Bias in the Case--Crossover Design: Implications for Studies of Air Pollution

Thomas Lumley Drew Levy

NRCSE

Technical Report Series

NRCSE-TRS No. 031

Bias in the case–crossover design: implications for studies of air pollution

Thomas Lumley
thomas@biostat.washington.edu
National Research Center for Statistics and the Environment
University of Washington

Drew Levy
Department of Epidemiology
University of Washington

July 15, 1999

Abstract

The case–crossover study design in principle allows testing for an acute health effect of an exposure such as air pollution with restriction in time to remove seasonal confounding. We argue that much thinking about this design has been based on false analogies with matched case–control studies, and show that the usual design and analysis of case–crossover studies introduce a previously unrecognised bias by not conditioning correctly on the sampling scheme. We show that the usual conditional logistic regression analysis is a valid maximum likelihood method for a simple modification of the case–crossover design that divides time into strata. However, the bias from standard case–crossover designs will typically be small.

1 Introduction

The case–crossover study design proposed by Maclure (1991) is suited to the study of a transient effect of an intermittent exposure on the subsequent risk of a rare acute-onset disease hypothesized to occur a short time after exposure. The health effects of fine particulate matter air pollution is a topical epidemiologic issue for which the case-crossover design may be especially useful. Fine particulate matter air pollution (PM) is an exposure which varies over time and there is concern that PM may affect the incidence of acute cardiovascular and respiratory disease events. The principle of Maclure's case–crossover design is that the exposures of cases just prior to the event are compared using matched case-control methods to the distribution of exposure estimated from data for the case from some separate time period. This separate referent time period should be representative of the expected distribution of exposure for follow-up time that does not result in a case. Many different referent selection strategies will likely be available in any one study. The best strategy may depend on the quality of the data in any particular study, and involve trade-offs between bias and precision (Mittelman et al., 1995).

A disadvantage of the case–crossover design is the potential for bias due to time trends in the exposure time-series. Since case–crossover comparisons are made for each case between different points in time, the case–crossover analysis implicitly depends on an assumption of stationarity of the air pollution time series. If the exposure time-series is non-stationary and case exposures are compared with referent exposures systematically selected from a different period in time, there may be a bias introduced into estimates of the measure of association for the exposure and disease. Greenland (1996) has identified this as a form of selection bias. Short term autocorrelation in the exposure time-series may introduce a bias analogous to over-matching in a case-control study when exposures proximal in time to the index case exposure are used as referents. Characteristics of time-series of PM, including long term time trend, seasonal trends, and short term autocorrelation, require that referent selection in a case–crossover study be considered carefully and adapted to minimize bias.

Navidi (1998) suggests that when subsequent exposures are not influenced by failures, as in studies of environmental exposures such as air pollutants, it is possible to determine at times postfailure what a subjects level of exposure would have been had the subject not failed. He proposes that ambidirectional sampling — sampling referents from time both before and after failure times — would control for linear long-term time trend.

We had undertaken simulations to explore the nature and degree of time-selection bias and to examine the ability of various referent selection strategies to counter biases in a case–crossover analysis of the association of PM and out-of-hospital primary cardiac arrest in Seattle, Washington. We felt that the problem of seasonality and long-term time trend in the PM time-series might be be dealt with by restriction of the sample frame for referents to a period short enough to be free of significant seasonal transitions. Our simulation studies to compare different referent selection designs showed that the gross biases from trend and season effects were indeed removed. However, we found persistent small to moderate biases, prompting us to review the theoretical justification for conditional logistic regression analysis of case–crossover designs. We found that the standard justification of conditional logistic regression in these designs was not correct and that in some cases the estimators were inconsistent.

In section 2 we explain why the justification of conditional logistic regression by analogy with case–control studies is not valid. In section 3 we analyse the conditional logistic regression estimating functions and show that there is only one standard case–crossover design in which they are unbiased, one which is undesirable in our problem for other reasons. We show that a slight modification of standard ambidirectional case–crossover studies does produce an unbiased analysis.

In section 4 we study various forms of the standard case–crossover design and the biases from conditional logistic regression analysis and from using referent data from after a terminal event. We show by simulation and direct calculation that designs can be constructed for which these biases are while not zero, are reliably small. Either by using these designs or by computing and adjusting for the bias the standard case–crossover design can still be used when ambidirectional assessment of exposure is not possible. In particular, when the exposure series for different individuals are independent the bias will be asymptotically negligible.

2 Case–control and case–crossover designs

The case–control and case–crossover designs can both be analysed in terms of a proportional hazards model for a rare disease. The case–crossover design, in order to compare hazards for the same person at different times, requires a parametric assumption about the baseline hazard, which is invariably that it is constant over time for each individual. We also allow this constant baseline hazard to vary arbitrarily between individuals, thus incorporating the effects of any time-constant covariates. The resulting model for the hazard $\lambda_i(t; z_{it})$ of person i at time t based on time-varying covariates z_{it} is

$$\lambda_i(t; z_{it}) = \lambda_i \exp(\beta' z_{it}).$$

These analogies between the matched case–control and case–crossover designs are relied on by many people wanting to think heuristically about the design properties. However these analogies are much weaker than seems to be appreciated. Although both designs can be considered as sampling from a full cohort there are a number of fundamental differences between case–control and case–crossover design schemes. Some of these are general and others are important specifically in studing air pollution.

First there is the role of time and the resulting autocorrelation. In a matched case–control study the exposure observations within each stratum are independent given the choice of stratum. In a case–crossover study there will typically be autocorrelation in the exposure over time. Also, in the case of air pollution and similar environmental exposures two cases that occur on the same day must have the same or similar exposure, a between-stratum constraint that is not present in case–control studies.

In an ambidirectional case–crossover study of an event such as death that can only occur once some of the referent information is taken from times when the subject is not at risk for the event. In a case–control study all controls must be at risk for the event.

Finally, in a matched case–control study the division of the population into strata depends only on covariates and not on the response. As we will show in section 3·2, in a case–crossover design the division into strata depends on the response. This is related to "overlap bias", discussed by Austin et al. (1989) in that failure to use disjoint strata that partition the population can invalidate assumptions about independent sampling.

All four of these differences can in principle lead to bias. The last, the improper creation of strata, is responsible for the inconsistency of conditional logistic regression. The autocorrelation of the exposure series affects the size of the bias. The correlation between individuals produces a bias which is non-zero conditional on the observed exposure series but may be zero unconditionally. The use of control days when a subject is not actually at risk turns out to produce negligible bias in any realistic situation.

The bias is zero, apart from the negligible contribution from sampling controls after an event, when there is no association between response and exposure, but can be towards or away from the null hypothesis when there is a true association.

3 The conditional logistic regression analysis

The data for the case–crossover design for a rare or terminal event consist of the exposure process $\langle z_{it}\rangle_{t=1,i=1}^{T,n}$, the case response times $\langle t_i\rangle_{i=1}^{n}$ and the referent sets, which we denote either $\mathcal{W}_i$ or $\mathcal{W}(t_i)$ to emphasize the dependence on individual or on event time. The estimating equations are the conditional logistic regression equations for a matched case–control study, but the expectations are taken over different variables, so the theory does not go through automatically.

We will initially ignore the potential for bias caused by using exposures after a death, and will assume that all subjects remain at risk for the entire time period but that the outcome is rare enough that no multiple events are observed.

3·1 *Navidi's design*

Navidi (1998) described a case–crossover analysis in which the referent window for each case is the entire available time period: $\mathcal{W}_i = \{1, 2, 3, \ldots, T\}$. Conditioning on the exposure series and on exactly one event being observed creates a conditional likelihood that is formally identical to that for a matched case–control study, showing that conditional logistic regression is valid in this case.

It is informative to consider the relationship of this design to a Poisson time series analysis based on the same cases. We begin with Navidi's estimating equations, which can be written as

$$\sum_{i=1}^{n} U_i(\beta) \equiv \sum_{i=1}^{n} \left(z_{it_i} - \sum_{t=1}^{T} z_{it} \frac{e^{\beta' z_{it}}}{\sum_{s=1}^{T} e^{\beta' z_{is}}} \right) = 0.$$

There are no individual-level time-constant covariates, as their coefficients are unidentifiable under Navidi's model. If we also assume that there are no individual-level time-varying covariates then $z_{it} \equiv z_t$. Defining Y_t to be the number of events on day t we have

$$\begin{aligned} \sum_{i=1}^{n} \left(z_{it_i} - \sum_{t=1}^{T} z_{it} \frac{e^{\beta' z_{it}}}{\sum_{s=1}^{T} e^{\beta' z_{is}}} \right\} &= \sum_{i=1}^{n} \left\{ z_{t_i} - \sum_{t} z_t \frac{e^{\beta' z_t}}{\sum_s e^{\beta' z_s}} \right) \\ &= \sum_{t=1}^{T} z_t \left(Y_t - \frac{n}{T} \frac{e^{\beta' z_t}}{\sum_s e^{\beta' z_s}} \right) \end{aligned}$$

Now we observe that $n/T \sum_s \exp(\beta' z_s)$ is a constant and so may be dropped at the cost of including an intercept parameter. The estimating function is then

$$\sum_{t=1}^{T} z_t \left(Y_t - e^{\tilde{\beta}' z_t} \right)$$

where an intercept has been added to β to get $\tilde{\beta}$. These are the estimating functions for a Poisson regression model. Similar analysis shows that the covariance estimate is the nominal-dispersion parametric covariance estimate from Poisson regression.

This case–crossover design can be validly analysed by conditional logistic regression but has the same sensitivity to seasonal confounding, overdispersion and autocorrelation in the outcome variable as a naive Poisson time series design. Its advantage over a time series analysis is that it offers a convenient way to incorporate information on individual-level effect modifiers, whether time-varying or time-constant.

3·2 *Referent window designs*

Here we use a fixed number of referent days before and possibly after the case in a short time frame. The motivation for this design is to restrict the referents in time to reduce seasonal confounding. This differs from the usual approach to seasonal confounding by restriction in the frequency domain, either explicitly by Fourier decomposition (Kelsall et al., 1999) or implicitly by smoothing (eg Samet et al., 1995).

The conditional likelihood approach used by Navidi does not translate to this design. Knowing the referent window $\mathcal{W}_i$ completely determines the case time t_i, so the conditional likelihood of t_i given exactly one case in $\mathcal{W}_i$ contains no information about β. What the design actually does is sample referent windows from the fixed exposure series with probability proportional to the case risk function at the window center $\exp(\beta' z_{t_i} + \gamma_{t_i})$, where z_{t_i} are the covariates of interest and γ_t represents the season and trend effects that we wish to exclude by restriction. Using a narrow referent window means that γ_t is roughly constant on the window.

The likelihood of the full data, and thus the score equation, is the same as in Navidi's design (except that we now explicitly include the unmeasured and low-frequency covariate effects γ_t) as the model generating the cases is the same. The score function is

$$U_i(\beta) = z_{it_i} - \sum_{t=1}^{T} \frac{z_{it} e^{\beta' z_{it} + \gamma_t}}{\sum_{s=1}^{T} e^{\beta' z_{is} + \gamma_s}}.$$

This score function depends on all the data, and involves the unwanted confounders γ_t. We approximate it by

$$\tilde{U}_i(\beta) = z_{it_i} - \sum_{t \in \mathcal{W}_i} \frac{z_{it} e^{\beta' z_{it} + \gamma_t}}{\sum_{s \in \mathcal{W}_i} e^{\beta' z_{is} + \gamma_s}}$$

where $\mathcal{W}_i$ is the referent region for a case at t_i and now assume γ_t is approximately constant over $\mathcal{W}_i$ to get

$$\tilde{U}_i(\beta) = z_{it_i} - \sum_{t \in \mathcal{W}_i} \frac{z_{it} e^{\beta' z_{it}}}{\sum_{s \in \mathcal{W}_i} e^{\beta' z_{is}}}.$$

This is an appealing estimating function as it depends only on β and data nearby in time, and is identical to the conditional logistic regression score equation that would arise if this were a matched case–control study. It is not, however, the derivative of an actual loglikelihood so we cannot automatically assume it has mean zero. Its expectation is given by integrating out the discrete random variable t_i. We will assume that γ_t is roughly constant on each window $\mathcal{W}_i$. This is reasonable if the referent windows are not too long. This ensures that γ_t and z_{it} are approximately independent within each window.

$$\begin{aligned}
\mathbf{E}_{t_i}\left[\tilde{U}(\beta)\right] &= \sum_{t_i=1}^{T} \frac{e^{\beta' z_{t_i} + \gamma_{t_i}}}{\sum_{s=1}^{T} e^{\beta' z_{is} + \gamma_s}} \left(z_{t_i} - \frac{\sum_{u \in \mathcal{W}_i} z_u e^{\beta' z_u}}{\sum_{u \in \mathcal{W}_i} e^{\beta' z_u}} \right) \\
&= \frac{\sum_{t=1}^{T} z_{it} e^{\beta' z_{it} + \gamma_t}}{\sum_{s=1}^{T} e^{\beta' z_{is} + \gamma_s}} - \sum_{t=1}^{T} \frac{\sum_{u \in \mathcal{W}(t)} z_u e^{\beta' z_u + \beta' z_{it} + \gamma_t}}{\sum_{s=1}^{T} \sum_{u \in \mathcal{W}(t)} e^{\beta' z_{is} + \gamma_s + \beta' z_u}} \\
&\approx \frac{\sum_{t=1}^{T} z_{it} e^{\beta' z_{it}}}{\sum_{s=1}^{T} e^{\beta' z_{is}}} - \sum_{t=1}^{T} \frac{\sum_{u \in \mathcal{W}(t)} z_u e^{\beta' z_u + \beta' z_{it}}}{\sum_{s=1}^{T} \sum_{u \in \mathcal{W}(t)} e^{\beta' z_{is} + \beta' z_u}} \\
&= \frac{\sum_{t=1}^{T} z_{it} e^{\beta' z_{it}}}{\sum_{s=1}^{T} e^{\beta' z_{is}}} - \frac{1}{\sum_{s=1}^{T} e^{\beta' z_{is}}} \sum_{t=1}^{T} \frac{\sum_{u \in \mathcal{W}(t)} z_u e^{\beta' z_u + \beta' z_{it}}}{\sum_{u \in \mathcal{W}(t)} e^{\beta' z_u}}
\end{aligned}$$

This is not identically zero, so the estimating functions are not unbiased conditional on the exposure history $\langle z_{it} \rangle$ as they would be in a true maximum likelihood analysis. Note that if the referent window $\mathcal{W}(t)$ is the same for all t, as in Navidi's design, the last term factors into a sum over $\mathcal{W}(t)$, which equals the first term, and a sum over $1, \ldots, T$, which reduces to 1. The estimating equations are thus unbiased if $\mathcal{W}(t)$ is the same for all t.

Further insight into the reasons for this bias can be obtained by comparing it to a related conditional likelihood score equation. If we partition the times $1, \ldots, T$ *a priori* into disjoint strata $\mathcal{S}(t)$ we can condition on there being exactly one event in the stratum $\mathcal{S}(t)$. This is different from the case–crossover design where the referent window $\mathcal{W}(t)$ is different for every t and in the ambidirectional design is centered at t. The score function for this conditional likelihood is

$$U_i(\beta) = z_{t_i} - \sum_{t \in \mathcal{S}(t_i)} z_{it} \frac{e^{\beta z_{it}}}{\sum_{s \in \mathcal{S}(t_i)} e^{\beta z_{is}}}$$

This unbiased score function is formally identical to the biased estimating function for the case–crossover design. The difference is that the strata $\mathcal{S}(t)$ do form a partition, whereas the referent windows $\mathcal{W}(t)$ are overlapping and different for each t. The choice of which days go together in a referent window depends on the outcome t_i and is not a valid stratification.

The bias can thus be removed by dividing time into strata and using the remainder of the days in each stratum as the referents for a case in that stratum. Navidi's design is given by the special case of a single stratum. When ambidirectional sampling of exposure is possible we recommend this stratified design as a simple modification that makes the conditional logistic regression analysis valid.

Although the bias can be removed by proper stratification it is still of interest to investigate the size of the bias in the referent window case–crossover designs. As the bias of the estimating functions has a known form depending only on β and $\langle z_{it} \rangle$ it is possible to remove it either by direct computation or by the parametric bootstrap. In the latter case simulations are used to compute the bias in $\hat{\beta}$ directly rather than in $\tilde{U}(\beta)$.

For this purpose we treat $\langle z_{it} \rangle$ as a realisation of some random process and ask what conditions on this process would make the bias small with high probability, so that conditional logistic regression approximates the distribution of $\hat{\beta}$ conditional on $\langle z_{it} \rangle$. Alternatively if the conditional bias has close to zero mean, inference could be done using the unconditional distribution of $\hat{\beta}$ if this can be calculated.

In air pollution epidemiology the exposures of interest and the most important potential confounders are identical or very similar for different people at the same time, so we focus primarily on this situation.

4 Design and analysis to reduce bias

The attributable risk due to air pollution is thought to be small for most health outcomes. If the same is true for other covariates then $\beta' z$ is small and we can approximate $\exp(\beta' z)$ by $1 + \beta' z$. Now if $\langle Z_{it} \rangle$ is a realisation of an independent sequence the bias in $\tilde{U}_i(\beta)$ is $o_P(\beta/\sqrt{T})$ and so will be negligible with high probability. When the process generating $\langle Z_{it} \rangle$ is autocorrelated the bias for small β is $O_p(\beta\bar{\rho})$ where $\bar{\rho}$ is the average autocorrelation between a case and its referents.

When $\beta' z$ is not small the bias is harder to characterise, but it may still be computed easily from the formulae derived in section 3·2. This allows us to explore the bias of $\tilde{U}_i(\beta)$ without requiring extensive simulation.

Figure 1 shows estimates of β from simulations based on the King Country cardiac arrest data. The exposure series is approximately 6 years (2092 days) of the daily PM_{10} series averaged over three monitors in and near Seattle. 362 cases, the number observed in the study by Siscovick et al. (1995) that provided our case data, are simulated from a proportional hazards model

$$\log \mathbf{E}\left[Y(t)\right] = \alpha + \beta \times PM_{10}(t)$$

where $\beta = 0.795$ is chosen to give a relative risk of 1.5 between the upper and lower quartiles of air pollution and α is the individual baseline hazard. Conditioning on the number of observed cases makes the value of α irrelevant. Varying numbers of controls were taken from the referent window of ± 30 days with days within ± 6 days of the case excluded. Conditional logistic regression was used to estimate β and Figure 1 shows the mean of $\hat{\beta}$ over 1000 simulations and a 95% confidence interval. There is a definite bias ranging from about 2.5% to nearly 10%. This bias is not due to residual confounding by season, which is not present in this simulation, but to the bias in the estimating functions.

The referent selection strategies of Figure 1 are not directly comparable with a stratified design. In Figure 2 we compare the mean of $\hat{\beta}$ for referent window case–crossover designs with 2, 4, 6 and 8 controls at weekly intervals centered around the case day and for stratified designs with 3, 5, 7 and 9 days at weekly intervals in each stratum, based on 15000 simulations. We can see from this picture that in this particular example the finite-sample bias, present in both methods, is of similar magnitude to the bias from the incorrect analysis with 2 or 4 controls but is smaller with 6 or 8 controls. Figure 3 compares the mean of the standardised estimating function

$$\bar{U}(\beta) = \frac{1}{\sqrt{n}} U(\beta)$$

for the referent window and stratified designs in the same 15000 simulations. This confirms that the referent window design does give rise to biased estimating functions but that the stratified design does not.

Figure 1: Mean and 95% confidence interval of regression coefficient β simulated from King county air pollution data and a true value of $\beta = 0.795$.

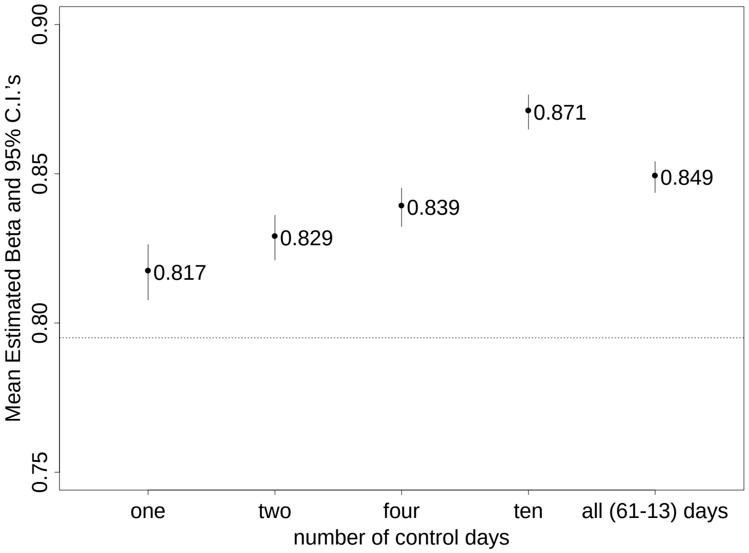


Figure 2: Mean and 95% confidence interval for regression coefficient β simulated from King county air pollution data and a true value of $\beta = 0.795$ with stratified (solid line) and referent window case–crossover (dashed line) sampling.

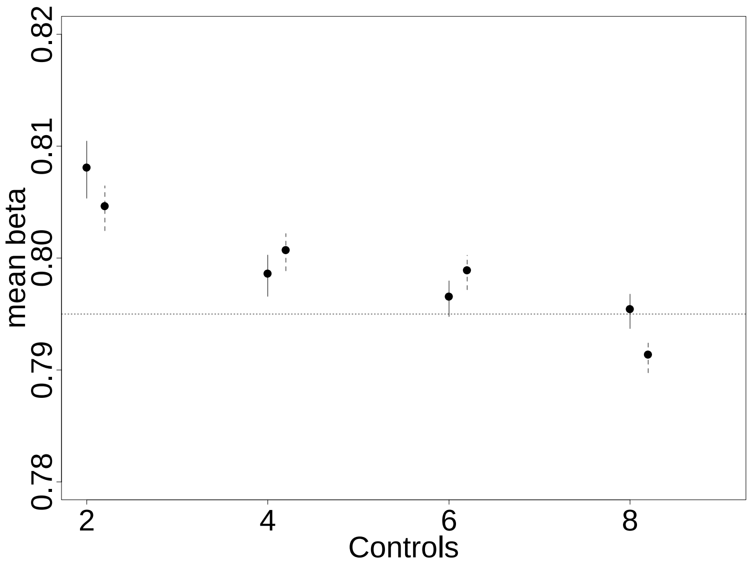
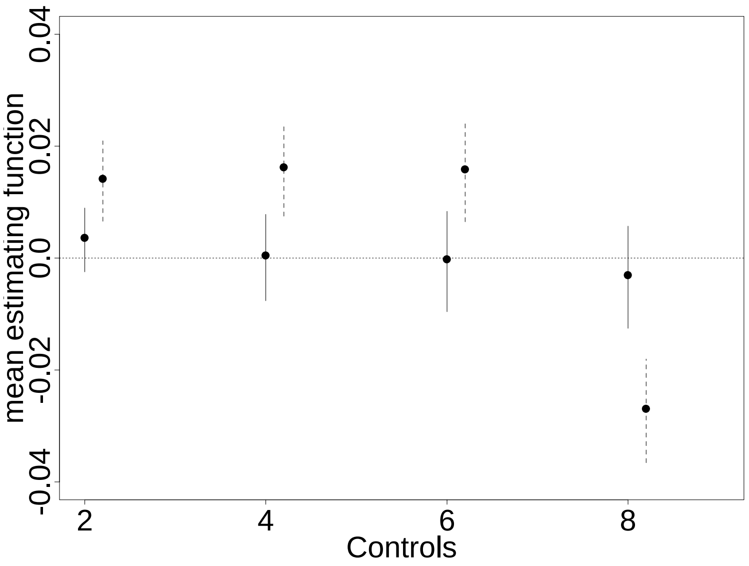


Figure 3: Mean and 95% confidence interval for estimating function with data simulated from King county air pollution data and a true value of $\beta = 0.795$ using stratified (solid line) and referent window case–crossover (dashed line) sampling.



We also used simulated exposure series to examine the range of bias that could be obtained from different sets of data. We present the bias in the standardised estimating function

$$\bar{U}(\beta) = \frac{1}{\sqrt{n}} U(\beta)$$

rather than in $\hat{\beta}$ as this avoids confusing the qualitatively different effects of biased estimating equations and small-sample deviations from asymptotic behaviour. For ease of interpretation we note that the bias in $\hat{\beta}$ is roughly 10 times that in the estimating functions.

The data were generated to fit the roughly lognormal distribution and 1 week short-term autocorrelation observed in PM series. Independent Normals with standard deviation 0.1 were generated and a six-term moving average with coefficients $(0.1, 0.2, 0.2, 0.3, 0.6, 1)$ was taken. The resulting variables were exponentiated to give 1000 days of simulated exposures $Z(t)$. A proportional hazards model

$$\mathbf{E}[Y(t)] = \lambda \exp(\beta Z(t))$$

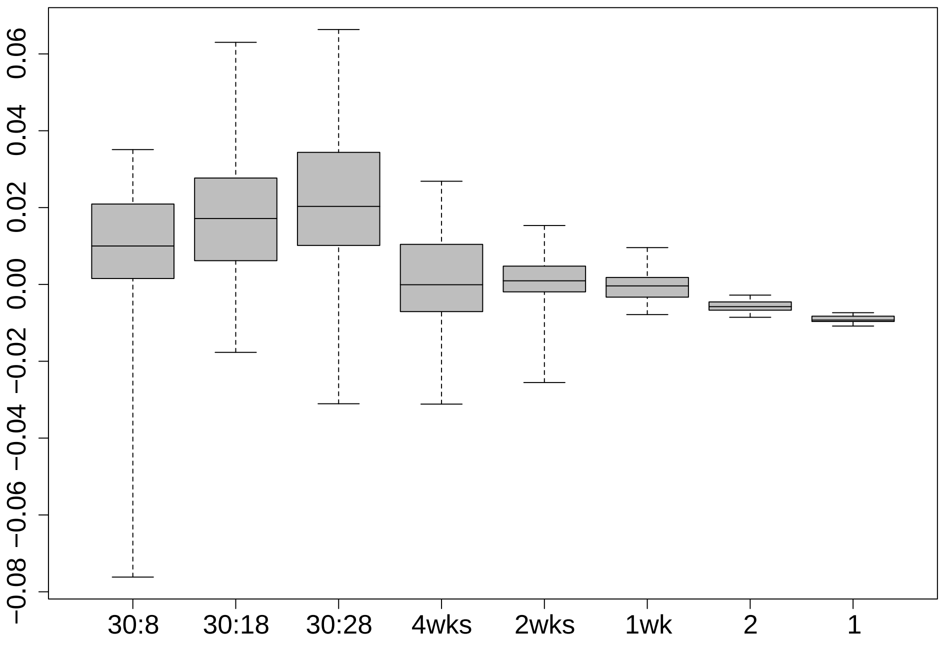
was then used to generate 100 cases. The regression coefficient was set at $\beta = 1$. We examined eight possible referent selection methods. The first three use all days from an outer limit of 30 days to an inner limit of 8, 18 or 28 respectively, in either direction. The next three use the same day of the week for the following and preceding 4, 2 or 1 weeks. The final two use the case day ±2 days and ±1 day respectively.

Figure 4 shows boxplots of the bias based on 50 realisations of the exposure series. In these simulations, and others not reported here, the bias is typically small when the 1, 2 or 4 week controls are used. In most cases the bias can be either towards or away from the null. Bias is still present when the exposure series is a realisation of an independent sequence, but is small and is equally likely to be towards or away from the null. The bias is smaller when the regression coefficient is smaller, and is zero when the regression coefficient is zero.

When there are multiple covariates it is possible that bias in the estimating function for one covariate will cause bias in the parameter estimate for other covariates correlated with it, even if their true regression coefficients are zero. This means that adjustments for meteorological factors or for copollutants such as sulfur dioxide, while important to remove confounding, may introduce additional bias.

We have considered the situation where the main exposure of interest is common to all subjects. In other areas of epidemiology where the case–crossover design is used the exposure histories for different subjects may be independent. In this case if there is also independence over time the biases conditional on the different exposure histories will be symmetrically distributed about zero and will average out over subjects. More precisely, as both the number of cases and the length of the time series increase the bias in the

Figure 4: Bias in estimating function for $\beta = 1$ and moderate autocorrelation with various referent selection strategies.



estimating function $\bar{U}(\beta)$ is $O_p(n^{-1/2})$ times the bias in an individual estimating function $U_i(\beta)$. In this case it may be more natural not to condition on the observed exposures. The unconditional distribution of $\hat{\beta}$ is unbiased, and we can regard the problem as an incorrect variance estimate for this unconditional distribution.

When exposure is independent across individuals the unconditional variance of $\hat{\beta}$ can be consistently estimated by the familiar sandwich method

$$\widehat{\mathrm{var}}\left[\hat{\beta}\right] = I^{-1}JI^{-1}$$

where I is the derivative of the estimating functions, so that I^{-1} is the usual variance estimate, and

$$J = \sum_{i=1}^{n} U_i(\hat{\beta})U_i(\hat{\beta})^T.$$

When there is a single exposure series or a correlation between individuals it is not possible to estimate the unconditional variance without further assumptions that decompose Z_t into deterministic parts that will be common to all realisations and stochastic parts that vary across realisations.

In our example we could assume that the air pollution measurements can be decomposed into a deterministic seasonal component and a stochastic component with short-term autocorrelation that would be independent across realisations.

5 Ambidirectional sampling and fatal events

To avoid serious selection biases from trends in exposure or risk over time both these methods use as referents some days after the event has occurred. Our analysis, and that of Navidi, have assumed that subjects are at risk even after an event. This is unreasonable for many interesting events such as death or primary cardiac arrest.

Navidi argues informally that as the air pollution history is known and fixed it cannot be affected by the occurence of an event and no bias is created by using this information. We find this argument plausible, but not completly compelling. The difficulty is that on referent days before a fatal event the subject was exposed and at risk but did not experience an event, whereas on referent days after the fatal event the subject was not exposed and was not at risk. In Navidi's analysis this distinction is ignored. The reason why this is valid is not explained, but is hidden in the rare disease assumption.

We can most easily consider this question analytically by introducing a model in which subjects are at risk at all times but the risk depends on the number of previous events. The situation of primary interest is then the limiting case where the risk drops to zero after an event. We can formulate this model as

$$\log \frac{p_{it}}{1-p_{it}} = \lambda_i + \beta z_t + \theta N_{it} \tag{5·1}$$

where p_{it} is the probability that subject i has an event on day t, λ_i is the constant baseline log odds for subject i, Z_t is air pollution on day t and N_{it} is the number of events subject i has experienced before time t.

When $\theta = 0$ a subject is at the same risk after an event as before and the analysis reduces to exactly the single event analysis described in section 3·1, as would be expected.

A fatal event, where the subject is no longer exposed and at risk is represented by the limiting case as $\theta \to -\infty$. When there is a trend in exposure N_{it} is correlated with Z_t and with the response and so is a potential confounder. If we fit the true model with N_{it} included then as $\theta \to -\infty$ observations after the event receive less and less weight. In the limit the referents taken after the event are ignored, and the protection from selection bias is lost.

The usual model, where N_{it} is omitted, is misspecified when $\theta \neq 0$. It still gives less biased results in the presence of trend than the model including N_{it}. When Z_t is stationary there will be no confounding by N_{it} and $\hat{\beta}$ will be at least approximately unbiased. When Z_t tends to increase over time the confounding will bias $\hat{\beta}$ downwards; when Z_{it} is decreasing $\hat{\beta}$ will be biased upwards. This is the reverse of the usual selection bias described by Greenland (1996). Estimating the magnitude of the bias requires a more thorough calculation, given in the Appendix. Under the null hypothesis $\beta = 0$ the resulting bias in $\hat{\beta}$ is approximately $-\Lambda\rho_{zt}$, where ρ_{zt} is the correlation between Z_t and t, a measure of the degree of trend, and Λ is the unconditional probability of death on any given day, which is typically very small.

We conclude that there is a bias from treating dead subjects as if they were still at risk, but that this bias is very small if the population rate of death is small, even when there is a trend in the exposure. Using ambidirectional referents can be expected to dramatically reduce the bias from trends in exposure in all reasonable examples, however the intuitive feeling that this procedure is not precisely valid is correct.

6 Conclusions

The standard conditional logistic regression analysis of case–crossover studies using restriction in time is not correct. Alternative designs that correctly condition on the exposure series provide conditionally unbiased relative risk estimates and valid conditional variances for these estimates. For example, time could be stratified by month and by day of week to create partitions with 3 or 4 referent days for each case. This stratification has the same robustness to trend in exposure as the more usual ambidirectional design. A minor disadvantage of the stratified design is that it gives fewer controls for the same maximum distance between case and control. A maximum 4 week gap leads to 3 controls in the stratified design and 6 controls in the windowed design.

The conditional logistic regression analysis will be approximately valid when the autocorrelation in exposure among case and referent days or between individuals is weak. For the specific case of particulate air pollution choosing 2–6 referent days at weekly intervals, symmetrically before and after the case day appears likely to give a small bias in the case–control analysis with the added benefit of matching on day of the week. In this example the bias was not much larger than the finite-sample bias, but in many case–crossover studies the finite-sample bias will be much smaller and the design bias may be larger.

It is also important to note that we have only discussed two forms of bias peculiar to the case–crossover study. Other forms of bias in the sampling of case and referent information, many of which are common to all observational studies, may be orders of magnitude larger.

This analysis has completely ignored the issue of exposure error and variability, which may be very important in studies of health effects of air pollution. Measurement error adjustment of the case–crossover design would be analytically challenging and would require assumptions about within-individual error distributions.

7 Acknowledgments

We are grateful to Lianne Sheppard and Scott Emerson for illuminating discussions on this topic. Part of this research is based on the PhD dissertation of the second author, funded by the Health Effects Institute. The first author is supported in part by the US Environmental Protection Agency through agreement CR825173–01–0 to the University of Washington, but this research 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.

A Calculation of bias for ambidirectional sampling with fatal events

Navidi (1998, p601) develops the multiple events model and explicitly considers the possibility that the risk may change with a subject's history although his calculations are not completely correct in this case as we will see. We use his more general model

$$\log \frac{p_{it}}{1-p_{it}} = \lambda_i + \beta^T X_{it} \qquad \text{(A·1)}$$

where X_{it} is the covariate vector for subject i at time t. In this model events are not necessarily rare, so we let A_i be the set of days on which subject i experiences an event, with n_i being the number of events. The likelihood of observing A_i is

$$\Pr(A_i) = \left(\prod_{t\in A_i} p_{it}\right)\left(\prod_{t\notin A_i} 1-p_{it}\right).$$

At this point it becomes useful to add another index to the daily probabilities and write p_{itA} for the unconditional probability that subject i has an event on day t if s/he also has events on the days in the set A that precede day t. This allows for the fact that X_{it} may depend on the subject's history and change as A changes. The conditional probability of A_i given n_i is then

$$\Pr(A_i \mid n_i) = \frac{\left(\prod_{t\in A_i} p_{itA}\right)\left(\prod_{t\notin A_i} 1-p_{itA}\right)}{\sum_{A\in\mathcal{A}_{n_i}} \left(\prod_{t\in A} p_{itA}\right)\left(\prod_{t\notin A} 1-p_{itA}\right)}$$

where $\mathcal{A}_{n_i}$ is the set of all sets of n_i times. If the risk does not depend on previous events, so $p_{itA} \equiv p_{it}$, the conditional likelihood simplifies to

$$\Pr(A_i \mid n_i) = \frac{\exp\left(\beta^T \sum_{t\in A_i} X_{it}\right)}{\sum_{A\in\mathcal{A}_{n_i}} \exp\left(\beta^T \sum_{t\in A} X_{it}\right)},$$

the conditional logistic regression likelihood obtained by Navidi.

When X_{it} depends on the subject's history, as in our case, this simplification does not occur. However, as $\theta \to -\infty$ the probability of observing more than one event per person goes to zero. Returning to the special case of equation 5·1 and assuming that all $n_i = 1$ we can simplify the conditional likelihood to

$$\Pr(A_i = \{t\} \mid n_i = 1) = \frac{\exp\left(\lambda_i + \beta Z_t\right) / \prod_{s\leq t}\left\{1+\exp\left(\lambda_i+\beta Z_s\right)\right\}}{\sum_{u=1}^{T}\exp\left(\lambda_i + \beta Z_u\right) / \prod_{s\leq u}\left\{1+\exp\left(\lambda_i+\beta Z_s\right)\right\}} \tag{A·2}$$

This is still complicated, so we consider the case when the null hypothesis is true: $\beta = 0$. We consider only one subject at a time and write Λ for $\exp(\lambda_i)$, suppressing the dependence on i. The conditional likelihood is then

$$\Pr((A_i = \{t\} \mid n_i = 1) = \frac{\Lambda/(1+\Lambda)^t}{\sum_{s=1}^{T}\Lambda/(1+\Lambda)^s}$$

We are interested in fitting the model in section 3·1, so we calculate the expectation of that score statistic under the true model

$$\begin{aligned}
\mathbf{E}\left[U(0)\right] &= \frac{1}{\sum_{s=1}^{T}\Lambda/(1+\Lambda)^s}\sum_{t=1}^{T}\left(z_t - \frac{1}{T}\sum_{u=1}^{T} z_u\right)\left\{\frac{\Lambda}{(1+\Lambda)^t}\right\} \\
&\approx \frac{1}{T\Lambda}\sum_{t=1}^{T}\left(z_t - \frac{1}{T}\sum_{u=1}^{T} z_u\right)\left(\Lambda - t\Lambda^2\right) \\
&= \frac{-\Lambda}{T}\sum_{t=1}^{T}\left(z_t - \frac{1}{T}\sum_{u=1}^{T} z_u\right) t
\end{aligned}$$

The approximation in the second line comes from a Taylor expansion and is accurate if $T\Lambda$, the risk of dying during the study period, is small. Using the fact that the information matrix is the covariance matrix of Z the resulting bias in $\hat{\beta}$ is approximately $-\Lambda\rho_{zt}$, where ρ_{zt} is the correlation between Z_t and t, a measure of the degree of trend.

References

Austin, H., Flanders, W. D., & Rothman, K. J. (1989). Bias arising in case–control studies from selection of controls from overlapping groups. *International Journal of Epidemiology* **18**, 713–716.

Greenland, S. (1996). Confounding and exposure trends in case–crossover and case–time–control designs. *Epidemiology* **7**, 231–239.

Kelsall, J. E., Zeger, S. L., & Samet, J. M. (1999). Frequency domain log-linear models; air pollution and mortality. *Applied Statistics* to appear.

Maclure, M. (1991). The case–crossover design: A method for studying transient effects on the risk of acute events. *American Journal of Epidemiology* **133**, 144–153.

Mittelman, M. A., Maclure, M., & Robins, J. M. (1995). Control sampling strategies for case–crossover studies: An assessment of relative efficiency. *American Journal of Epidemiology* **142**, 91–98.

Navidi, W. (1998). Bidirectional case–crossover designs for exposures with time trends. *Biometrics* **54**, 596–605.

Samet, J. M., Zeger, S. L., & Berhane, K. (1995). The association of mortality and particulate air pollution. In *Particulate Air Pollution and Daily Mortality: Replication and Validation of Selected Studies*, pages 1–104. Health Effects Institute: Cambridge, MA.

Siscovick, D. S., Raghunathan, T. E., King, I., Weinmann, S., Wicklund, K. G., Albright, J., Bovbjerg, V., Arbogast, P., Smith, H., Kushi, L. H., et al. (1995). Dietary intake and cell membrane levels of long-chain n-3 polyunsaturated fatty acids and the risk of primary cardiac arrest. *Journal of the American Medical Association* **274**, 1363–7.