I. Introduction

The case-crossover design was proposed as a method for estimating the association between a short-term exposure and the risk of a rare event.[1] Only cases are required for such a study. For each case, exposure at the "index" time (the hazard or at-risk period prior to the event) is compared to exposure at comparable control, or "referent" times. The design can be viewed as a hybrid of the case-control and the crossover designs: selection is based on the outcome, similar to the case-control design, but the case serves as his/her own control, as in a crossover study. By making within-person comparisons, confounding due to time-independent factors is eliminated and, with proper selection of referents, time-dependent confounders can also be controlled.

The case-crossover design has been widely used to study the effect of air pollution exposure on the risk of an adverse health event. A variety of different referent selection strategies have been used in implementing this design (see, for example [2-8]). However, the choice of a referent scheme is of particular importance with air pollution exposure data. Representative referents are more difficult to define in an exposure series with substantial structure; air pollution often has a long term time trend, and varies with season and day of the week. These factors are also often associated with the outcome of interest, and hence confounding is a major issue. In addition, the design relies on the assumption of a constant distribution of exposure across referent times.[6;9] Yet the choice of referent selection strategy is important for a more fundamental reason. Lumley and Levy [5] showed that certain referent schemes induce bias in the estimating equations that are typically used. We call this bias in the estimating equations "overlap bias"; it is a version of the bias induced from choosing non-disjoint strata to partition the population in a matched case-control study.[5;10;11] We begin this paper with a description of the case-crossover design,

including estimation and referent selection methods. We discuss the bias associated with variation in the exposure over time and confounding under certain referent selection strategies, and focus in particular on considerations important in choosing the referent selection strategy with air pollution exposure series. In Section 4, we give a derivation of overlap bias, and describe the situations in which it is a concern, as well as how it can be avoided. In Section 5, we calculate the magnitude of this bias for a given exposure series, and explore how the bias depends on properties of the exposure series in the absence of confounding. We conclude with a discussion of the implications of these results on the use of the case-crossover design.

2. Overview of the Method

The case-crossover design compares exposures at index times, when the exposure triggering an event may occur, with exposures at referent times, when an event is not triggered. The design is suitable for assessing the risk of an acute-onset event during a brief period following a transient exposure, and for exposures that are thought not to have carryover effects. Its main advantages are the fact that that it achieves control over fixed (time-invariant) confounders, and, with selection of suitable referents, controls for some effects of time-varying exposures by design, rather than by statistical modeling. However, there are many challenges in implementing this design, including, but not limited to, determination of the duration and timing of the hazard period prior to the event, defining the exposure, and selecting referent exposures that are representative of the underlying distribution of exposure.

The underlying model for the case-crossover design is typically the proportional hazards model for a rare disease with a constant baseline hazard for each individual. The model for the hazard

rate of person i at time t given time-varying covariates x_{it} is $\lambda_i(t; x_{it}) = \lambda_i \exp(x_{it}\beta)$ (see [6] for a logistic regression formulation applicable to more common outcomes). Over a short time period, the constant baseline hazard assumption (λ_i) is often reasonable, and is equivalent to assuming smooth seasonal effects in a time series analysis.

Assuming an appropriate referent sampling scheme, an unbiased estimate the effect of short-term exposure can be obtained using conditional logistic regression. This is similar to the analysis of matched case-control studies; here, the exposure at each index time is part of a "matched set" of exposures consisting of exposures for the same subject at his/her referent times. Consider the situation in which we have a shared exposure series, z, defined at times $t = 1,2,\ldots,T$ common to all $i = 1,2,\ldots,n$ subjects. Denote the index time for subject i by t_i, the exposure at the index time by z_{t_i}, and let W_i represent the referent window for subject i (including the index time and all referent times). The conditional logistic regression estimating equations are

$$\sum_{i=1}^{n} U_i(\beta) = \sum_{i=1}^{n} z_{t_i} - \sum_{t \in W_i} z_t \frac{e^{z_t \beta}}{\sum_{s \in W_i} e^{z_s \beta}}. \qquad (1)$$

For each subject, the estimating equation is the difference between exposure at the index time and a weighted average of exposure at all times in the referent window; hence, the estimate of β obtained represents the change in the risk of an event associated with a short-term unit increase in exposure. We will discuss the use of the conditional logistic regression estimating equations in Section 4; it is only for some referent selection strategies that these equations have mean zero.

Typically, exposure data for air pollution studies come from centrally located monitors of ambient pollution, and thus exposure is common to all individuals. Consequently, research on the case-crossover design in the air pollution exposure context should focus on properties of the effect estimate conditional on a known and shared exposure series. We want to know that the estimate and the corresponding estimating equations are unbiased for *a given* exposure series; knowing that they are unbiased when averaged over all possible exposure series is of little value when there is a single exposure. In this paper, we restrict our attention to models that condition on exposure.

3. Referent Selection Strategies

In a typical air pollution time series study, daily event counts (e.g. mortality or hospital admissions) are regressed on the shared exposure series using a Poisson regression model (see [12] for a review of this approach). Strong confounding effects of season and weather must be controlled statistically in the model. It is instructive to note that the case-crossover design is equivalent to a Poisson regression analysis except that confounding is controlled for by design (by matching) instead of in the regression model.[5] Restricting referents to the same day of week and season as the index time controls for these confounding effects by design.

Distinct from confounding is another concern regarding time trend in the exposure series. If there is a long-term time trend, choosing referents only prior to the index day may lead to bias. Navidi [6] proposed that the time trend bias could be eliminated by choosing referents both before and after the index time, a strategy called *bidirectional* referent selection. Technically, bidirectional sampling is only valid when cases are still at risk after the event, an assumption that

is certainly violated when the event is death. However, Lumley and Levy [5] showed that, with a rare event, the bias due to sampling referents after the at-risk period is very small. Most importantly, the bias from sampling referents outside of the time at risk is smaller than the bias associated with the alternative of only sampling referents prior to the index time in the presence of a time trend. Because of concerns over exposure trends, bidirectional referent selection strategies have been almost universally adopted among researchers performing case-crossover analyses of air pollution data.

Figure 1 displays an array of possible referent sampling strategies used in air pollution studies. Figure 1a shows the total history unidirectional referent sample, in which referents are all days in the exposure series prior to the index day. Because referents can be long before the index day, this design will be subject to bias due to non-stationarity of the exposure, even in the absence of confounding by time trend, season, and/or day of the week. This strategy could be improved by restricting referents to be closer to the index time, such as no more than 30 days prior to the index time. Figures 1b through 1d give examples of restricted unidirectional referent sampling designs. All of these approaches fix the location of the referents to be prior to and relative to the index time. Figure 1d shows referents placed 7, 14, and 21 days prior to the index day in order to control for day-of-week confounding.

Figures 1e through 1g show examples of bidirectional referent sampling. Figure 1e shows the full stratum bidirectional referent sample [6] where the referents are all days in the exposure series other than the index day. With this design, season must be modeled or the air pollution effect estimates will be confounded. A popular alternative to modeling is restriction in time using

symmetric bidirectional sampling. Referents are selected at fixed intervals before and after the index time (see e.g. [2;3;8;13] for applications). This referent selection strategy controls for bias due to time trend, season, and day of the week, if lags are multiples of seven.[14] Figure 1f shows the symmetric bidirectional design with two referents 7 days before and after the index time, and panel g shows a strategy with four referents 7 and 14 days before and after the index time.

The main problem we focus on with respect to the referent schemes in Figure 1a-1d, and 1f-1g is overlap bias. These designs are all examples of *non-localizable referent window strategies*. With these designs, an unbiased estimating equation that is restricted to the referent windows does not exist. Thus, the conditional logistic regression estimating equations, which are restricted to exposures within the referent windows, are biased. We call this bias *overlap bias*, and discuss its magnitude in Section 5.

In contrast, we define *localizable referent designs* as those for which there exists an unbiased estimating equation restricted to the referent windows. Localizability is a desirable property since it allows us to obtain unbiased estimates by making comparisons within referent windows. Since these are usually relatively short periods of time, we are able to control for confounding in this way. The time stratified [5] and the semi-symmetric bidirectional [16] designs are both localizable designs.

The time stratified referent strategy divides time into disjoint strata, uses the index time to determine which stratum is selected for a given case, and selects all or a sample of the remaining times in the stratum as referents. This strategy maintains control over confounders by design

since days within each stratum are matched on important confounders. A common stratification in the air pollution setting is to select referents that fall on the same day of the week and in the same month and year as the index day. This controls for confounding due to season and day of the week. In addition, there is no bias due to time trend since there is no pattern in the placement of referents relative to the index time. The full stratum bidirectional referent strategy is a special case of the time stratified design where there is only one stratum; with this stratification, however, confounders such as season cannot be assumed to be constant within stratum.

With semi-symmetric bidirectional referent selection, one referent is randomly chosen from the pair of days a fixed lag pre- and post-event; if only one is available (e.g. at the beginning and end of a time series), it serves as the referent. Confounding is controlled by design when an appropriate lag is selected.

The semi-symmetric bidirectional design is not a time stratified design, but a slight modification of it is. We call this the adjusted semi-symmetric bidirectional referent strategy. For events at the beginning or end of the time series, the referent strategy still randomly selects from the pair of days pre- and post-event. However, when the referent chosen is outside the series, the case is excluded. This modified design can also be thought of as randomly choosing between two overlapping sets of disjoint strata, where each set is based on a different time stratification. For each day (except those at the end of the series) there are two possible strata (e.g. if the lag is one day, a case on day two has days one and two as one possible stratum, and days two and three as another possible stratum). One stratum is randomly selected for each case, and cases that fall outside the selected stratum are excluded.

Though the time stratified and semi-symmetric bidirectional designs are both localizable, they differ in one important respect. We further partition the localizable designs into two groups called *ignorable* and *non-ignorable* designs to highlight this difference. With an ignorable design, such as the time stratified design, the referent sampling scheme can be ignored in conducting the analysis, and the conditional logistic regression estimating equations are unbiased. However, with a non-ignorable design, such as the semi-symmetric bidirectional design, the referent sampling scheme cannot be ignored. The likelihood of the data depends on the sampling scheme, and must be used for an unbiased analysis. Alternatively, conditional logistic regression could be used if an appropriate offset were specified. (See Section 4 for a derivation of the likelihood of the data under the different types of referent strategies.) Thus, the distinction between ignorable and non-ignorable designs refers to whether standard conditional logistic regression can be used for an unbiased analysis.

4. Overlap Bias

The use of conditional logistic regression for analysis of the case-crossover design has been motivated by the analogy to matched case-control designs, where conditional logistic regression maximizes the true (conditional) log-likelihood of the data. The effect estimate, β, is found by setting the conditional logistic regression score equations (1) equal to zero. With localizable and ignorable designs, such as the time stratified design, equation (1) is the derivative of the true conditional log-likelihood. We derive the likelihood as follows. Since the index times are random conditional on the referent windows, we can use the likelihood of the index times conditional on the referent windows to estimate β. Consider again the situation in which we

have an exposure series, $\mathbf{z}$, defined at times $t = 1,2,\ldots,T$ common to all $i = 1,2,\ldots,n$ subjects. Denote the index time for subject i by t_i, the exposure at the index time by z_{t_i}, and let W_i represent the referent window for subject i (including the index time and all referent times). Let Y_{it} be an indicator of whether subject i's index time was on day t. We assume that events follow a proportional hazards model in which the likelihood of an event occurring at time t for subject i is $\lambda_i e^{z_t \beta}$. We condition on the exposure series, the referent windows, and assume that events are rare. We note that subjects without events do not contribute information. The likelihood is

$$\prod_{i=1}^{n} P(T_i = t_i \mid \mathbf{z}, W_i, \sum_{s=1}^{T} Y_{is} = 1) = \prod_{i=1}^{n} \frac{P(T_i = t_i, \sum_{s=1}^{T} Y_{is} = 1 \mid \mathbf{z}, W_i)}{\sum_{t=1}^{T} P(T_i = t, \sum_{s=1}^{T} Y_{is} = 1 \mid \mathbf{z}, W_i)}$$

$$= \prod_{i=1}^{n} \frac{\lambda_i e^{z_{t_i} \beta}}{\sum_{t \in W_i} \lambda_i e^{z_t \beta}}$$

$$= \prod_{i=1}^{n} \frac{e^{z_{t_i} \beta}}{\sum_{t \in W_i} e^{z_t \beta}} . \qquad (2)$$

The λ_i's cancel, so we can use cases only to estimate β, and need not estimate λ_i. We note that the likelihood depends only on exposures at times within the referent windows. In addition, we find that the derivative of the logarithm of (2) is exactly the conditional logistic regression estimating equation (1), and thus the conditional logistic regression estimating equations have mean zero if localizable, ignorable referents are used.

If, however, non-localizable referent windows are used, the conditional logistic regression estimating equations are not the derivatives of the log-likelihood of the data. In order to see this,

we first note that the non-localizable designs are exactly those for which the location of the index times within the referent windows are fixed (i.e., with symmetric bidirectional referents, the index time is the time in the center of the referent window). Hence, with non-localizable referents, the likelihood of the index times conditional on the referent windows is

$$\prod_{i=1}^{n} P(T_i = t_i \mid \mathbf{z}, W_i, \sum_{s=1}^{T} Y_{is} = 1) = \prod_{i=1}^{n} \frac{P(T_i = t_i, \sum_{s=1}^{T} Y_{is} = 1 \mid \mathbf{z}, W_i)}{\sum_{t=1}^{T} P(T_i = t, \sum_{s=1}^{T} Y_{is} = 1 \mid \mathbf{z}, W_i)} = \prod_{i=1}^{n} \frac{1}{\sum_{t \in W_i} I_{[t=t_i]}} = 1,$$

where $I_{[A]} = 1$ if A is true, and zero otherwise. This likelihood is uninformative, since knowing a referent window gives the associated index time exactly. With non-localizable designs, the random variables are actually the referent windows, which fall around each index time with probability $\lambda_i e^{z_i \beta}$. The appropriate likelihood to use is that of the referent windows, unconditional on the index times. Again, we condition on there being one event for each case. The marginal likelihood of the referent windows is

$$\prod_{i=1}^{n} P(W_i = w_i \mid \mathbf{z}, \sum_{s=1}^{T} Y_{is} = 1) = \prod_{i=1}^{n} \frac{P(W_i = w_i, \sum_{s=1}^{T} Y_{is} = 1 \mid \mathbf{z})}{\sum_{w} P(W_i = w, \sum_{s=1}^{T} Y_{is} = 1 \mid \mathbf{z})}$$

$$= \prod_{i=1}^{n} \frac{\lambda_i e^{z_{t_i} \beta}}{\sum_{w} \lambda_i e^{z_{t_w} \beta}}$$

$$= \prod_{i=1}^{n} \frac{e^{z_{t_i} \beta}}{\sum_{t=1}^{T} e^{z_t \beta}}. \qquad (3)$$

Note that the likelihood depends on exposure at all times in the exposure series. Thus, the derivative of (3) is not the conditional logistic regression estimating equations (1), which depend

only on exposure within the referent windows. It can be shown that (1) does not have mean zero for non-localizable referent window sampling schemes (see Section 5). Hence, there is overlap bias associated with the use of the conditional logistic regression estimating equations under non-localizable referent selection.

We want to emphasize the difference in the likelihoods for the non-localizable and localizable, ignorable referent schemes. Note that there are two random variables in each referent scheme: the index times, t_i, and the referent windows, W_i. With a non-localizable referent window scheme, t_i and W_i are simple functions of each other, and hence the likelihood of one conditional on the other is uninformative. Thus, we use the marginal likelihood of the referent windows to estimate β. A marginal likelihood could also be calculated for the localizable, ignorable designs (instead of the conditional likelihood), but it would not yield a useful estimate of β. The estimate would be entirely confounded, due to variables such as season which are associated both with index times (and hence the referent windows selected) and with exposure.

We also examine the likelihood of the data under the semi-symmetric bidirectional design, an example of a localizable, but non-ignorable design. With this strategy, the referent window for a subject with an index time at t_i is $W_i = \{ t_i - \delta, t_i \}$ with probability 0.5, and $W_i = \{ t_i, t_i + \delta \}$ with probability 0.5, for some lag δ.[16] We can again use the conditional likelihood to estimate β. The conditional likelihood can be written as

$$\prod_{i=1}^{n} P(T_i = t_i \mid \mathbf{z}, W_i, \sum_{s=1}^{T} Y_{is} = 1) = \prod_{i=1}^{n} \frac{P(W_i = w_i \mid \mathbf{z}, T_i = t_i, \sum_{s=1}^{T} Y_{is} = 1) \cdot P(T_i = t_i, \sum_{s=1}^{T} Y_{is} = 1 \mid \mathbf{z})}{\sum_{t=1}^{T} P(W_i = w_i \mid \mathbf{z}, T_i = t, \sum_{s=1}^{T} Y_{is} = 1) \cdot P(T_i = t, \sum_{s=1}^{T} Y_{is} = 1 \mid \mathbf{z})}$$

$$= \prod_{i=1}^{n} \frac{P(W_i = w_i \mid \mathbf{z}, T_i = t_i, \sum_{s=1}^{T} Y_{is} = 1) \cdot P(T_i = t_i, \sum_{s=1}^{T} Y_{is} = 1 \mid \mathbf{z})}{\sum_{t \in W_i} P(W_i = w_i \mid \mathbf{z}, T_i = t, \sum_{s=1}^{T} Y_{is} = 1) \cdot P(T_i = t, \sum_{s=1}^{T} Y_{is} = 1 \mid \mathbf{z})}$$

$$= \begin{cases} \prod_{i=1}^{n} \frac{\pi(W_i \mid t_i) \cdot \lambda_t e^{z_{t_i}\beta}}{\pi(W_i \mid t_i) \cdot \lambda_t e^{z_{t_i}\beta} + \pi(W_i \mid t_i - \delta) \cdot \lambda_t e^{z_{t_i - \delta}\beta}}, W_i = \{t_i - \delta, t_i\} \\ \prod_{i=1}^{n} \frac{\pi(W_i \mid t_i) \cdot \lambda_t e^{z_{t_i}\beta}}{\pi(W_i \mid t_i) \cdot \lambda_t e^{z_{t_i}\beta} + \pi(W_i \mid t_i + \delta) \cdot \lambda_t e^{z_{t_i + \delta}\beta}}, W_i = \{t_i, t_i + \delta\} \end{cases} \quad (4)$$

where $\pi(W_i \mid t) = P(W_i = w_i \mid \mathbf{z}, T_i = t, \sum_{s=1}^{T} Y_{is} = 1)$. We note that the likelihood depends only on times within the referent windows, but that the form of the likelihood depends on the referent sampling scheme. With semi-symmetric bidirectional referent selection, $\pi(W_i \mid t_j) = 0.5$ if t_j is in the middle of the exposure series, but index times at the beginning or end of the exposure series have only one possible referent window. (If $\pi(W_i \mid t_j) = \pi(W_i \mid t_k)$ for all $t_j, t_k \in W_i$, the likelihood would reduce to the conditional logistic regression likelihood (2).) This likelihood, (4), must be used in order to obtain an unbiased estimate of β. Yet, it is apparent from equation (4) that if we use an offset term of log2 for days at the beginning and end of the series in a conditional logistic regression analysis, we will get the same estimate of β as if we had used (4) to estimate β, since $0.5 \cdot e^{\log(2) + z_t\beta} = e^{z_t\beta}$.

A final observation concerns the adjusted semi-symmetric bidirectional design, a localizable and ignorable design. With this strategy, if we randomly select a referent window at the beginning or end of the series that doesn't exist, we drop the case from the analysis. This is equivalent to weighting cases at the beginning and end of the series by 0.5, which would ensure that the weights (π) in (4) cancel. Thus, with adjusted semi-symmetric bidirectional referent selection,

the likelihood of the data is exactly the conditional logistic regression likelihood, and (1) can be used to obtain an unbiased estimate of β.

5. Magnitude of the Overlap Bias

Lumley and Levy [5] were the first to point out the existence of overlap bias with non-localizable referent window selection. With both non-localizable and localizable but non-ignorable designs, the conditional logistic regression estimating equations do not have mean zero. The expected value of the estimating equation for one subject, regardless of the referent selection strategy, is

$$E_{t_i}(U_i(\beta)) = \sum_{t_i=1}^{T} \frac{e^{z_{t_i}\beta}}{\sum_{s=1}^{T} e^{z_s\beta}} \left(z_{t_i} - \sum_{u \in W_i} z_u \frac{e^{z_u\beta}}{\sum_{v \in W_i} e^{z_v\beta}} \right). \quad (5)$$

This expectation is the sum over all possible times of the estimating equation weighted by the *marginal* distribution of index times, $P(T = t \mid \mathbf{z}, \sum_{s=1}^{T} Y_{is} = 1) = e^{z_t\beta} / \sum_{s=1}^{T} e^{z_s\beta}$. Note here that we do not want to condition on the referent window, since we want to compute the expectation over all index times. If we rewrite (5) with $t = t_i$ and $W_i = W_t$, we obtain

$$E_{t_i}(U_i(\beta)) = \sum_{t=1}^{T} \nu_t \left(z_t - \overline{z(W_t)} \right),$$

where $\nu_t = e^{z_t\beta} / \sum_{s=1}^{T} e^{z_s\beta}$ are weights, W_t is the referent window to which t belongs, and $\overline{z(W_t)} = \sum_{u \in W_t} z_u e^{z_u\beta} / \sum_{v \in W_t} e^{z_v\beta}$ is a referent window-weighted average of exposure.

For the time stratified design, we can rewrite the sum over all times as a sum over strata $s = 1, \ldots, S$ and time within strata, since time is partitioned into a complete set of disjoint strata. Thus, we can write (5) as

$$E_{t_i}(U_i(\beta)) = \sum_{s=1}^{S} \frac{1}{\sum_{u=1}^{T} e^{z_u\beta}} \left(\sum_{t \in s} z_t e^{z_t\beta} - \sum_{u \in s} z_u e^{z_u\beta} \cdot \frac{\sum_{t \in s} e^{z_t\beta}}{\sum_{l \in s} e^{z_l\beta}} \right) = 0 ,$$

since the term in parentheses is zero.

With other referent selection strategies, time is not partitioned into disjoint sets, and there is no simplification that reduces the expectation to zero. However, expression (5) gives us the magnitude of the overlap bias in the conditional logistic regression estimating equations under any referent strategy for a given exposure series. Examination of this expression reveals that, when $\beta = 0$, bias occurs only if the referent strategy differs among times in the exposure series. (This occurs, for example, in the symmetric bidirectional design, in which cases occurring at the beginning of the exposure series do not have pre-event referents, and cases at the end do not have post-event referents.) Bateson and Schwartz [15] suggest subtracting off this bias, but this only accounts for the overlap bias at $\beta = 0$. We also note that an unbiased estimating equation could be obtained by subtracting off the bias, (5), from the conditional logistic regression estimating equations, (1). However, this unbiased estimating equation is not local; it depends on exposures at all times in the exposure series.

We would like to examine the overlap bias on the β scale, rather than the estimating equation scale, so we use the approximation

$$\hat{\beta} - \beta \approx \frac{E_{t_i}(U_i(\beta))}{-E_{t_i}(\frac{\delta}{\delta\beta}U_i(\beta))}. \quad (6)$$

The denominator can be found by differentiating (1) with respect to β and taking the expectation over all possible times as in (5). We use the approximation in (6) to show, for a given exposure series, the large sample bias in the effect estimate, $\hat{\beta}$.

In order to explore bias in the estimate of β under various referent selection strategies, we simulated three different types of exposure series. This also allows us to determine how the bias depends on properties of the exposure series. Note that, because we consider the exposure series to be fixed rather than random, properties of the exposure series, such as autocorrelation, should be thought of as mathematical descriptions of the series, rather than as parameters of the distribution of exposures. We simulated exposures to mimic PM_{10} (particulate matter less than 10 μm in diameter) in Seattle over a one year period. Three lognormal exposures were generated, with 1. No temporal structure (Figure 2); 2. Serial correlation on adjacent days (ρ = 0.6; Figure 3); and 3. Serial correlation and a time-dependent mean and variance (Figures 4 and 5). All lognormal exposures had a mean of 3.6 and a variance of 0.2. Mean and variance structure added included a decreasing temporal trend, day of the week effect, seasonality (modeled with sine and cosine terms with periods of $2, 1, \frac{1}{2}, \frac{1}{3}, \text{and} \frac{1}{4}$ of a year), and seasonal variance (modeled with sine and cosine terms with periods of 1 and $\frac{1}{2}$ of a year). Lognormal exposures were then exponentiated to yield simulated exposures, z_t.

We show, in Figures 2, 3, 4, and 5, the large sample bias for three randomly chosen realizations of each type of exposure series as a function of β. For each type of exposure, the same three series are shown across all referent strategies. The bias is shown for each of the referent window strategies in Figure 1, as well as for the time stratified and semi-symmetric bidirectional designs. We don't show the bias for the full stratum bidirectional design, however, since this is a special case of the time stratified approach, and hence the bias is zero for all β. Note that, with the semi-symmetric bidirectional design, we show the bias averaged over all realizations of the referent strategy; for a given realization, the bias would be larger. For the series with time-varying mean and variance, we show the bias on two different scales: a larger window in which β varies from -0.2 to 0.2, and a smaller window in which β varies from -0.01 to 0.01 (Figures 4 and 5), while for the unstructured (Figure 2) and autocorrelated (Figure 3) series we only show bias on the smaller scale. The smaller scale shows effect estimates on the order of those typical in air pollution studies, since a β of 0.0048 corresponds to a relative risk of 1.1 for one interquartile range change in exposure (where IQR = 20). It should be emphasized that the bias shown in Figures 2, 3, 4 and 5 is overlap bias only. These figures do not include bias due to confounding, or address the issue of which referent strategies best deal with this problem.

Figures 2, 3, 4, and 5 suggest a few points regarding overlap bias. As expected, there is no overlap bias for the time stratified and full stratum bidirectional designs. In some cases, there is substantial bias for the unidirectional designs, but the bias with the various symmetric bidirectional designs is small. The bias in the effect estimate is generally small, especially for small β, but tends to increase with more structure in the exposure series. However, these results depend on the particular realization of the exposure series. For some exposures, such as the three

shown with time varying mean and variance, there is substantial bias in the effect estimate with unidirectional referents, even for small β. It is possible that, for some exposures, the bias may be larger than β itself. In addition, for some exposures, there may be bias even when $\beta = 0$. Finally, for some exposures, there is bias apparent with certain referent selection strategies, but not with others. Hence, it is clear that the magnitude and direction of overlap bias cannot be predicted before the exposure is known. It seems logical that one should choose a referent selection strategy that avoids this bias entirely.

6. Discussion and Conclusions

The case-crossover design is well suited to the study of the acute health effects of air pollution, but proper selection of referents is particularly important. These control exposures must be representative of the exposure that generated the outcome, and also must be chosen to control for confounding (or confounding effects must be otherwise included in the model). The point we have emphasized here, however, is that an analysis method must be paired with an appropriate referent selection strategy in order to ensure unbiased estimating equations.

We have presented a new taxonomy of referent selection strategies. Most commonly used referent designs (e.g. the symmetric bidirectional design) are non-localizable, and are subject to overlap bias. Within the localizable designs are the non-ignorable designs, such as the semi-symmetric bidirectional design. These strategies require analysis using the true likelihood of the data (or specifying an appropriate offset in the conditional logistic regression analysis). If conditional logistic regression estimating equations are used (without an offset), overlap bias will

be incorporated. With localizable, ignorable designs, such as the time stratified design, there is no overlap bias associated with a standard conditional logistic regression analysis.

Our studies of overlap bias indicate that it is usually small, as was suggested by Lumley and Levy.[5] However, we have also seen that the bias is quite unpredictable in magnitude and direction; it depends on the particular shared exposure series. The bias will be exacerbated by model shopping on referent series, since the magnitude and direction of the bias depends on the particular realization of the exposure series. We have shown that it is possible to have bias even for small β, and, indeed, even when $\beta = 0$. While it is possible to estimate the expected bias as a function of β for any specific exposure series, it is very difficult to predict in advance the extent of the problem. Therefore, it seems most prudent for researchers to choose a referent selection strategy that avoids overlap bias entirely.

Two of the main strengths of the case-crossover design in the air pollution context are that it controls fixed confounders and allows for control over time-dependent confounders by design. Choosing referents that are restricted to the same day of the week and season as the index time will control for these effects. Though we have not examined the degree to which various referent selection strategies control bias due to confounding, several points are clear. Referents that are closer to the index time will achieve greater control of seasonal confounding. However, sampling referents closer to the index time will also result in a loss of power due to less heterogeneity in the exposure series, and a smaller number of referents will also result in lower power.

Acknowledgements: This work was supported by the EPA Northwest Center for Particulate Matter and Health grant number R827355. The contents of this article do not necessarily reflect the views and policies of the EPA.

References

(1) Maclure M. The Case-Crossover Design - A Method for Studying Transient Effects on the Risk of Acute Events. *American Journal of Epidemiology* 1991; **133**(2):144-153.

(2) Kwon HJ, Cho SH, Nyberg F, Pershagen G. Effects of Ambient Air Pollution on Daily Mortality in a Cohort of Patients with Congestive Heart Failure. *Epidemiology* 2001; **12**:413-419.

(3) Lee JT, Schwartz J. Reanalysis of the effects of air pollution on daily mortality in Seoul, Korea: A case-crossover design. *Environmental Health Perspectives* 1999; **107**:633-636.

(4) Levy D, Sheppard L, Checkoway H, Kaufman J, Lumley T, Koenig J et al. A case-crossover analysis of particulate matter air pollution and out-of-hospital primary cardiac arrest. *Epidemiology* 2001; **12**:193-199.

(5) Lumley T, Levy D. Bias in the case-crossover design: implications for studies of air pollution. *Environmetrics* 2000; **11**:689-704.

(6) Navidi W. Bidirectional case-crossover designs for exposures with time trends. *Biometrics* 1998; **54**:596-605.

(7) Peters A, Dockery DW, Muller JE, Mittleman MA. Increased particulate air pollution and the triggering of myocardial infarction. *Circulation* 2001; **103**:2810-2815.

(8) Sunyer J, Schwartz J, Tobias A, Macfarlane D, Garcia J, Anto JM. Patients with chronic obstructive pulmonary disease are at increased risk of death associated with urban particle air pollution: A case-crossover analysis. *American Journal of Epidemiology* 2000; **151**:50-56.

(9) Greenland S. Confounding and exposure trends in case-crossover and case time-control designs. *Epidemiology* 1996; **7**:231-239.

(10) Austin H, Flanders WD, Rothman KJ. Bias arising in case-control studies from selection of controls from overlapping groups. *International Journal of Epidemiology* 1989; **18**:713-716.

(11) Robins J, Pike M. The validity of case-control studies with nonrandom selection of controls. *Epidemiology* 1990; **1**:273-284.

(12) Dominici F, Sheppard L. Health effects of air pollution: a statistical review. *International Statistical Review* 2003 (in press).

(13) Neas LM, Schwartz J, Dockery D. A case-crossover analysis of air pollution and mortality in Philadelphia. *Environmental Health Perspectives* 1999; **107**:629-631.

(14) Bateson TF, Schwartz J. Control for seasonal variation and time trend in case crossover studies of acute effects of environmental exposures. *Epidemiology* 1999; **10**:539-544.

(15) Bateson TF, Schwartz J. Selection bias and confounding in case-crossover analyses of environmental time-series data. *Epidemiology* 2001; **12**:654-661.

(16) Navidi W, Weinhandl E. Risk set sampling for case-crossover designs. *Epidemiology* 2002; **13**(1):100-105.

Figure 1 Referent selection strategies

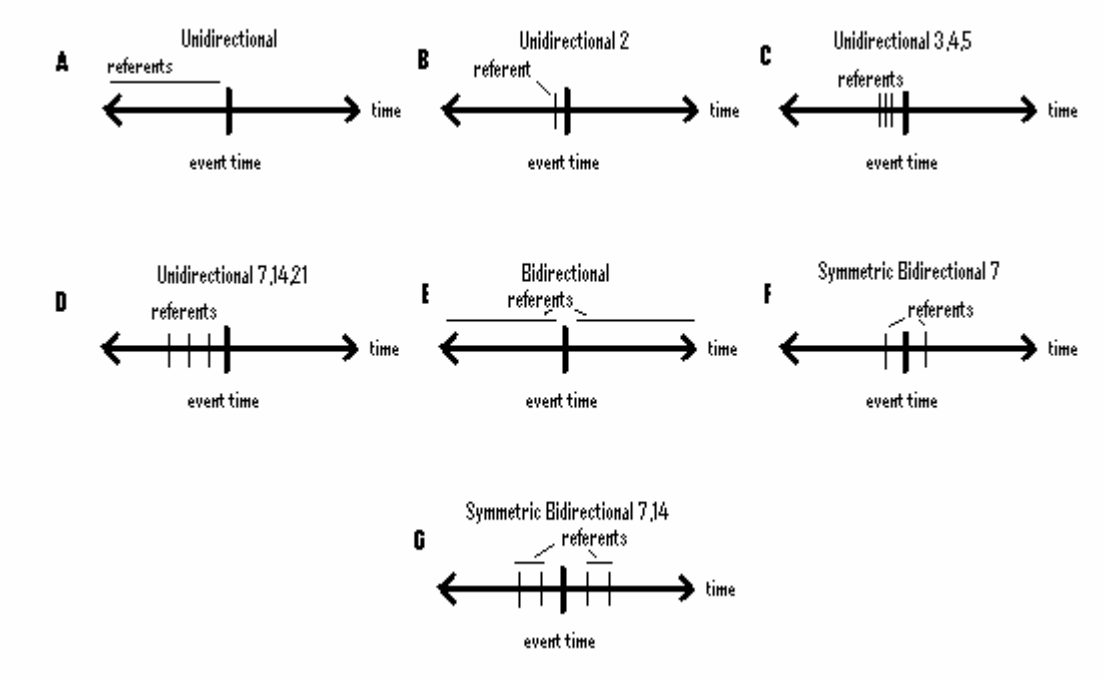

Figure 2 Large sample bias as a function of beta for three realizations of an exposure series with random noise, for eight different referent selection strategies: total history unidirectional, unidirectional 2, unidirectional 3,4,5, unidirectional 7,14,21, symmetric bidirectional 7, symmetric bidirectional 7,14, semi-symmetric bidirectional 7, and time stratified. For each referent strategy, we show the bias for β in (-0.01, 0.01).

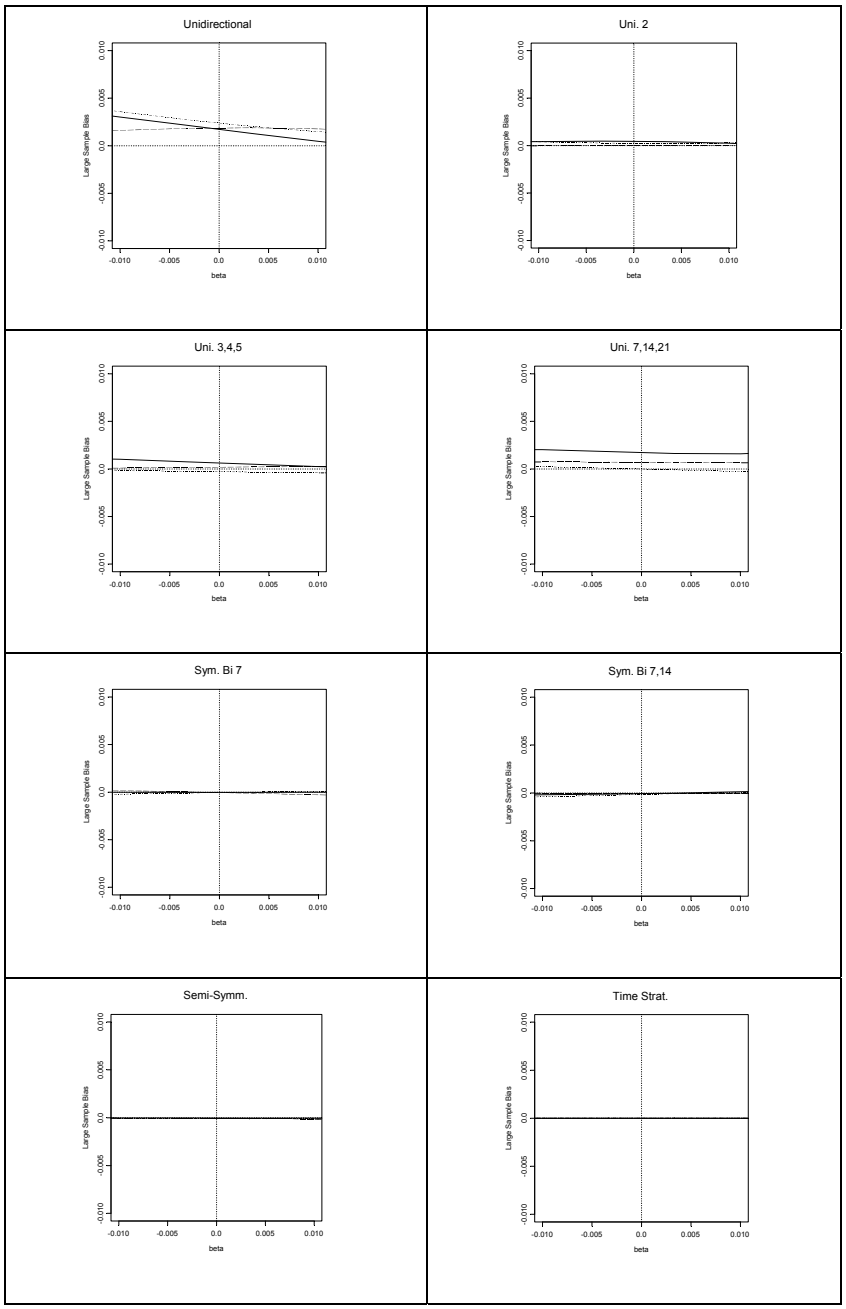

Figure 3 Large sample bias as a function of beta for three realizations of an exposure series with autocorrelation 0.6, for eight different referent selection strategies: total history unidirectional, unidirectional 2, unidirectional 3,4,5, unidirectional 7,14,21, symmetric bidirectional 7, symmetric bidirectional 7,14, semi-symmetric bidirectional 7, and time stratified. For each referent strategy, we show the bias for β in (-0.01, 0.01).

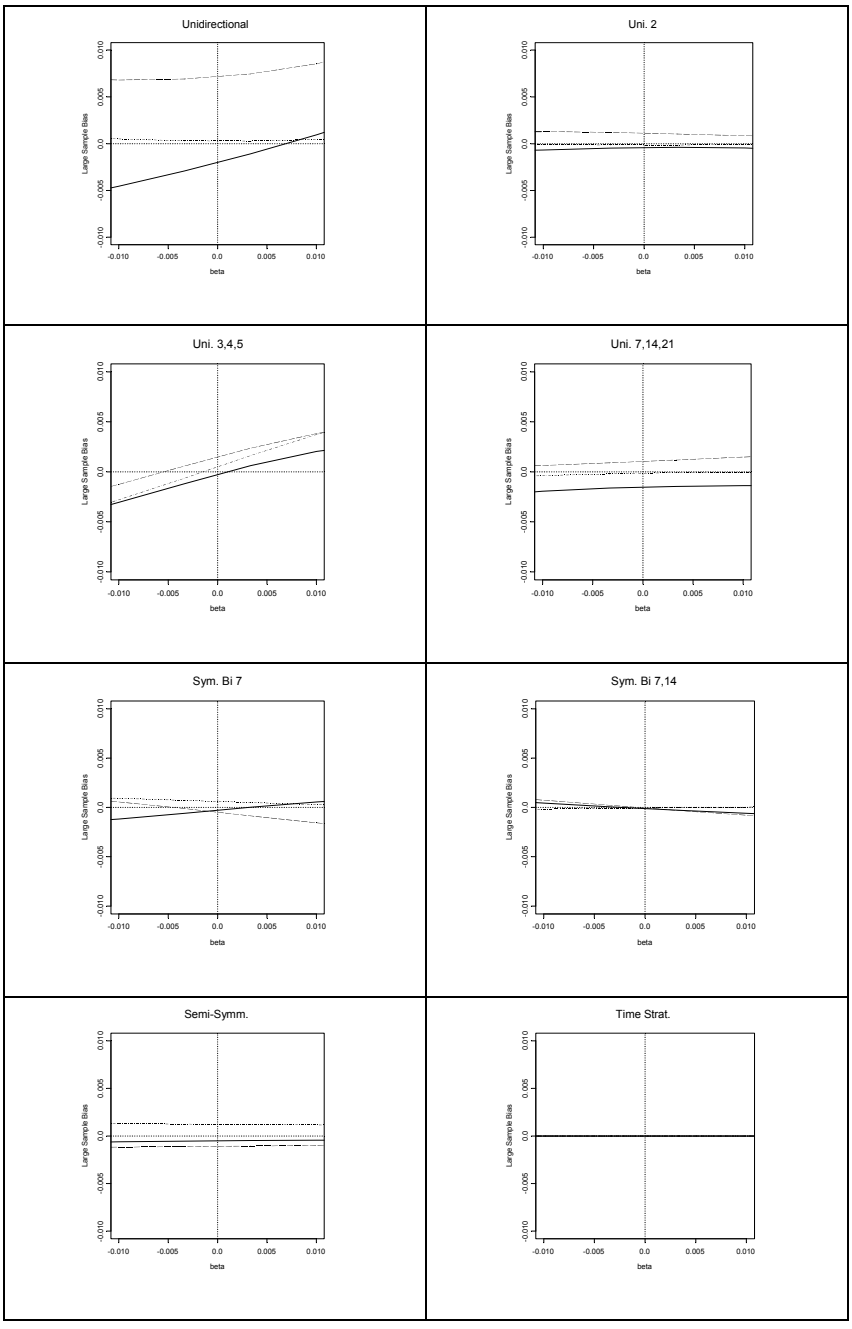

Figure 4 Large sample bias as a function of beta for three realizations of an exposure series with time varying mean and variance, for eight different referent selection strategies: total history unidirectional, unidirectional 2, unidirectional 3,4,5, unidirectional 7,14,21, symmetric bidirectional 7, symmetric bidirectional 7,14, semi-symmetric bidirectional 7, and time stratified. For each referent strategy, we show the bias for β in (-0.2, 0.2).

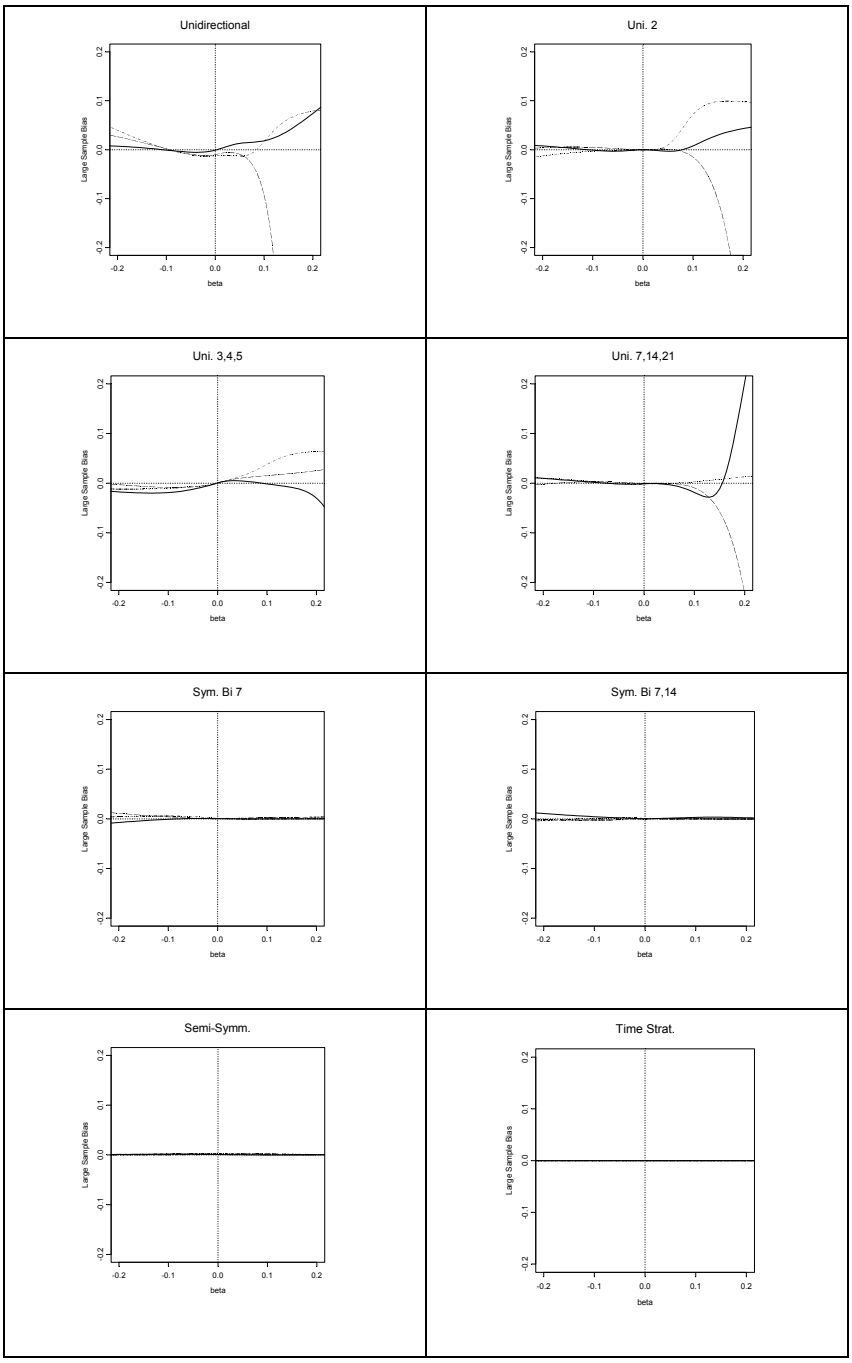

Figure 5 Large sample bias as a function of beta for three realizations of an exposure series with time varying mean and variance, for eight different referent selection strategies: total history unidirectional, unidirectional 2, unidirectional 3,4,5, unidirectional 7,14,21, symmetric bidirectional 7, symmetric bidirectional 7,14, semi-symmetric bidirectional 7, and time stratified. For each referent strategy, we show the bias for β in (-0.01, 0.01).