Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Anran Liu Date

Study designs for estimating association between time-varying exposures during pregnancy and preterm birth: a simulation study

By

Anran Liu

Master of Science of Public Health

Emory University

Rollins School of Public Health

Department of Biostatistics and Bioinformatics

Howard H. Chang
Committee Chair

Lyndsey Darrow
Committee Member

Study designs for estimating association between time-varying exposures during pregnancy and preterm birth: a simulation study

By

Anran Liu

B.S., Sun Yat-Sen University, 2013

MSPH, Emory University

Rollins School of Public Health

2015

Thesis Committee Chair: Howard H. Chang, PhD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Department of Biostatistics and Bioinformatics

2015

Abstract

Study designs for estimating association between time-varying exposures during pregnancy and preterm birth: a simulation study

By Anran Liu

Preterm birth, defined as birth occurring before 37 weeks of gestation, is the leading cause of perinatal morbidity and mortality, and long-term neurological disabilities. An increasing number of studies have investigated the association between environmental exposures during pregnancy and preterm birth. However, the results are inconsistent across studies and across exposure windows because ambient air pollution levels and temperature have strong seasonal patterns and there is no standard analytic method to study time-varying exposures. The purpose of our simulation study is to examine the performances of 4 commonly used study designs, including logistic regression, case-crossover design, time-series analyses, and discrete time survival model. We first simulated the outcome of gestational age using a discrete time survival model with true relative risk for the exposures of interest equal to 1.000, 1.002, 1.004, and 1.010 for the exposure of interest. Then we used the 4 methods to estimate the risk. We compared the root mean square error (RMSE), power, coverage of 95% confidence interval, bias, and average standard error from the first 3 designs to those from discrete time survival model. We found that logistic regression is as good as the discrete time survival model when examining time invariant exposure windows; but it would overestimate the risks when examining the time varying exposure windows. The longer the exposure window, larger bias is associated with logistic regression. Case-crossover design and time-series analyses were used to examine the 1-week lag exposure and we found that case-crossover design would introduce large bias, large average standard errors, large RMSEs, small power and small 95% coverage. We also found that time-series analyses with and without stratification give similar results to the discrete time survival model.

Study designs for estimating association between time-varying exposures during pregnancy and preterm birth: a simulation study

By

Anran Liu

B.S., Sun Yat-Sen University, 2013

MSPH, Emory University

Rollins School of Public Health

2015

Thesis Committee Chair: Howard H. Chang, PhD

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Department of Biostatistics and Bioinformatics

2015

Acknowledgements

I want to thank the faculty, advisors, and staff of the Department of Biostatistics and Bioinformatics at Rollins School of Public Health for the dynamic two years of learning that I have had. This thesis is only a sample of the vast knowledge that was attained and applied through my two years here at Rollins. I would especially like to thank Dr. Howard Chang for all of his advice and support to help me write this thesis. Also a special thanks to Dr. Lyndsey Darrow for taking the time to read my thesis.

Lastly I want to give a special thanks to my parents for their loving support and encouragement that motivated me to pursue a degree at the graduate level.

Contents

1 Introduction **1**

2 Methods **5**

2.1 Data Generation 5

2.2 Model for the 4 approaches 7

2.2.1 Logistic Regression 7

2.2.2 Discrete Time Survival Model 8

2.2.3 Case-Crossover Design 8

2.2.4 Time-series Analyses 9

2.3 Simulation Evaluation 10

3 Results **11**

4 Dicussions **14**

References **15**

Appendices **16**

Tables 16

Sample R code 20

1 Introduction

Preterm birth, which is defined as birth occurring before 37 weeks of gestation, is the leading cause of perinatal morbidity and mortality (Goldenberg et al., 2008). Among the survivors, infants born preterm are at risk of developing a variety of medical complications in the long term, such as neurodevelopmental disabilities (Saigal et al., 2008; Lorenz et al., 1998), decreased motor and cognitive functioning, increased behavioral disorders (Anderson et al., 2003), visual and hearing loss, and chronic lung disease (Swamy et al., 2008). Among those who do not have medical disabilities, the gestational age at birth is also associated with socioeconomic status. Specifically, infants born preterm can have lower educational attainment, income and receipt of Social Security benefits in adulthood when compared with those born at full term (Moster et al., 2008).

Preterm birth is considered to be a syndrome initiated by multiple mechanisms and researchers have been trying to identify the risk factors, such as maternal demographic characteristics, and biological and genetic markers, that are associated with preterm birth. More recently, an increasing number of studies have investigated the association between environmental exposures during pregnancy and preterm birth.

For example, a case-control survey nested within a birth cohort was conducted in Los Angeles. Logistic regression was used and they concluded that exposure to traffic-related pollutants during the first trimester is associated with preterm birth (Ritz et al., 2007). Another study in the Republic of Korea found a significant association between air pollution and preterm birth during the third trimester of pregnancy using logistic regression (Leem et al., 2006). A population-based cohort study was conducted in Vancouver, Canada and they found positive associations between traffic-related air pollution and preterm birth during the overall pregnancy period (Brauer et al., 2008). Temperature was also found

to have a short-term association with preterm birth independent of air pollutants in California using case-crossover design (Basu et al., 2010). Time-series approach was used in a retrospective cohort study conducted in Atlanta and they concluded that most of the ambient air pollutants they studied did not have late pregnancy effects on preterm birth (Darrow et al., 2009a). The results were inconsistent across studies. Some of them found excess risks of environmental exposures to preterm birth while others found no meaningful associations. Even in studies that found significant associations, the associations were inconsistent across exposure levels and windows (Bosetti et al., 2010). There are 4 areas that have been identified that contribute to the variation in the findings in the published studies, including confounding and effect modification, spatial and temporal exposure variations, vulnerable windows of exposure, and multiple pollutants (Woodruff et al., 2009). Thus, the epidemiologic evidence remains limited and inconsistent.

Another possible factor that contribute to the variation of these published articles is that investigators are using different analytic strategies in their studies. The most common analytic approach is logistic regression in which preterm birth and full term births are treated as binary outcomes. This approach is most appropriate to examine time invariant exposure windows such as first trimester, second trimester, and first 4-weeks since conception. For time variant exposures, logistic regression is no longer appropriate because exposure windows closer to the end of pregnancy are more challenging to assess due to the changing risk set of ongoing pregnancies across time. For example, If we use logistic regression to study the long-term air pollution effect (exposure window from conception to birth), bias in risk estimates may increase because preterm births and full-term births experience different lengths of exposure. For pregnancies starting from winter, preterm births are more likely to experience lower average exposure because full term births have longer exposure window ex-

tending in to the summer when ambient air pollution levels are normally higher. Similarly, for pregnancies starting from summer, preterm births are more likely to experience higher average exposure compared to full-term birth. Logistic regression is also inappropriate for short-term time variant exposure windows (e.g. 1 week or 4 weeks before births), because this exposure metric, a full-term pregnancy is no longer at risk of being preterm. Also, using logistic regression on short-term time variant exposures cannot fully use information from earlier weeks (Chang et al., 2013).

Case-crossover method is a statistical technique well suited to examine 1-week lag exposures (average of the exposure on the case day and the previous 6 days) with acute outcomes. It is a modification of the matched case-control study. In the case-crossover design, each preterm birth serves as its own control to adjust for unknown time-invariant confounders. Post-event control periods are used in this study design even though a pregnant woman is theoretically no longer at risk of giving birth. Bidirectional and unidirectional sampling of control days may introduce bias. But we cannot exclude the control period after the case event because it would typically introduce selection bias when there are within-window trends in exposure. However, any bias from using control periods after the risk period is small if the event is rare (Lumley et al., 2000). Another problem with this design is that it cannot adjust for time-variant risk factors if there are no available measures of the factors; thus, one of the key assumptions is that aside from the environmental exposure under study, no other risk factors for preterm birth should be time-variant, unless these factors that vary systematically within the time window are controlled for (Basu et al., 2010).

Time-series analysis is also widely used in studying the effect of environmental exposure. In time-series analyses, the outcome of interest is the daily number of preterm births. The offset is the daily number of births that are still at risk of being preterm. By observing

the population over time, we could remove the influence of known and unknown time-invariant risk factors that vary across individuals over short periods of time. However, a time-series analysis has limited power to detect long-term effects because considerable temporal variation in the exposure is removed when controlling for seasonality in preterm births.

Finally, another way is to consider the gestational age as a time-to-event (survival) data. In this approach, each pregnancy enters the risk set at 27th gestational week or as early as the 20th week of gestation if early preterm births are of interest. Then the pregnancy is followed until a birth occurs before 37th week (preterm), or reaches 37th week (full-term). This way, pregnancies are only compared with each other when they are at risk of being preterm (i.e. 27-36). Therefore, the long-term effects can be examined using a time varying cumulative average instead of an average over the entire pregnancy and the short-term effects can be examined using a time varying lagged average instead of a fixed period (Chang et al., 2013). Compared to logistic regression, case-crossover design and time-series analyses, discrete time survival model is an appropriate method to study the exposure effects on preterm births for 2 reasons, including the strong seasonality of ambient pollution and temperature, and changes in the risk set of pregnancies at risk for preterm birth. The limitation of the discrete time survival model is it cannot control potential confounding by individual-level risk factors like in case-crossover design and time-series analyses.

We conducted a simulation study to compare the root mean square error (RMSE), bias, coverage, average standard error, and power associated with the first three methods to that from the discrete time survival approach because the outcome of gestational age was simulated using a discrete time survival model. We studied 3 environmental exposures including fine particulate matter ($PM_{2.5}$), temperature, and ozone because each of these 3

exposures have their unique seasonal patterns and temporal correlation. We used logistic regression and discrete time survival model to study both the time-invariant exposures (first trimester, second trimester, and first 4 weeks since conception) and time-variant exposures (third trimester, 1-week lag, 4-weeks lag, and the cumulative exposure from conception to births). We used case-crossover design and time-series analyses to study the 1-week lag exposure because these 2 methods are specifically designed for short term exposures. For each exposure, we did the simulation under 4 RRs (1.000, 1.002, 1.005, 1.010) for preterm birth per one unit increase in the different exposure metrics. Because this is a simulation study, we only include exposure and gestational week in our models for simplicity, and no other confounders are considered.

2 Methods

2.1 Data Generation

Our models are defined through $p_{it} \in [0, 1]$, which is the hazard rate of birth during at-risk window week t for pregnancy i. We model the hazard rate using logistic regression.

$$p_{it} = \frac{\frac{h_0(t)}{1-h_0 t} e^{\beta_1 z_{it}}}{1 + \frac{h_0(t)}{1-h_0 t} e^{\beta_1 z_{it}}} \tag{1}$$

where $h_0(t)$ is the baseline hazard rate of birth for each gestational week t. It is estimated using Atlanta birth record data. Specifically, $h_0(t)$ is the probability of birth for pregnancy i at week t given that pregnancy i reaches week t when exposure level is 0. The exposures used in our simulation come from the daily air pollutant level and temperature in Atlanta. We used the records starting from January 1st 2001 and for each simulation we generated 1000 days and 50 conceptions per day.

First, for each birth i, we calculated weekly exposure x_{it} for gestational week t, ($t \in (1, 44)$. Week 44 is the latest gestational week that a birth can occur based on the Atlanta birth record data we used. x_{it} is the average exposure concentrations during the 7 days leading up to the date that gestational week t was completed. For example, for the 50 conceptions on 2001-01-01, we use the exposure data starting from 2001-01-01 to estimate the weekly exposure (exposure for gestational week 1 is the average of the exposures from 2001-01-01 to 2001-01-07 and exposure for gestational week 2 is the average of the exposures from 2001-01-08 to 2001-01-14); and for the 50 conceptions on 2001-01-02, we use the exposure data starting from 2001-01-02 to estimate the weekly exposure (exposure for gestational week 1 is the average of the exposures from 2001-01-02 to 2001-01-08 and exposure for gestational week 2 is the average of the exposures from 2001-01-09 to 2001-01-15).

We investigated average exposures defined over 7 exposure windows (z_{it} in equation (1)). We considered 3 time invariant exposure windows. Each one of these effects is identical for all ongoing pregnancies during the entire at-risk window (week 27-36). a) Trimester 1: average of the exposures from gestational week 1 to week 13. b) Trimester 2: average of the exposures from gestational week 14 to week 27. c) 4 weeks since conception: average of the exposures from gestational week 1 to week 4. Given that a pregnancy completed at gestational week t, we also considered 4 time-varying exposures. a) trimester 3: average of the exposures from gestational week 27 to week t. b) cumulative: average of exposures from conception to gestational week t. c) 4-week lag: average of the exposures from gestational week $t-3$ to week t. d) 1-week lag: exposure at gestational week t.

We also used different RRs for each exposure and exposure metric. β_1= 0, 0.002, 0.005, 0.010 correspond to an approximate 1.000, 1.002, 1.005, and 1.010 increase in hazard ratio per one unit increase in weekly exposures. Given β_1 and the weekly exposure during the at-

risk window, the probability of having a preterm birth for pregnancy i at gestational week t was calculated using formula (1). Then, for each gestational week t during the at-risk window, whether preterm births occurred (0 for preterm birth not occurred, 1 for preterm birth occurred) for pregnancy i are generated using a binomial distribution.

Generally, for each exposure metric, we have three matrices, one is the exposure matrix that records the exposure for each gestational week t and each pregnancy i; one is the probability matrix that the probability of preterm birth for each gestational week t and each pregnancy i; and the other one is the birth matrix filled with 0s, 1s, and NAs. For pregnancy i, once a preterm birth occurred at gestational week t, NAs were assigned to week $t+1$ to week 36. Each matrix has 50000 rows and 10 columns. In order to apply different models to estimate risk, different modifications are needed for the simulated dataset that we describe below.

2.2 Model for the 4 approaches

2.2.1 Logistic Regression

In logistic regression, we needed to add up each row of the birth matrix to identify whether preterm birth occurred for pregnancy i. The time invariant exposures have the same effect for all ongoing pregnancies after week 27. For the time variant exposures, we need to expand the matrices to gestational week 44. The hazard rate during the at-risk window is still the same and the hazard rate after week 36 equals to the baseline hazard rate since they are no longer at risk of preterm birth. The exposure used for pregnancy i is the one when the birth occurred (either preterm birth or not). The model for logistic regression is,

$$logit(p_i) = \beta_0 + \beta_1 x_i \quad i = 1, \ldots, 50000$$

where β_1 represents the coefficients for exposure vector x_i.

2.2.2 Discrete Time Survival Model

In this design, we viewed gestational age as time-to-event data and each pregnancy enters the risk set at the 27th week of gestation and is followed until either a birth occurs before the 37th week (preterm) or it reaches the 37th week and a full-term birth is expected. Full-term births are censored at week 36, and no censoring occurs within the at-risk window. We aligned the exposure matrix and birth matrix such that pregnancies are compared with each other only during a window at risk of being preterm. We fitted the discrete-time survival model with a logistic link. Specifically, let y_{it} denote the indicator of whether a birth occurs during gestational week t for pregnancy i. We modeled

$$logit(p_{it}) = \beta_0 + \alpha_t + \beta_1 x_{it} \quad t = 1, \ldots, 10 \quad i = 1, \ldots, 500000$$

where p_{it} represents $Pr(y_{it} = 1|$ no birth before week t), α_t represents the week-specific intercepts and β_1 represents the coefficients for exposure vector x_{it}.

2.2.3 Case-Crossover Design

Case-crossover design is only used for the 1-week lag exposure. We consider a time-stratified design where the control days are defined as the same day of week and the same month of the case date. To simplify the simulation, we generated our own date. It has 12 months a year and 28 days a month. In this approach, each preterm birth serves as its own control. Control periods are limited to the rest 3 weeks in the preterm birth's month. If a case of preterm birth occurred in the middle of the month, control periods occurred both before and after the case period. If a case of preterm birth occurred at the end of the month, all control periods occurred prior to the case. If a case of preterm delivery occurred at the

beginning of the month, all control periods would be selected after the case. For example, if a preterm birth occurred at 2001-03-15, the control periods are 2001-03-01, 2001-03-08, and 2001-03-22. For each preterm birth, the case and 3 controls form a group. Different simulated data can have different number of groups. Instead of using dummy variables in a logistic regression, conditional logistic regression is used for this design. Conditional logistic regression still has a linear term for exposures and the log(odds) of preterm delivery (yes/no) served as the outcome measure. It estimates a logistic regression model by maximizing the conditional likelihood. We used the *clogit* function in package *survival* in *R*.

2.2.4 Time-series Analyses

Poisson regression is used in the time-series analyses. Our outcome of interest is the daily number of preterm births. For this design, we generated 1063 days and 50 conceptions per day. For the first 7 days, at each day, only 50 pregnancies reached the at-risk window. From day 8 to day 14, at each day, another 50 pregnancies reached the at-risk window and the original 50 pregnancies moved to gestational week 28. Thus, only until day 64, we can have a total number of 500 pregnancies per day, which means 50 pregnancies for each gestational week per day. Similarly, starting from day 1064, there are no more new pregnancies reaches the at-risk window; thus, we no longer have pregnancies in gestational week 27. Thus, we need to discard the first 63 days and last 63 days, such that our simulated data have 1000 days and 50 pregnancies per gestational week per day.

The dependent variables are the daily number of preterm births. The offset is the daily number of pregnancies that are still at-risk of being preterm (between gestational week 27-36 and haven't been born). We fitted two models for time-series analyses: one stratified by gestational week and one doesn't. The two models are given as follows,

$$log(\lambda_{tk}) = \beta_0 + \alpha_t + \beta_1 x_{tk} + \text{offset}(log(N_{tk}))$$

$$log(\lambda_k) = \beta_0 + \beta_1 x_k + \text{offset}(log(N_k))$$

where y_{tk} = number of preterm births on day k at gestational week t, $y_{tk} \sim Poisson(\lambda_{tk})$, N_{tk} = number of ongoing births on day k at gestational week t, x_{tk} = 1-week lag exposure on day k at gestational week t, and α_t = gestational week t. And y_k = number of preterm births on day k, $y_k \sim Poisson(\lambda_k)$, N_k = number of ongoing births on day k, x_k = 1-week lag exposure on day k.

For logistic regression, discrete time survival analyses, and case-crossover approach, we simulated 1000 times for each exposure metric and each RRs. For the time-series analyses, we did the simulations for 200 times due to the long time that is needed to generate each simulated data set.

2.3 Simulation Evaluation

We examined at the following statistics for each designs:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\beta}_i - \beta_{true})^2}{N}}$$

$$95\% \text{ Coverage} = Pr(\beta_{true} \in (\hat{\beta}_i \pm 1.96\hat{se}_i))$$

$$\text{Average SE} = \frac{\sum_{i=1}^{N}\hat{s}_i}{N}$$

$$\text{Bias} = \frac{\sum_{i=1}^{N}(\hat{\beta}_i - \beta_{true})}{N}$$

$$\text{Power} = Pr(|\frac{\hat{\beta}_i}{\hat{s}_i}| > 1.96)$$

where N is the total number of simulations. We compare the statistics from logistic regression, case-crossover design, and time-series analyses to those from discrete time survival model. For bias, we presented the raw bias times 1000. For RMSE and average standard error, we use the ones from the 3 testing designs divided by the ones got from the discrete

time survival model. For 95% coverage, and power, we use the ones from the 3 other designs to subtract the ones got from the discrete time survival model. If the difference for RMSE and average standard error is close to 1 and the difference for 95% coverage, and power is close to 0, it means that the design perform similarly to the discrete survival model. All of the simulations were done using R and the code is attached in the appendix.

3 Results

Table 1. shows the results of the differences between logistic regression and discrete time survival model of all the exposure metrics except the 1-week lag exposure. For RMSE, coverage, average standard error, and power we presented the statistics relative to the survival model; for bias, we presented the raw bias times 1000.

First, we look at the time invariant exposure windows. For trimester 1, the RMSEs and average standard errors from logistic regression are almost the same as the ones from the survival model, only slightly larger; the 95% coverages and powers are also similar to the survival result; the biases are small but for temperature and ozone, they have an increasing trend when RR increases. For trimester 2 and first 4 weeks, the RMSEs for $PM_{2.5}$ of logistic regression are similar to the ones from survival model; the RMSEs for temperature and ozone slightly increases when relative risk increases; the 95% coverage, averages standard error, and power from the logistic regression are nearly the same as the ones from the survival model; and the risk estimates from the logistic regression are very close to the true risks.

In conclusion, for the time invariant exposure windows, logistic regression can perform almost as good as the survival model.

Then, we look at the time-varying exposure windows. For trimester 3, we can see that

the relative RMSEs and average standard errors are quite large even when the relative risk equals to 1 and they increase when the relative risk increases for each exposure. We can also see that compared to the survival model, the 95% coverage of logistic regression is smaller and it decreases when the relative risk increases. The raw biases are all positive which means logistic regression would over estimate the risks and the bias also increases with increasing relative risk and it is much larger compared to the time invariant exposures. For temperature and ozone, logistic regression loses the most power when RR=1.002, but loses almost no power for the other RRs likely due to the positive bias. For $PM_{2.5}$, when RR=1.005, logistic regression loses 41.3% power than the survival model.

For the 4-week lag exposure window, the relative RMSE of temperature and ozone when RR=1 is quite large and it increases when RR increases; but the RMSE of $PM_{2.5}$ is very similar to that from the survival model. For temperature and ozone, the 95% coverage of logistic regression is smaller than the survival model; but for $PM_{2.5}$, the 95% coverage is almost the same. The average standard error of logistic regression is almost the same as the one from the survival model only slightly larger. For the temperature, the bias has an increasing trend when RR increases. The biases have no increasing trend for ozone and $PM_{2.5}$, but they are larger than the biases from the time-invariant exposure. For temperature and ozone when RR=1.002, logistic regression loses the most power compared to the survival model. For $PM_{2.5}$, the power of logistic regression is almost the same as the power from survival model.

For the cumulative exposure metric, the statistics have the same pattern as the statistics in trimester 3. The relative RMSEs and raw biases are even larger and logistic regression loses up to 90% of the power than the survival model. In conclusion, logistic regression is not appropriate for time-variant exposures because it introduces bias and loses a lot of

power. And the longer the time-varying exposure window is, the poorer the performance of logistic regression.

Table 2. shows the results of the 1-week lag exposure. First we look at the biases. They are presented as raw bias times 1000. From the table, we can see that case-crossover design has the largest biases and all of them are negative, which means case-crossover design tend to underestimate the risks. For each exposure, the biases increase when the relative risk increases in the case-crossover design. The biases of $PM_{2.5}$ in the logistic regression are also all negative and increase when the RR increases. Time-series analyses stratified by week or not have similar results. The biases from these two time-series models are small but there is still a slightly increasing pattern when the RR increases.

The average standard errors and RMSEs in the case-crossover design are much larger than those from the survival model even when RR=1. And the average standard errors and RMSEs from logistic regression and the 2 time-series analyses are very similar to those from the survival model. And the relative 95% coverage in the logistic regression and case-crossover design are almost all negative, which means that the 95% coverage of logistic regression and case-crossover design are smaller than the one from survival model. For time-series analyses, the 2 models have similar results which are also close to those from the survival model. The power in the case-crossover design are almost all negative, which means that case-crossover design loses a lot of power compared to the survival model. The powers for logistic regression and the 2 time-series analyses are almost the same as those from a survival model.

In conclusion, for the 1-week lag exposure, case-crossover design has the poorest performance among all the study designs. It will lead to large bias, large average standard errors, large RMSEs, small power, and small 95% coverage. Logistic regression is not as good as

the discrete time survival model when testing the time-variant exposures. The 2 models of time-series analyses have similar results, which means whether stratify by gestational week doesn't influence the risk estimation very much. And the time-series analyses is almost as good as the survival model when examining the 1-week lag exposure.

4 Dicussions

There are several limitations of our simulation study. One limitation is that in order to simplify the simulation, we generated equal number of conceptions for each day. However, there are seasonal patterns of birth, such as an overall peak in August-September and a bottom during April–May. It may influence the result of time-series analyses because the seasonal pattern of birth would change the population at risk in a time-series of pregnancy outcomes. That could create confounding when examining a seasonally varying exposure because the risk of preterm birth could differ across seasons due to changing distributions of risk factors in the pregnancy risk set (Darrow et al., 2009b). Thus, for further simulations, the number of conceptions each day should be simulated according to the seasonal pattern of birth.

The second limitation is that how we generated the simulated data for time-series analyses is not very efficient. We only run 200 times for the time-series analyses but 1000 times for the other study designs. We need to improve the efficiency of our R code so that we can run the time-series analyses more times which can make the final results more reliable. The third limitation is that for case-crossover design, we only used a time-stratified approach to sample the control days. We can also examine the unidirectional or bidirectional sampling of control days and see how does sampling methods affect the performance of the case-crossover design.

References

Basu, R., Malig, B. and Ostro, B. (2010). High ambient temperature and the risk of preterm delivery. *American Journal of Epidemiology* **172**, 1108–1117.

Bosetti, C., Nieuwenhuijsen, M. J., Gallus, S., and Cipriani, S., La Vecchia, C. and Parazzini, F. (2010). Ambient particulate matter and preterm birth or birth weight: A review of the literature. *Archives of Toxicology* **84**, 447–460.

Brauer, M., Lencar, C., Tamburic, L., Koehoorn, M., Demers, P. and Karr, C. (2008). A cohort study of traffic-related air pollution impacts on birth outcomes. *Environmental Health Perspectives* **116**, 690–686.

Chang, H. H., Reich, B. J. and Miranda, M. L. (2013). A spatial time-to-event approach for estimating associations between air pollution and preterm birth. *Journal of the Royal Statistical Society Series C 62*, 167–179.

Darrow, L. A., Klein, M., Flanders, W. D., Waller, L. A., Correa, A., Marcus, M., Mulholland, J. A., Russell, A. G. and Tolbert, P. E. (2009a). Ambient air pollution and preterm birth: a time-series analysis. *Epidemiology* **20**, 689–698.

Darrow, L. A., Strickland, M. J., Klein, M., Waller, L. A., Flanders, W. D., Correa, A. and Tolbert, P. E. (2009b). Seasonality of birth and implications for temporal studies of preterm birth. *Epidemiology* **20**, 699–706.

Goldenberg, R. L., Culhane, J. F., Iams, J. D. and Romero, R. (2008). Epidemiology and causes of preterm birth. *Lancet* **371**, 75–84.

Leem, J. H., Kaplan, B. M., Shim, Y. K., P, H. R., Gotway, C. A., Bullard, S. M., Rogers, J. F., Smith, M. M. and Tylenda, C. A. (2006). Exposures to air pollutants during pregnancy and preterm delivery. *Environmental Health Perspectives* **114**, 905–910.

Lorenz, J. M., Wooliever, D. E., Jetton, J. R. and Paneth, N. (1998). A quantitative review of mortality and developmental disability in extremely premature newborns. *Archives of Pediatrics & Adolescent Medicine* **152**, 425–435.

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

Ritz, B., Wilhelm, M., Hoggatt, K. J. and Ghosh, J. K. C. (2007). Ambient air pollution and preterm birth in the environment and pregnancy outcomes study at the University of California, Los Angeles. *American Journal of Epidemiology* **166**, 1045–1052.

Saigal, S. and Doyle, L. W. (2008). An overview of mortality and sequelae of preterm birth from infancy to adulthood. *Lancet* **371**, 261–269.

Woodruff, T. J., Parker, J. D., Darrow, L. A., Slama, R., Bell, M. L., Choi, H., Glinianaia, S., Hoggatt, K. J., Karr, C. J., Lobdell, D. T. and Wilhelm, M. (2009). Methodological issues in studies of air pollution and reproductive health. *Environmental Research* **109**, 311–320.

Appendices

Tables

Table 1: Discrete Time Survival Model vs. Logistic Regression

Trimester 1						
Exposure	**Relative risk**	**RMSE**	**95% Coverage**	**Average SE**	**Raw Bias × 1000**	**Power**
Temperature	1.000	1.045	0.005	1.042	-0.003	-0.005
	1.002	1.019	0.005	1.048	0.102	0.016
	1.005	1.071	-0.016	1.060	0.270	0.003
	1.010	1.396	-0.080	1.087	0.870	0.000
Ozone	1.000	1.008	0.014	1.042	0.015	-0.014
	1.002	1.065	-0.012	1.046	0.113	-0.001
	1.005	1.086	0.001	1.052	0.261	0.002
	1.010	1.195	-0.031	1.065	0.639	0.000
$PM_{2.5}$	1.000	1.047	-0.004	1.042	-0.320	0.004
	1.002	1.026	0.007	1.044	0.151	-0.012
	1.005	1.040	0.011	1.045	0.093	-0.005
	1.010	1.071	-0.006	1.050	0.388	0.001
Trimester 2						
Exposure	**Relative risk**	**RMSE**	**95% Coverage**	**Average SE**	**Raw Bias × 1000**	**Power**
Temperature	1.000	1.094	-0.015	1.042	0.054	0.015
	1.002	1.008	0.003	1.048	0.052	-0.058
	1.005	1.101	-0.003	1.060	0.302	-0.003
	1.010	1.397	-0.068	1.086	0.823	0.000
Ozone	1.000	0.997	0.016	1.042	-0.053	-0.016
	1.002	1.006	0.003	1.046	0.035	0.004
	1.005	1.151	-0.026	1.052	0.249	0.000
	1.010	1.248	-0.054	1.065	0.609	0.000
$PM_{2.5}$	1.000	1.073	-0.014	1.042	0.147	0.014
	1.002	1.014	0.007	1.043	0.173	-0.004
	1.005	1.056	-0.004	1.046	0.100	-0.009
	1.010	1.034	-0.010	1.050	0.553	-0.026
First 4 weeks						
Exposure	**Relative risk**	**RMSE**	**95% Coverage**	**Average SE**	**Raw Bias × 1000**	**Power**
Temperature	1.000	0.963	0.039	1.042	-0.014	-0.039
	1.002	1.066	-0.003	1.048	0.107	-0.014
	1.005	1.099	-0.012	1.059	0.248	0.001
	1.010	1.472	-0.096	1.086	0.834	0.000
Ozone	1.000	1.074	-0.006	1.042	-0.014	0.006
	1.002	1.104	-0.016	1.046	0.106	-0.020
	1.005	1.162	-0.023	1.053	0.253	0.000
	1.010	1.371	-0.082	1.065	0.668	0.000
$PM_{2.5}$	1.000	1.049	-0.003	1.042	-0.203	0.003
	1.002	1.029	0.011	1.044	-0.052	0.001
	1.005	1.031	-0.010	1.046	0.223	-0.007
	1.010	1.011	0.030	1.051	0.053	-0.036

Trimester 3

Exposure	Relative risk	RMSE	95% Coverage	Average SE	Raw Bias × 1000	Power
Temperature	1.000	1.298	-0.036	1.122	0.680	0.036
	1.002	1.570	-0.096	1.127	1.125	0.232
	1.005	2.020	-0.255	1.137	1.617	0.002
	1.010	3.434	-0.732	1.160	2.789	0.000
Ozone	1.000	1.256	-0.016	1.132	0.400	0.016
	1.002	1.445	-0.082	1.135	0.831	0.172
	1.005	2.071	-0.264	1.141	1.484	0.000
	1.010	3.160	-0.677	1.153	2.625	0.000
$PM_{2.5}$	1.000	1.883	-0.144	1.208	6.473	0.144
	1.002	1.955	-0.212	1.209	7.504	0.269
	1.005	2.187	-0.251	1.212	8.407	0.413
	1.010	2.458	-0.353	1.216	10.353	0.365

4-week lag

Exposure	Relative risk	RMSE	95% Coverage	Average SE	Raw Bias × 1000	Power
Temperature	1.000	1.130	-0.015	1.048	0.442	0.015
	1.002	1.132	-0.016	1.055	0.525	0.148
	1.005	1.283	-0.054	1.066	0.705	0.000
	1.010	1.651	-0.144	1.093	1.011	0.000
Ozone	1.000	1.204	-0.041	1.048	0.612	0.041
	1.002	1.219	-0.035	1.052	0.647	0.224
	1.005	1.335	-0.070	1.059	0.741	0.000
	1.010	1.655	-0.145	1.073	0.976	0.000
$PM_{2.5}$	1.000	0.990	0.017	1.042	0.815	-0.017
	1.002	1.006	0.006	1.044	0.568	-0.007
	1.005	0.993	0.007	1.047	0.793	0.022
	1.010	1.055	0.006	1.051	0.643	-0.001

Cumulative

Exposure	Relative risk	RMSE	95% Coverage	Average SE	Raw Bias × 1000	Power
Temperature	1.000	4.688	-0.796	1.368	11.624	0.796
	1.002	5.557	-0.892	1.369	12.928	0.848
	1.005	6.965	-0.943	1.373	14.832	0.407
	1.010	9.488	-0.966	1.383	17.311	0.000
Ozone	1.000	4.319	-0.685	1.372	8.757	0.685
	1.002	5.272	-0.846	1.374	10.519	0.825
	1.005	6.155	-0.906	1.376	12.337	0.340
	1.010	8.274	-0.944	1.381	15.615	0.001
$PM_{2.5}$	1.000	5.555	-0.889	1.334	66.916	0.889
	1.002	5.367	-0.898	1.335	67.857	0.900
	1.005	5.640	-0.908	1.336	69.369	0.902
	1.010	6.196	-0.926	1.338	72.224	0.856

Table 2: 1-week Lag Exposures: Logistic Regression, Case-Crossover Design, and Time-series Analyses vs. Discrete Time Survival Model

Raw Bias × 1000

Exposure	Relative risk	Logistic Regression	Case-crossover Design	Time-series w/o factor week	Time-series with factor week
Temperature	1.000	-0.135	-0.489	0.028	0.056
	1.002	-0.137	-0.465	-0.269	-0.200
	1.005	-0.004	-0.818	-0.184	-0.169
	1.010	0.232	-1.154	-0.653	-0.580
Ozone	1.000	0.354	-1.174	0.147	-0.121
	1.002	0.383	-1.204	-0.038	-0.067
	1.005	0.429	-1.317	-0.272	-0.290
	1.010	0.533	-1.534	-0.554	-0.431
$PM_{2.5}$	1.000	-0.976	-2.554	-0.098	0.216
	1.002	-1.039	-2.710	0.292	-0.175
	1.005	-1.014	-2.852	-0.106	-0.839
	1.010	-1.190	-3.179	-0.637	-0.665

Average Standard Error

Exposure	Relative risk	Logistic Regression	Case-crossover Design	Time-series w/o factor week	Time-series with factor week
Temperature	1.000	1.041	2.969	0.987	0.988
	1.002	1.047	2.966	0.985	0.985
	1.005	1.060	2.960	0.983	0.983
	1.010	1.086	2.937	0.975	0.976
Ozone	1.000	1.050	1.966	0.992	0.993
	1.002	1.054	1.957	0.992	0.990
	1.005	1.062	1.941	0.9892	0.990
	1.010	1.078	1.918	0.985	0.984
$PM_{2.5}$	1.000	1.045	1.322	0.993	0.990
	1.002	1.046	1.324	0.989	0.990
	1.005	1.049	1.327	0.990	0.992
	1.010	1.055	1.335	0.990	0.989

RMSE

Exposure	Relative risk	Logistic Regression	Case-crossover Design	Time-series w/o factor week	Time-series with factor week
Temperature	1.000	1.059	3.075	0.917	0.938
	1.002	0.993	2.847	0.854	1.005
	1.005	1.044	3.050	0.979	0.912
	1.010	1.137	3.323	1.264	1.125
Ozone	1.000	1.148	2.382	1.057	0.918
	1.002	1.148	2.481	0.968	0.958
	1.005	1.170	2.598	0.946	0.970
	1.010	1.321	2.912	1.360	1.191
$PM_{2.5}$	1.000	1.048	1.562	0.984	1.061
	1.002	1.174	1.631	1.040	0.987
	1.005	1.055	1.581	0.976	0.879
	1.010	1.121	1.734	0.930	0.985

95% Coverage

Exposure	Relative risk	Logistic Regression	Case-crossover Design	Time-series w/o factor week	Time-series with factor week
Temperature	1.000	-0.011	-0.009	-5.00E-03	0.015
	1.002	0.011	0.009	3.60E-02	0.006
	1.005	-0.009	-0.017	-1.00E-03	-0.001
	1.010	-0.018	-0.04	-1.03E-01	-0.023
Ozone	1.000	-0.012	-0.05	-2.90E-02	0.011
	1.002	-0.016	-0.079	3.20E-02	0.002
	1.005	-0.026	-0.093	2.10E-02	0.001
	1.010	-0.053	-0.146	-1.12E-01	-0.072
$PM_{2.5}$	1.000	-0.011	-0.055	-3.00E-02	-0.02
	1.002	-0.026	-0.041	-1.90E-02	-0.009
	1.005	0.001	-0.054	-3.00E-03	0.017
	1.010	-0.011	-0.084	1.80E-02	-0.022

Power

Exposure	Relative risk	Logistic Regression	Case-crossover Design	Time-series w/o factor week	Time-series with factor week
Temperature	1.000	0.011	0.009	0.005	-0.015
	1.002	-0.091	-0.427	-0.095	-0.025
	1.005	0.001	-0.68	0.001	0.001
	1.010	0	-0.041	0	0
Ozone	1.000	0.012	0.05	0.029	-0.011
	1.002	0.09	-0.636	0.001	0.001
	1.005	0	-0.299	0	0
	1.010	0	0	0	0
$PM_{2.5}$	1.000	0.011	0.055	0.03	0.02
	1.002	-0.022	-0.043	0.027	-0.043
	1.005	-0.139	-0.296	0.034	-0.106
	1.010	-0.088	-0.512	-0.006	-0.026

Sample R code

```
### Get Simulation Parameters
load ("GW.RData") ## Gestational age by birth date in ATL
gestAge = subset (gestAge, gestweeks >= 27)

### Estimate baseline hazards from week 27 to 44
HZ = rep(NA, 18)
#hazard=#(X=t)/#(X>=t)
#the probability of birth for pregnancy i at week t given that pregnancy i reaches t
for(i in 1:18){
  HZ[i] = sum(gestAge$gestweeks == i+26)/sum(gestAge$gestweeks >= i+26)
}
rm (gestAge)
## Daily air pollutant level in ATL
aq = read.csv ("atl_aq_met.csv")
aq$DATE = as.Date (aq$DATE, "%m/%d/%Y")
aq = subset (aq, DATE >= as.Date ("2001-01-01"))
### Simulate Exposure Data
aq = aq[, c("DATE","PM25_CMS", "O3_M8_CMS", "TEMPMX")]
ozone = aq$O3_M8_CMS ##Ozone
pm=aq$PM25_CMS ##PM
temp=aq$TEMPMX ##Temperature

poll = ozone
#poll = pm
#poll = temp

dat.sim = data.frame (Day = rep(1:N.day, each = N.con))
## Get daily pollution concentration from week 1 to week 36
X.sim = matrix (NA, nrow = N.con*N.day, ncol = 36)
for (j in 1:N.day){
  poll.j = poll[j:(j+36*7-1)]
  #7*36,calculate the average pollution of each week
  poll.j = apply (matrix (poll.j, ncol = 36), 2, mean)
  X.sim[ dat.sim$Day == j, ] = rep(poll.j, each = N.con)
}
## Calculate exposure at each gestational week
T1 = T2 = T3 = Lag1 = Lag4 = First4 = Total = matrix (NA, nrow=N.con*N.day, ncol =
    10)

#exposure from week 1 to week 13 influence week 27 to 36 equally
for (g in 1:10){
  ## Time-invariant exposure
  T1[,g] = rowMeans (X.sim[,1:13])
  T2[,g] = rowMeans (X.sim[,14:26])
  First4[,g] = rowMeans (X.sim[,1:4])
  ## Time-variant exposure
  T3[,g] = rowMeans (as.matrix(X.sim[,27:(26+g)]))
  Lag4[,g] = rowMeans (X.sim[,(23+g):(26+g)])
  Lag1[,g] = X.sim[,26+g]
  Total[,g] = rowMeans(X.sim[,1:(26+g)])
}
P_T1 = P_T2 = P_T3 = P_Lag1 = P_Lag4 = P_First4 = P_Total =
  matrix (NA, nrow = nrow (dat.sim), ncol = 10) ## Probability of birth
###################T1####################
for (g in 1:10){
  tmp = HZ[g]*exp(T1[,g]*beta)/(1-HZ[g])
  P_T1[,g] = tmp / (1+tmp)
}
for (g in 1:10){
  tmp = HZ[g]*exp(T2[,g]*beta)/(1-HZ[g])
  P_T2[,g] = tmp / (1+tmp)
}
for (g in 1:10){
  tmp = HZ[g]*exp(T3[,g]*beta)/(1-HZ[g])
  P_T3[,g] = tmp / (1+tmp)
}
```

```
for (g in 1:10){
  tmp = HZ[g]*exp(Lag1[,g]*beta)/(1-HZ[g])
  P_Lag1[,g] = tmp / (1+tmp)
}
for (g in 1:10){
  tmp = HZ[g]*exp(Lag4[,g]*beta)/(1-HZ[g])
  P_Lag4[,g] = tmp / (1+tmp)
}
for (g in 1:10){
  tmp = HZ[g]*exp(First4[,g]*beta)/(1-HZ[g])
  P_First4[,g] = tmp / (1+tmp)
}
for (g in 1:10){
  tmp = HZ[g]*exp(Total[,g]*beta)/(1-HZ[g])
  P_Total[,g] = tmp / (1+tmp)
}

Y = matrix (NA, nrow = N.con*N.day, ncol = 10)
for (g in 1:10){
  Y[,g] = rbinom ( nrow(Y), 1, P_T1[,g] )
}
#Fill in NA
for (g in 1:9){
  Y[ Y[,g] == 1, (g+1):10]  <- NA
}
### Take trimester 1 as an example
### Fit discrete time survival model
time = rep (c(1:10), each = N.con*N.day)
fit = glm (c(Y)~c(T1)+factor(time),family = "binomial")

### Fit logistic regression
dat.sim$PTB = rowSums(Y==1, na.rm = T)
fit = glm (dat.sim$PTB~T1[,1],family = "binomial")

### Fit case-crossover design
### Simulate birth at gestational week 24 to 39
Y = matrix (NA, nrow = N.con*N.day, ncol = 16)
for (g in 1:16){
  Y[,g] = rbinom (nrow(Y), 1, P_Lag1[,g])
}
#Fill in NA
for (g in 1:15){
  Y[ Y[,g] == 1, (g+1):16]  <- NA
}
Lag1.16<-Lag1
Lag1.10<-Lag1[,4:13]
Y.16<-Y
Y.10<-Y[,4:13]

table10<-matrix(NA,nrow=10*N.con*N.day,ncol=7)
colnames(table10)<-c("ID","Exposure","Year","Month","Date","Week","Birth")
table10[,1]<-rep(1:(N.con*N.day),10)
for (ii in 1:10){
  mm<-N.con*N.day*(ii-1)+c(1:(N.con*N.day))
  table10[mm,2]<-Lag1.10[,ii]
}
for (ii in 1:10){
  mm<-N.con*N.day*(ii-1)+c(1:(N.con*N.day))
  table10[mm,7]<-Y.10[,ii]
}

table16<-matrix(NA,nrow=16*N.con*N.day,ncol=7)
colnames(table16)<-c("ID","Exposure","Year","Month","Date","Week","Birth")
table16[,1]<-rep(1:(N.con*N.day),16)
for (ii in 1:16){
  mm<-N.con*N.day*(ii-1)+c(1:(N.con*N.day))
  table16[mm,2]<-Lag1.16[,ii]
}
```

```
for (ii in 1:16){
  mm<-N.con*N.day*(ii-1)+c(1:(N.con*N.day))
  table16[mm,7]<-Y.16[,ii]
}

### a1-a16, b1-b16, c1-c16, d1-d16 are the dates we generated
table10[,3]<-c(a4,a5,a6,a7,a8,a9,a10,a11,a12,a13)
table10[,4]<-c(b4,b5,b6,b7,b8,b9,b10,b11,b12,b13)
table10[,5]<-c(c4,c5,c6,c7,c8,c9,c10,c11,c12,c13)
table10[,6]<-c(d4,d5,d6,d7,d8,d9,d10,d11,d12,d13)

table16[,3]<-c(a1,a2,a3,a4,a5,a6,a7,a8,a9,a10,a11,a12,a13,a14,a15,a16)
table16[,4]<-c(b1,b2,b3,b4,b5,b6,b7,b8,b9,b10,b11,b12,b13,b14,b15,b16)
table16[,5]<-c(c1,c2,c3,c4,c5,c6,c7,c8,c9,c10,c11,c12,c13,c14,c15,c16)
table16[,6]<-c(d1,d2,d3,d4,d5,d6,d7,d8,d9,d10,d11,d12,d13,d14,d15,d16)

#Find the IDs that had a preterm birth
table10[is.na(table10[,7]),7]<-0
table10.1<-table10[order(table10[,1]),]
table10.2<-table10.1[table10.1[,7]==1,]
table10.2<-data.frame(table10.2)

table16[is.na(table16[,7]),7]<-0
table16.1<-table16[order(table16[,1]),]
table16.1<-data.frame(table16.1)

table16.2<-subset(table16.1,ID%in%table10.2$ID)

table16.2$label<-paste(table16.2$ID,table16.2$Month)
junk=subset(table16.2,Birth==1)
table16.3<-subset(table16.2,label%in%junk$label)

table16.4<-cbind(rep(1:length(table10.2$ID),each=4),table16.3[,c(2,7)])
colnames(table16.4)<-c("group","exposure","ptb")
table16.4<-data.frame(table16.4)

fit<-clogit(ptb~exposure+strata(group),data=table16.4)

### Fit time-series analyses
Y = matrix (NA, nrow = N.con*N.day, ncol = 10)
for (g in 1:10){
  Y[,g] = rbinom ( nrow(Y), 1, P_Lag1[,g] )
}
#Fill in NA
for (g in 1:9){
  Y[ Y[,g] == 1, (g+1):10]  <- NA
}
#Create a data frame that help to find cases and controls
table10<-data.frame(matrix(NA,nrow=10*N.con*N.day,ncol=5))
colnames(table10)<-c("ID","Exposure","Date","Week","Birth")
table10[,1]<-rep(1:(N.con*N.day),10)
for (ii in 1:10){
  mm<-N.con*N.day*(ii-1)+seq(1,N.con*N.day,by=1)
  table10[mm,2]<-Lag1[,ii]
}
Date.week1<-rep(seq(as.Date("2001-01-05"),as.Date("2001-01-05")+N.day-1,by=1),each=N.
    con)
for (ii in 1:10){
  mm<-N.con*N.day*(ii-1)+seq(1,N.con*N.day,by=1)
  table10[mm,3]<-as.character(Date.week1+7*(ii-1))
}
table10[,4]<-rep(27:36,each=N.con*N.day)
for (ii in 1:10){
  mm<-N.con*N.day*(ii-1)+seq(1,N.con*N.day,by=1)
  table10[mm,5]<-Y[,ii]
}
table10$Date<-as.Date(table10$Date)
table10$label<-paste(table10$Date,table10$Week,sep="/")
```

```
table10<-table10[order(table10$Date),]

y<-tapply(table10$Birth,table10$label,sum,na.rm=T)
cases=data.frame(label=names(y),cases=y)
cases[1]<-lapply(cases[1],as.character)
xx<-strsplit(cases$label,"/")
xx1<-do.call(rbind,xx)
cases$Date<-as.Date(xx1[,1])
cases$Week<-xx1[,2]

exp<-table10[,c(2,3)]
exp<-unique(exp)

table10<-table10[order(table10$ID),]
junk<-subset(table10,Birth==1)
table10.1<-subset(table10,ID%in%junk$ID)
table10.1$B_Date<-rep(junk$Date,each=10)
table10.2<-subset(table10,!(ID%in%junk$ID))
table10.2$B_Date<-max(table10$Date)+7
table10.3<-rbind(table10.1,table10.2)
table10.3<-table10.3[order(table10.3$ID,table10.3$Date),]

ind<-seq(from=1,to=10*N.con*N.day,by=10)
ind1<-ind+9

table10.3$Start<-rep(table10.3$Date[ind],each=10)
table10.3$End<-rep(table10.3$Date[ind1],each=10)

control<-data.frame(matrix(NA,nrow=nrow(cases),ncol=3))
colnames(control)=c("Date","Week","control")
control[,1:2]<-cases[,3:4]
for (jj in 1:nrow(cases)){
  print(c(kk,jj))
  temp<-table10.3[table10.3$Date==cases$Date[jj] & table10.3$Week==cases$Week[jj],]
  control[jj,3]<-sum(temp$Date>=temp$Start & temp$Date<=temp$End & temp$Date<temp$B_
      Date)
}
case.control<-merge(cases,control,by=c("Date","Week"))
final<-merge(case.control,exp,by="Date")
ttt<-which(table(table10$Date)==500)
tt<-names(ttt)
final.new<-final[final$Date%in%as.Date(tt),]
fit.t<-glm(cases~Exposure+offset(log(control))+factor(Week),family="poisson",data=
    final.new)
```