data are described in detail by Sheppard et al. (1999), including the imputation techniques used when one or more monitors had missing data.

2·2 *Simulation methods*

The basis of the simulation design is a model for generating realistic $PM_{2.5}$ series based on season and current and recent temperature. There are two components to this model. The systematic component involves a linear regression model for $\log PM_{2.5}$; the random component is a model for the autocorrelated residuals.

After smoothing, Seattle temperatures vary seasonally in a way that is well approximated by a sine wave. For this reason the seasonal component of the model was based on a sine and cosine term. Additional higher frequency terms were also investigated. To account for variation in $PM_{2.5}$ at shorter time scales we used an indicator variable for each day of the week and smooth functions of temperature. Initially a smoothing spline with four degrees of freedom for current temperature was used, and similar smooth functions were added for lagged temperatures whenever they caused a substantial reduction in the residual variance, starting with one day lag and working back to 30 days. Finally, the smooth functions of temperature were reduced to 1 or 2 degrees of freedom, which did not have a substantial impact on prediction.

The resulting model included 2-df smooths of temperature on the same day and with a three day lag, and linear terms for temperature at 6 and 21 day lags as well as the

day-of-week and seasonal sine wave terms. The random part of the model for $\log \mathrm{PM}_{2.5}$ is a Gaussian AR-1 process with autocorrelation 0.6 and standard deviation 0.556. The observed $\mathrm{PM}_{2.5}$ series and three simulated realisations are shown in Figure 1.

In the simulation study we generated 400 realisations of this model and calculated the log relative risk for a 1 unit increment in $\mathrm{PM}_{2.5}$ at each of lags 0–6. This analysis was performed unadjusted, adjusted for a 64df regression spline (8df/year) in time and adjusted for both this regression spline and a 4df spline in current day's temperature. The degrees of freedom match those previously used for these data. The latter two models gave extremely similar results in all cases and so only the unadjusted and fully adjusted models are presented here. Using 400 realisations allows a reasonably accurate estimate of the upper and lower 2.5 percentiles, each of which excludes 10 observations.

A second study added extra simulated events to the outcome to achieve a relative risk of 1.1 over the interquartile range of $\mathrm{PM}_{2.5}$. This is a positive control study designed to estimate the bias when a true association is present. It is worth noting that to create this relative risk, which is larger than that actually seen in King County, required adding less than one extra asthma case every two days.

3 Results

Table 1 gives a summary of all simulation results, while for each condition the distribution of simulation estimates is shown in Figures 2-5. In all the analyses, the log relative risks estimated without adjusting for season and temperature are positively biased. This indicates that the uncontrolled $\mathrm{PM}_{2.5}$ effect estimates incorporate the important seasonal confounding that is present in the data. Both $\mathrm{PM}_{2.5}$ and asthma morbidity are higher in the winter than in the summer, causing a spurious association over and above any true effect of pollution.

Figure 1: $PM_{2.5}$ (μgm^{-3}) series from Seattle, Washington, and three simulated series. The genuine data are in the lower left panel.

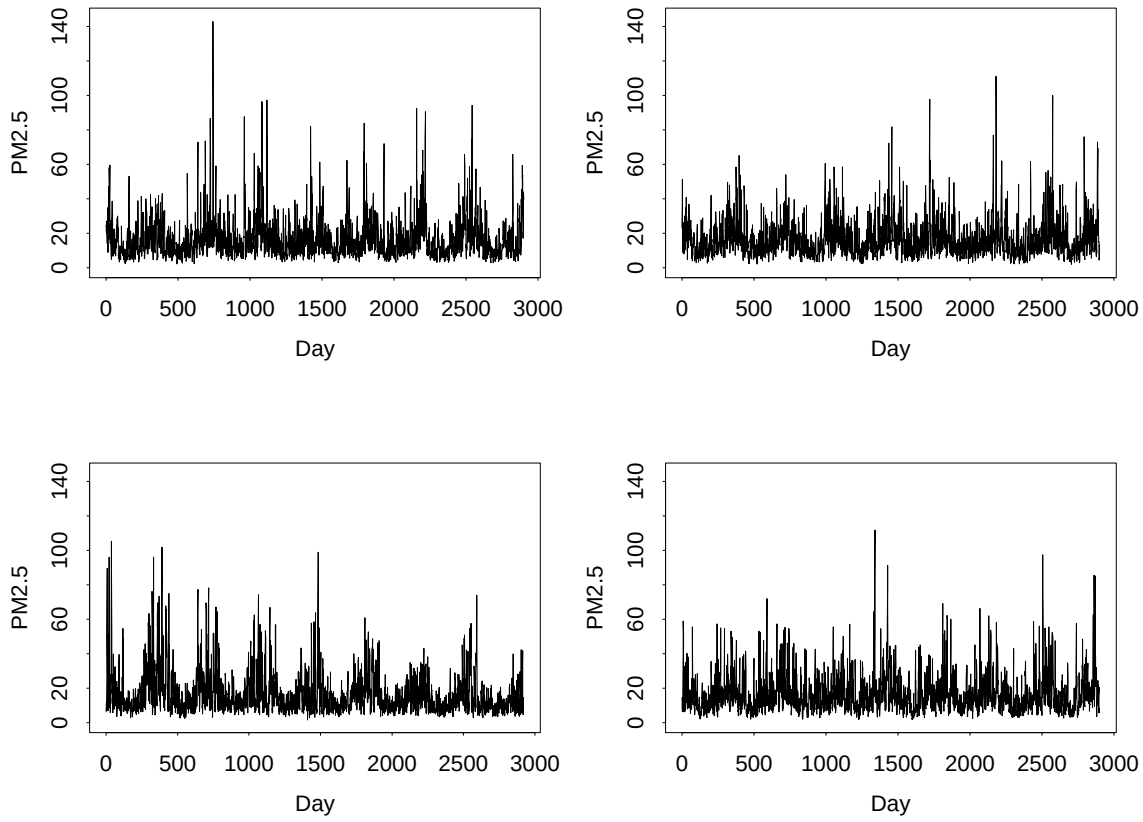

Table 1: Median and upper and lower 2.5 percentiles of 400 log relative risks (per μgm^{-3}) from simulated data sets with and without a true association between $PM_{2.5}$ and outcome.

Simulation	Adjustments	Median log(RR) (2.5%, 97.5%)	True log(RR)	Observed data estimate
Current day lag				
No association	none	0.0026 (0.0014, 0.0038)	0	0.0055
	season & temperature	−0.0002 (−0.0030,0.0021)	0	0.0027
Association	none	0.011 (0.009, 0.013)	0.0083	—
	season & temperature	0.008 (0.006, 0.010)	0.0083	—
Best of 7 lags				
No association	none	0.0038 (0.0019, 0.0057)	0	0.0055
	season & temperature	0.0013 (−0.0003,0.0032)	0	0.0027
Association	none	0.011 (0.009,0.013)	0.0083	—
	season & temperature	0.008 (0.006, 0.010)	0.0083	—

3·1 *Known lag structure*

The 400 log relative risks for simulations including only the association estimated at lag 0 are shown in Figure 2. In this case the bias in the adjusted analysis is small, with the median estimated log relative risk being -0.00021. In only two of the 400 realisations does the estimate exceed the log relative risk of 0.00275 estimated from the observed data, giving a Monte Carlo p-value of (2+1)/(400+1)=.007.

In the positive control model, where the true log relative risk is 0.0083, the bias is negligble, with the median estimated log relative risk being 0.0084. These 400 estimates are shown in Figure 3.

3·2 *Data-dependent lag structure*

To select the best lag in the simulations we fitted models with each lag from 0 to 6 days and chose the one that gave the largest relative risk for PM. The 400 log relative risks for simulations including the association estimated at the best lag are shown in Figure 4. When the largest relative risk at any lag from 0 to 6 is estimated, the bias in the adjusted analysis is substantially larger, with the median estimated log relative risk being $0.0013\mu g^{-}1m^{3}$. This median bias is approximately half the log relative risk of 0.002746 estimated from the observed data, and in 32 of the 400 realisations the simulated estimate exceeds the observed log relative risk giving a Monte Carlo p-value of (32+1)/(400+1)=.08.

In the positive control model, where the true log relative risk is 0.0083, the bias is still negligible, with the median estimated log relative risk being 0.0082. These 400 estimates are shown in Figure 5.

3·3 *Control data*

We also examined two sources of real control data. The first was to use PM from the same area but a different year. We fitted the same models to data sets using the first seven, six or five years of health outcomes and the last seven, six or five years of PM data. The

Figure 2: Histogram of 400 estimated log relative risks for $1\mu\text{gm}^{-3}$ increment in in current day's $PM_{2.5}$ when the true log relative risk is zero. The vertical line indicates the estimate from the observed data.

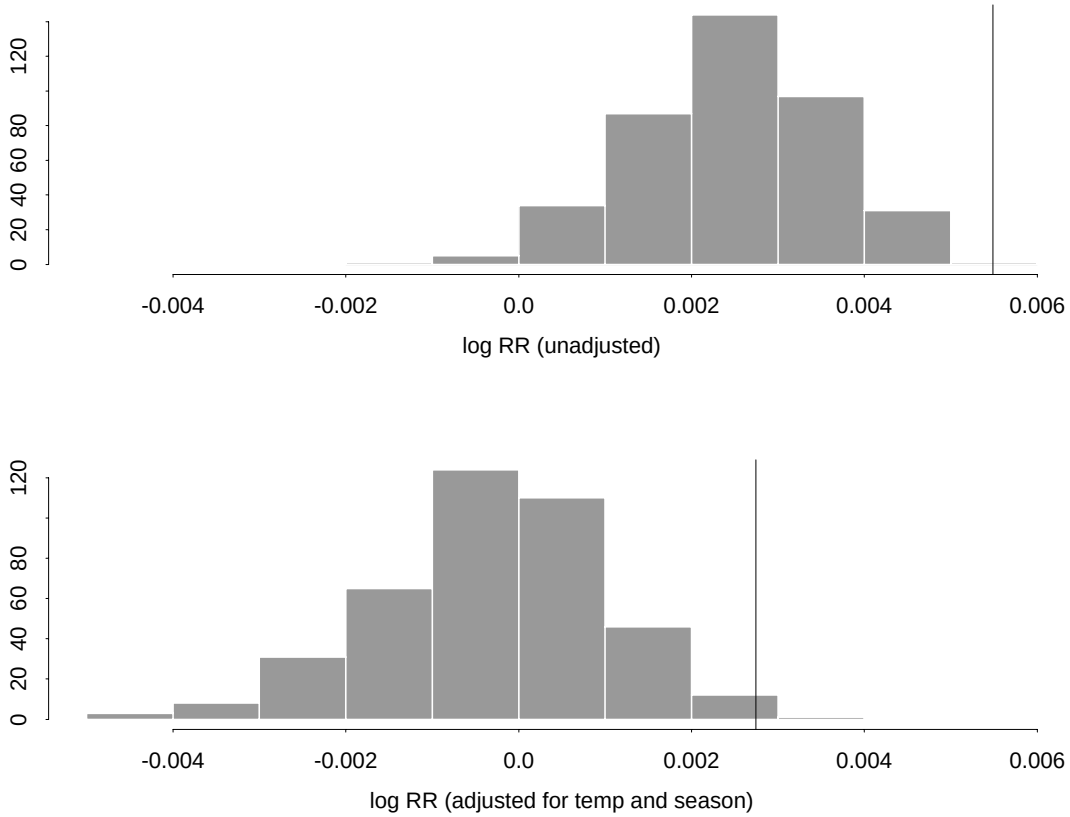

Figure 3: Histogram of 400 estimated log relative risks for $1\mu\text{gm}^{-3}$ increment in in current day's $\text{PM}_{2.5}$ when the true relative risk is 1.1 over the interquartile range. The vertical line indicates this true log relative risk $\beta = 0.0083$ per $1\mu\text{gm}^{-3}$.

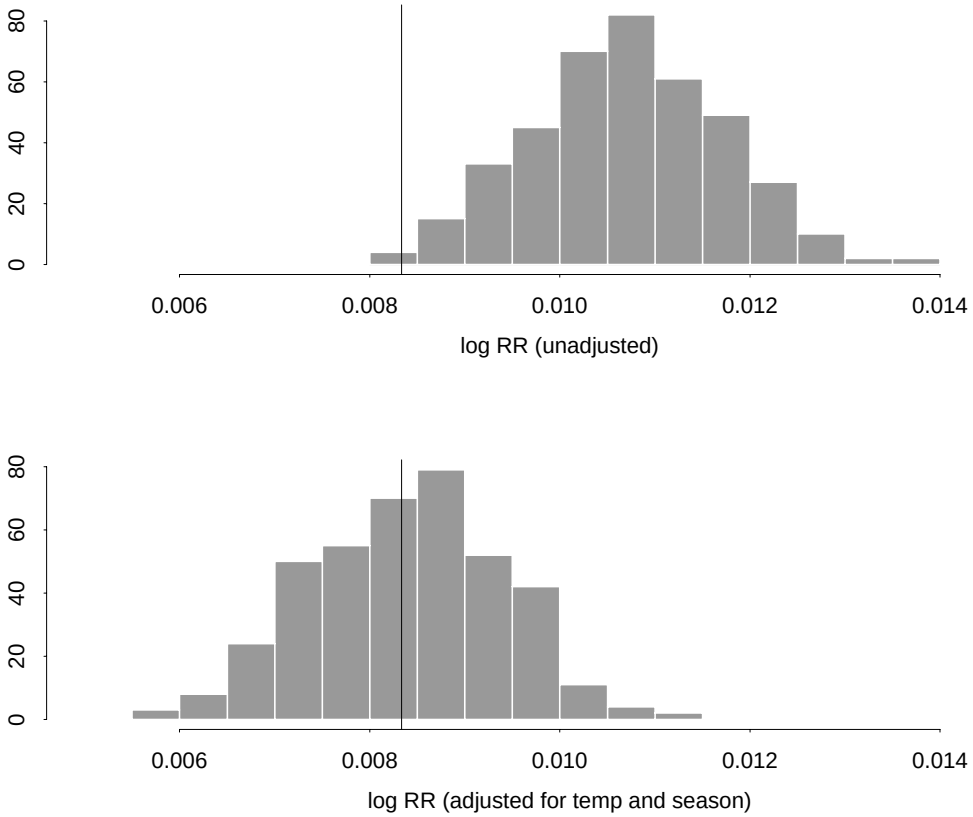

Figure 4: Histogram of 400 estimated log relative risks for $1\mu\text{gm}^{-3}$ increment in $\text{PM}_{2.5}$, choosing the lag that gives the greatest estimate, when the true log relative risk is zero. The vertical line indicates the estimate from the observed data.

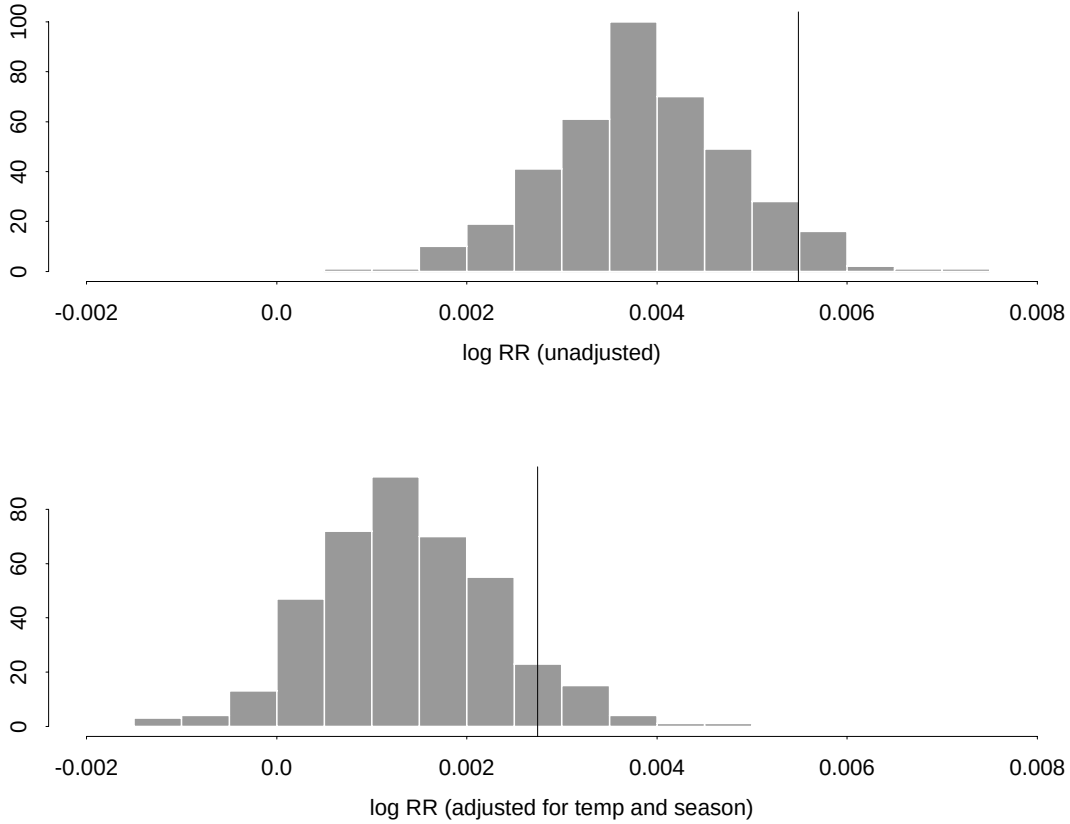

Figure 5: Histogram of 400 estimated log relative risks for $1\mu\text{gm}^{-3}$ increment in $\text{PM}_{2.5}$, choosing the lag that gives the greatest estimate, when the true relative risk is 1.1 over the interquartile range. The vertical line indicates this true log relative risk $\beta = 0.0083$ per $1\mu\text{gm}^{-3}$.

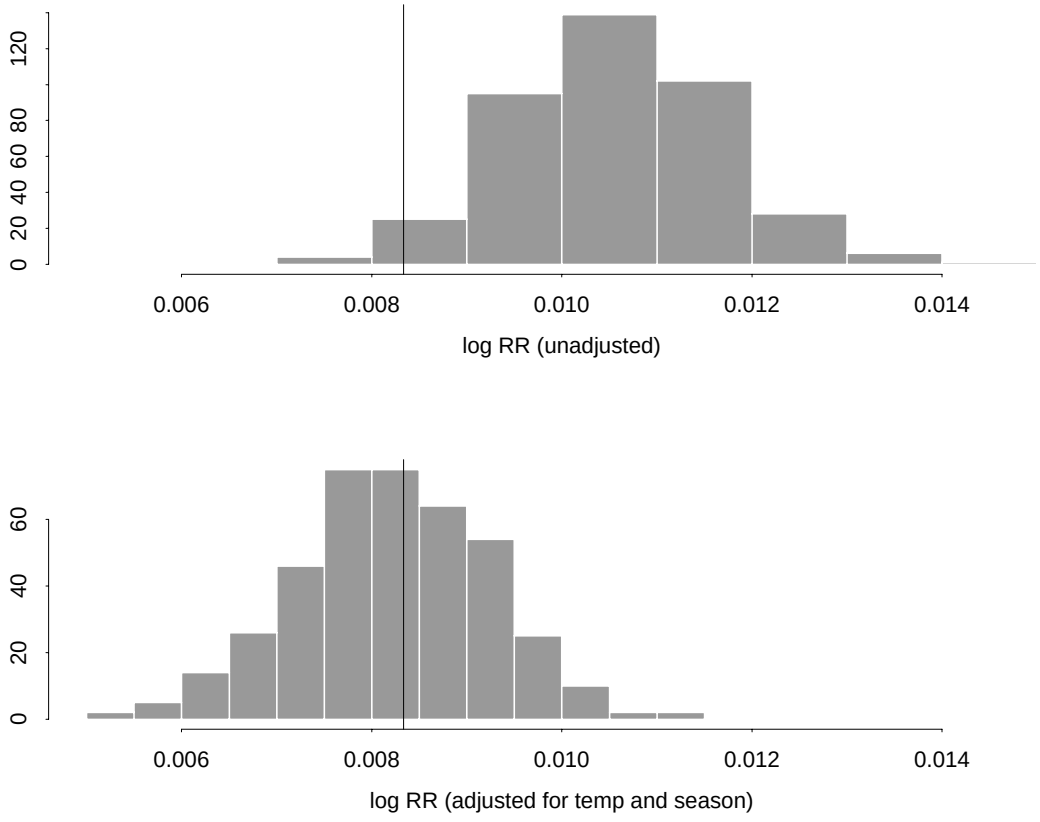

Table 2: Assocations (log relative risk per μgm^{-3}) between asthma admissions and $PM_{2.5}$ one, two and three years in the future, unadjusted and adjusted for season and temperature.

	Adjustments	
	none	season & temperature
One year offset:		
Lag 0	0.0048	−0.0005
Best lag	0.0048	−0.0005
Two year offset:		
Lag 0	0.0048	−0.0005
Best lag	0.0062	0.0021
Three year offset:		
Lag 0	0.0044	−0.0003
Best lag	0.0057	0.0018

causal assocation in these analyses, between asthma on a given day and $PM_{2.5}$ one to three years in the future must be zero, providing a valid negative control. The results are given in Table 2.

These results are consistent with the simulations. The bias at a fixed lag in the adjusted analyses is negligible. When the lag is selected from the data the bias is still negligible at a one year offset, but more than half as big as the observed association at two and three year offsets.

The second control analysis used PM data from Portland, Oregon. Portland is approximately 140 miles south of Seattle and has a similar climate, lifestyle and population density. The weather in the two cities is broadly similar but weather patterns that affect one may miss the other, or arrive at a different time. We obtained nephelometry data from Port-

Table 3: Assocations (log relative risk per μgm^{-3}) between Seattle asthma admissions and Seattle and Portland PM based on nephelometry, unadjusted and adjusted for season and temperature.

	Adjustments	
	none	season & temperature
Seattle:		
Lag 0	0.0048	0.0017
Best lag	0.0052	0.0020
Portland:		
Lag 0	0.0018	0.0003
Best lag	0.0020	0.0007

land for the same time period as the Seattle study. Nephelometry, which measures light scattering, gives readings highly correlated with direct measurements of $PM_{2.5}$ at least in Seattle. Table 3 shows the best lag and current day relative risks for an increase of $1\mu gm^{-3}$ in $PM_{2.5}$ based on the nephelometer readings, both unadjusted and adjusted for time and temperature.

This analysis shows a similar pattern to the others. The additional bias from model selection is again noticeable.

4 Discussion

We have presented three additional negative control analyses to allow the level of statistical uncertainty in the results of Sheppard et al. (1999) to be further assessed. These analyses indicate that the bias from a combination of residual seasonal confounding and model selection is not negligible. It is well-known that the bias due to model selection increases

with the number of candidate models. We found that whene selecting the best of only seven candidate model, the log relative risk based on the Seattle data is about twice the mean bias, and is only at the 90th percentile of the bias distribution in these control analyses. This demonstrates that in investigating the extremely weak associations between PM and health outcomes we need to be concerned about levels of statistical bias that would ordinarily be negligible.

Our positive control analyses indicate that the bias is much smaller when the true association is moderately large. This suggests that in most epidemiologic studies we would not need to be concerned about bias caused by selection from a handful of candidate models, and that even in studying PM the bias may be less important if we can identify more sensitive populations or refine our characterisation of exposure.

In creating data where the association between PM and health outcomes is known there is necessarily some distortion from reality. In simulations we assume that the effect of season and temperature on asthma incidence is not mediated by PM to any important degree and we rely on the extent to which we can duplicate the short-term patterns of PM by our simulation model. The lagged analyses and the Portland nephelometry analysis do not require modelling of PM. The lagged analyses will have somewhat weaker confounding than the true data, since PM and weather are linked only through their seasonal relationship. Confounding by weather in the Portland data is less likel to be removed since the cities share large-scale weather patterns. These analyses, in contrast to the simulations, provide only a small number of comparison estimates rather than a complete bias distribution.

We have not addressed the important issues of measurement error and of multiple pollutants. For these issues the ecological time series study cannot stand on its own. Measurement error modelling requires at least some information on actual exposures to calibrate a measurement error model. Disentangling the effects of multiple pollutants from the same source will always be difficult in a single time series, and either more basic toxicological information or comparisons of relative risks between cities with different pollution profiles

are more appropriate methods.

Investigators in this area need to be aware of the potential for bias from model selection, and it must be taken into account for analyses to be interpretable. One way to remove the bias is to select the model *a priori*, as is being done for the 100 largest US cities by the National Morbidity and Mortality Air Pollution Study. This will sometimes be too inflexible an approach. Control analyses, such as that in Sheppard et al. (1999), or those presented here are a valuable tool if control data or suitable simulated data can be produced. Bayesian model averaging techniques(Clyde & DeSimone-Sasinovska, 1999; Madigan & Raftery, 1994) provide another potentially valuable tool. Model averaging allows a large number of models to be fitted and the differences in conclusions between them to be incorporated explicitly in the statistical conclusions, though the bias characteristics of Bayesian model averaging in this context have not been thoroughly evaluated.

5 Acknowledgements

Although the research described in this article has been funded in part by the United States Environmental Protection Agency through agreement CR825173–01–0 to the University of Washington, it has not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred.

References

Chen, C., Chock, D. P., & Winkler, S. L. (1999). A simulation study of confounding in generalized linear models for air pollution epidemiology. *Environmental Health Perspectives* **107**, 217–222.

Clyde, M. & DeSimone-Sasinovska, H. (1999). Does particulate matter particularly matter? Accounting for model uncertainty in Poisson regression models. ISDS Discussion Paper, Duke University.

Lipfert, F. W. & Wyzga, R. E. (1995). Uncertainties in identifying responsible pollutants in observational epidemiology studies. *Inhalation Toxicology* **7**, 671–689.

Madigan, D. & Raftery, A. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's Window. *Journal of the American Statistical Association* **89**, 1535–1546.

Moolgavkar, S. H. & Lubeck, E. G. (1996). A critical review of the evidence on particulate air pollution and mortality. *Epidemiology* **7**.

Sheppard, L., Levy, D., Norris, G., Larson, T. V., & Koenig, J. Q. (1999). Effects of ambient air pollution on nonelderly asthma hospital admissions in seattle, washington, 1987–1994. *Epidemiology* **10**, 23–30.

The GUSTO III Investigators (1997). A comparison of reteplase with alteplase for acute myocardial infarction. *New England Journal of Medicine* **337**, 1118–23.

The GUSTO investigators. (1993). An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *New England Journal of Medicine* **329**, 673–82.