IEA

International Epidemiological Association

OXFORD

Education corner

Time-stratified case-crossover studies for aggregated data in environmental epidemiology: a tutorial

Aurelio Tobias [ID] ,[1,]* Yoonhee Kim[2] and Lina Madaniyazi [ID] [3]

[1]Institute of Environmental Assessment and Water Research (IDAEA), Spanish Council for Scientific Research (CSIC), Barcelona, Spain, [2]Department of Global Environmental Health, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan and [3]School of Tropical Medicine and Global Health, Nagasaki University, Nagasaki, Japan

*Corresponding author. Institute of Environmental Assessment and Water Research (IDAEA), Spanish Council for Scientific Research (CSIC), C/ Jordi Girona 18-26, 08034 Barcelona, Spain. E-mail: aurelio.tobias@idaea.csic.es

Abstract

The case-crossover design is widely used in environmental epidemiology as an effective alternative to the conventional time-series regression design to estimate short-term associations of environmental exposures with a range of acute events. This tutorial illustrates the implementation of the time-stratified case-crossover design to study aggregated health outcomes and environmental exposures, such as particulate matter air pollution, focusing on adjusting covariates and investigating effect modification using conditional Poisson regression. Time-varying confounders can be adjusted directly in the conditional regression model accounting for the adequate lagged exposure–response function. Time-invariant covariates at the subpopulation level require reshaping the typical time-series data set into a long format and conditioning out the covariate in the expanded stratum set. When environmental exposure data are available at geographical units, the stratum set should combine time and spatial dimensions. Moreover, it is possible to examine effect modification using interaction models. The time-stratified case-crossover design offers a flexible framework to properly account for a wide range of covariates in environmental epidemiology studies.

Keywords: Time-stratified case-crossover, environmental epidemiology, air pollution, conditional Poisson regression.

Key Messages

- In environmental epidemiology, the case-crossover design has been used as an effective alternative to the time-series regression design to estimate short-term associations of environmental exposures with a range of acute events.
- The time-stratified approach allows simultaneous control for the long-term trend and seasonality of unmeasured time-varying confounders and the impacts of the day of the week using conditional regression models.
- Conditional regression models can adjust environmental time-varying confounders, accounting for the adequate lagged exposure–response function. Subpopulation time-invariant covariates require reshaping the time-series data set into a long format and conditioning out in the expanded stratum set.
- Effect modification can be investigated by fitting the interaction between environmental exposure and the potential effect modifying covariate in the conditional regression model.

Introduction

The case-crossover design has been frequently used in environmental epidemiology to estimate short-term associations of daily variations in ambient environmental exposures, such as air pollution and temperature, with a range of acute events, including mortality and morbidity.[1] The rationale of the case-crossover design is to compare the exposure level on the day on which the health event occurs (case day) with the exposure levels on nearby days (control days) to look for differences in the exposure that might explain differences in the number of cases.[2]

In the past decades, there has been an ongoing effort to improve the case-crossover design in environmental epidemiology studies.[1] The unidirectional approach, which selects control days only before the case day, and the bidirectional approach, which selects control days both before and after the case day, can lead to biases from long-term trends and seasonal patterns in exposure, as well as non-independent selection of control days.[3] Conversely, the time-stratified approach for selecting referent control periods as per day-of-week within-month and year ensures least biased estimates by accounting for time trends in the environmental exposure. Moreover, it can be tailored to match on specific time-varying confounders using conditional regression models.[3]

This tutorial aims to offer readers a modern introduction to the time-stratified case-crossover design for investigating aggregated environmental exposures, and provide a practical overview of its application in environmental epidemiology

research. It can also be considered an extension of Bhaskaran *et al.*'s[4] tutorial addressing specific time-series regression issues in environmental epidemiology studies, such as long-term trends and seasonal patterns, lagged associations, and autocorrelation. We aim to assemble previous methodological advancements on the case-crossover design in environmental epidemiology studies (Supplementary material, Section A, available as Supplementary data at *IJE* online) in a practical and comprehensive tutorial illustrating the main concepts for statistical analysis through real example data sets. Stata and R codes to replicate the analyses are available in the Supplementary material (available as Supplementary data at *IJE* online).

Examples data sets

We collected daily mortality counts, mean temperature (°C), relative humidity (%), and 24-h average concentrations of particulate matter with aerodynamic diameter $<10\,\mu g/m^3$ (PM_{10}) for the cities of Valencia (Spain) and London (UK) between 2002 and 2006. For Valencia, mortality counts are classified into two age groups (<65 and ≥ 65 years), whereas for London, mortality counts are only available for all ages. Additional details can be found in Supplementary material, Section B (available as Supplementary data at *IJE* online). These data sets have been used in previous studies and are open-access.[4,5]

Implementation of the time-stratified case-crossover design

The first step in a case-crossover study is matching control days to case days in a reference window. The most common time-stratified approach compares the ambient exposure concentrations (e.g. air pollution) on the day on which the health event occurs (i.e. case days) with the concentrations on the same day-of-week within-month (i.e. control days) (Figure 1). We split the time series into equally sized non-overlapping strata defined as day-of-week within-month and year. This enables simultaneous control for the long-term trends and seasonality of unmeasured time-varying confounders and the influence of the day of the week.

The second step is to fit an appropriate regression model to investigate whether changes in environmental exposure can explain some of the short-term variations in the outcome. Here, the conventional approach in case-crossover studies to imply conditional logistic regression on an expanded data set similar to the individually matched case–control format. For each death occurring on a given day, that day of death is coded as a 'case', whereas the other days in the same stratum are coded as 'controls'. This allows conditional logistic regression to analyse individual-level exposure data. However, in the typical situation of working with aggregated-level exposure data, it is computationally less efficient than other alternatives, such as the weighted conditional logistic model (where the weights are the number of events in each day) and conditional Poisson regression. Notably, both conditional logistic and weighted conditional regression models cannot allow for overdispersion or autocorrelation in the original counts, which can mislead estimates of standard errors.[6] In contrast, the conditional Poisson model does not require transforming the data from the typical time-series format to an individually matched case–control format. This simplifies computation by conditioning the parameters on the total counts for the health outcome in each stratum.[6] Moreover, the conditional Poisson model can account for overdispersion and autocorrelation if the observations are not independent within and across strata. It is crucial to note, however, that the strata with all zero counts will be eliminated from the conditional regression model fit in Stata and R software because they do not contribute to the likelihood (Supplementary material, Section C, available as Supplementary data at *IJE* online).

Exposure–outcome association and environmental time-varying confounders adjustment

We first evaluated the association between PM_{10} and daily mortality using the Valencia data set. The 1826 observations were split into 420 stratum sets defined by day-of-week within-month and year (Table 1a). Conditioning on the time-stratified stratum sets allows for controlling long-term trends and seasonality by design.[3,6]

However, in air pollution studies, other time-varying confounders, such as temperature and relative humidity, should be considered.[7] Time-varying confounders can be adjusted in the conditional regression model accounting for the adequate lagged exposure–response function. We fitted a conditional Poisson regression model:

$$E[\text{Log}\,(Y_{t,s=\text{dow}\,\times\,\text{month}\,\times\,\text{year}})]$$
$$= \alpha_s + \text{ns}(\text{temp}_{t,l}) + \text{ns}(\text{hum}_{t,l}) + \beta PM10_{t,l}$$

where $Y_{t,s}$ is the mortality on day t assumed to follow a Poisson distribution with overdispersion (i.e. quasi-Poisson) conditioned on the sum of events in each stratum s defined by day-of-week within-month and year. We used two natural cubic spline functions with 6 degrees of freedom (df) for the 4-day moving average of temperature and 3 df for relative humidity. We fitted a linear function for PM_{10} to examine the short-term effects up to 3 preceding days.

We observed a positive association for PM_{10} at lag 0, gradually diminishing in magnitude and significance thereafter (Supplementary Figure S1, available as Supplementary data at *IJE* online). However, when analysing the London data
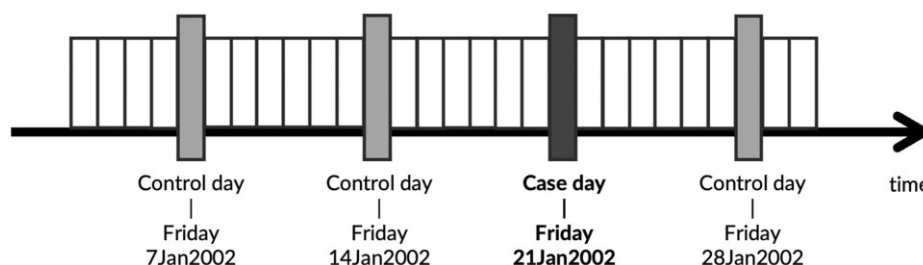
Figure 1. Illustrative example of time-stratified control day selection

Table 1. Excerpt from the example data set with stratum by day-of-week within-month and year

(a) Valencia data set in wide format

Date	Stratum	Death <65 years	Death ≥65 years	PM_{10} (µg/m³)	Temperature (°C)
07Jan2002	2002-January-Friday	6	19	48.9	8.8
14Jan2002	2002-January-Friday	2	15	51.6	12.6
21Jan2002	2002-January-Friday	1	14	57.6	9.3
28Jan2002	2002-January-Friday	2	21	45.8	12.1

(b) Valencia data set in long format

Date	Stratum	Age group	Death counts	PM_{10} (µg/m³)	Temperature (°C)
07Jan2002	2002-January-Friday	<65	6	48.9	8.8
07Jan2002	2002-January-Friday	≥65	19	48.9	8.8
14Jan2002	2002-January-Friday	<65	2	51.6	12.6
14Jan2002	2002-January-Friday	≥65	15	51.6	12.6
21Jan2002	2002-January-Friday	<65	1	57.6	9.3
21Jan2002	2002-January-Friday	≥65	14	57.6	9.3
28Jan2002	2002-January-Friday	<65	2	45.8	12.1
28Jan2002	2002-January-Friday	≥65	21	45.8	12.1

(c) Combined data set for Valencia and London in long format

Date	Stratum	City	Death counts	PM_{10} (µg/m³)	Temperature (°C)
07Jan2002	2002-January-Friday	Valencia	25	48.9	8.8
07Jan2002	2002-January-Friday	London	180	48.5	5.2
14Jan2002	2002-January-Friday	Valencia	17	51.6	12.6
14Jan2002	2002-January-Friday	London	184	48.0	9.3
21Jan2002	2002-January-Friday	Valencia	15	57.6	9.3
21Jan2002	2002-January-Friday	London	189	49.2	10.8
28Jan2002	2002-January-Friday	Valencia	23	45.8	12.1
28Jan2002	2002-January-Friday	London	204	54.1	10.3

PM_{10}, particulate matter with aerodynamic diameter <10 µg/m³.

set, we found significant associations at lags 0 and 1. To simplify the illustration of forthcoming examples and for comparability with previous studies,[8] we fitted a linear term for the 2-day moving average of PM_{10}, representing the average exposure over the current and previous days. In Valencia, a rise of 10 µg/m³ in the 2-day moving average of PM_{10} was associated with a mortality risk increase of 2.3% [relative risk (RR) = 1.023; 95% CI = 1.003, 1.043)], whereas in London, the risk was 0.7% (RR = 1.007; 95% CI = 1.003, 1.010).

Adjustment of subpopulation time-invariant covariates

Researchers frequently seek to incorporate subpopulation characteristics (e.g. age group) as covariates in environmental epidemiology studies. Here, the conventional time-series data format must be reshaped from a wide to a long format. In the typical scenario, when environmental exposures are summarized as daily average concentrations by city, the subpopulation levels (e.g. <65 and ≥65 years) share the same daily exposure concentrations in the stratum set (Table 1b). Time-invariant subpopulation characteristics are controlled by design, making within-patient comparisons[3] conditioning on the subpopulation covariate as well as the time variables (i.e. age-time-stratified):

$$\begin{aligned} & E[\text{Log}\,(Y_{t,s=\text{age}\times\text{dow}\times\text{month}\times\text{year}})] \\ & = \alpha_s + ns(\text{temp}_{t,l}) + ns(\text{hum}_{t,l}) + \beta PM10_{t,l} \end{aligned}$$

Note that, when using aggregated-level exposure data, this model produces the same estimates as those obtained by fitting the age variable into the typical time-stratified approach,

defined by day-of-week within-month and year. However, in the age-time-stratified model, the parameters for the age variable are not estimated since they are conditioned out in the expanded stratum set. Moreover, this adjustment will not alter the estimates obtained when analysing the aggregated data set by age group.

In the Valencia data set, we found the same mortality risk estimate when analysing the total death counts using the typical time-series data in wide format (RR = 1.023; 95% CI = 1.003, 1.043) since the parameters in the conditional Poisson model are conditioned on the sum of death counts in each stratum.[6]

Investigation of effect modification

More interesting research questions might include whether the health effects of environmental exposures differ according to subpopulation characteristics.[9] In the case-crossover design, stratified analyses to investigate effect modifiers can be conducted by fitting stratum-specific conditional Poisson models for each age group, <65 and ≥65 years. However, the stratified analysis does not provide a formal homogeneity test between the risk estimates. Alternatively, we can fit a conditional Poisson regression model in the age–time-stratified approach and include the PM_{10}–age group interaction terms to the model:

$$\begin{aligned} & \text{Log}\,[E(Y_{t,s=\text{age}\times\text{dow}\times\text{month}\times\text{year}})] \\ & = \alpha_s + ns(\text{temp}_{t,l}) + ns(\text{hum}_{t,l}) + \beta PM10_{t,l} + \gamma PM10_{t,l} \times \text{age}_t \end{aligned}$$

Note that the main effect for the age variable is not parametrized since the model is already conditioned on age. Moreover, the distribution of the target subpopulation

Table 2. Mortality risk for an increase of 10 μ/m^3 in PM$_{10}$ by age group in Valencia using stratified analysis and interaction

	Stratified analysis		Interaction model	
Age group	Relative risk	95% CI	Relative risk	95% CI
<65 years	1.012	0.965–1.063	1.013	0.967–1.061
≥65 years	1.025	1.003–1.047	1.025	1.003–1.047

PM$_{10}$, particulate matter with aerodynamic diameter <10 μg/m^3.

Table 3. Mortality risk for an increase of 10 μ/m^3 in PM$_{10}$ by city using stratified analyses and interaction

	Stratified analysis		Interaction model	
City	Relative risk	95% CI	Relative risk	95% CI
Valencia	1.023	1.003–1.043	1.023	1.003–1.043
London	1.006	1.003–1.010	1.007	1.003–1.010

PM$_{10}$, particulate matter with aerodynamic diameter <10 μg/m^3.

should be carefully reviewed to check for any strata with all zero outcomes or with only one observation, which would be dropped from the analysis. This may affect stratification by subpopulation characteristics with unusual acute events (e.g. child mortality). Additionally, this model assumes that temperature and relative humidity effects are homogeneous between the age groups.

Both stratification and interaction approaches provided similar results in our example (Table 2); however, the risk estimates from the interaction model were more precise, showing a narrower CI than the stratified analysis. Moreover, the interaction model allows testing for effect modification using a likelihood ratio test compared to the model without interaction. In our example, we found weak evidence of effect modification by age group ($P = 0.6435$).

The ways to investigate effect modification described here are similar to how it would be assessed using any regression model. Thus, as mentioned in a previous section, conditional logistic regression would also be helpful in assessing modification by individual-level continuous factors (e.g. body mass index, smoking pack-years) if this information is available and such modifications are of interest.

Multi-location studies

When environmental exposure data are available at any given geographic unit (e.g. city), exposure values will vary daily in the strata between geographical levels (Table 1c). Therefore, the strata should combine time and spatial dimensions, matching the geographical units by day-of-week within-month and year. This approach is referred to as a space-time-stratified case-crossover:[10]

$$\text{Log}\left[E(Y_{t,s=\text{city}\times\text{dow}\times\text{month}\times\text{year}})\right]$$
$$= \alpha_s + \text{ns}(\text{temp}_{t,l}) + \text{ns}(\text{hum}_{t,l}) + \beta \text{PM10}_{t,l}$$

The combined data set for Valencia and London comprises 3,652 observations split into 840 stratum sets defined day-of-week within-month and year for each city. Here, a 10 μg/m^3 increase in the 2-day moving average of PM$_{10}$ was associated with a mortality risk increase of 0.7% (RR = 1.007; 95% CI = 1.003, 1.011). However, we assume that the PM$_{10}$–mortality associations in Valencia and London are homogeneous when considering the city in the stratum set. Otherwise, we should assess for heterogeneity between locations. Based on the homogeneity assumption, the space–time-stratified approach might be more appropriate for small geographical units (e.g. neighbourhoods, districts).

We can derive location-specific risk estimates using stratified analysis or an interaction model for the space–time-stratified approach:

$$\text{Log}\left[E(Y_{t,s=\text{city}\times\text{dow}\times\text{month}\times\text{year}})\right]$$
$$= \alpha_s + \text{ns}(\text{temp}_{t,l}) + \text{ns}(\text{hum}_{t,l}) + \beta \text{PM10}_{t,l} + \gamma \text{PM10}_{t,l} \times \text{city}_t$$

Here, the interaction model assumes a similar confounding structure for temperature and humidity in Valencia and London, which may be quite a strong assumption. However, the risk estimates for each city slightly vary between the stratified analysis and the interaction model (Table 3). Nevertheless, the likelihood ratio test may provide some indication of effect modification by city ($P = 0.1059$). An alternative approach could involve incorporating random effects specific to each location for the environmental exposure variables. In this context, Barrera-Gómez *et al.*[11] recently introduced a Bayesian estimation procedure that allows for incorporating spatial patterns related to the environmental exposure of interest. However, their analyses lead to similar conclusions to the usual frequentist analysis.

Discussion

In this tutorial, we introduced the implementation of the time-stratified case-crossover design in environmental epidemiology studies, including covariates adjustment and investigation of effect modifiers.

The case-crossover study design was first developed for individual-level data to study transient effects of the risk of acute events (e.g. myocardial infarction).[12] However, in environmental epidemiology studies, ambient exposures (e.g. air pollution, temperature) are often assigned using central monitoring stations or gridded exposure surfaces, meaning that individuals who live close to each other share the same exposure. Therefore, the case-crossover design applied in environmental epidemiology is typically an aggregated exposure study at the population level.[2] The time-stratified case-crossover is often viewed as a competing design of the time-series regression. However, the main difference is that in a time-stratified case-crossover, long-term trends and seasonality are controlled by design through conditioning on the time-stratified stratum sets. In contrast, in a time-series regression study, we must adjust modelling long-terms and seasonality using functions of time (e.g. natural cubic splines).[4] The strata in the case-crossover design enable each patient to act as their own control; therefore, estimates will not be affected substantially when the number of observations is low compared with time-series regression analysis. However, Lu and Zeger[13] have already shown that the time-stratified case-crossover design is a particular case of time-series analysis when there is a shared environmental exposure. In our example, we found very close risk estimates for PM$_{10}$ using both designs (Supplementary Figure S1, available as Supplementary data at *IJE* online).

The time-stratified approach can be extended further to a space–time-stratified case-crossover design[10] to analyse multilevel data in one step as an alternative to the two-stage design.[14] In our example, we found similar risk estimates for PM10 between the space–time-stratified case-crossover estimates and combining the city-specific estimates for Valencia and London using a two-stage design (Supplementary Table S2, available as Supplementary data at *IJE* online). The time-stratified case-crossover design can also contribute to investigating effect modification due to subpopulation characteristics and geographical locations. Although the interaction model has a higher power to detect heterogeneity and better covariate control than stratified analysis,[15] it assumes a common confounding structure across strata. This assumption should be explored to determine its appropriateness in the data when including subpopulation characteristics and for multi-location studies.

However, our examples have focused on the typical situation in which the environmental exposure of interest is collected as aggregated data at the city level. In the case of the availability of individual exposure measurements, we recommend the recently developed case–time-series design.[16,17] Both are self-matched designs that allow the control of time-invariant confounders, but the case–time-series design is much more flexible to control for time-varying confounders. As stated in this tutorial, a case-time-stratified crossover design compares the exposure on the case day with their own control days from a different time period (i.e. day-of-week within-month and year). On the other hand, the case–time-series design involves multiple cases and controls for confounding factors using a within-individual comparison. The follow-up period is split into equally spaced time intervals, resulting in a set of multiple case-level time series, which is suitable for investigating the short-term effects of time-varying exposures on acute health outcomes and for the analysis of longitudinal data.[16]

In conclusion, the time-stratified case-crossover design offers an effective alternative to conventional time-series regression studies in environmental epidemiology, with further extensions to multilevel data and the assessment of effect modification.

Ethics approval

Not applicable as this is not research involving human patients.

Data availability

Stata do-file, R script file and example data sets can be downloaded via open-access from GitHub https://github.com/aureliotobias/casecrossover.

Supplementary data

Supplementary data are available at *IJE* online.

Author contributions

A.T. designed the study and directed its implementation. Y.K. and L.M. assisted in designing the analytical strategy and interpreting the findings.

Funding

Acknowledgements

Conflict of interest

None declared.

References

1. Carracedo-Martinez E, Taracido M, Tobias A, Saez M, Figueiras A. Case-crossover analysis of air pollution health effects: a systematic review of methodology and application. *Environ Health Perspect* 2010;**118**:1173–82.
2. Jaakkola JJ. Case-crossover design in air pollution epidemiology. *Eur Respir J Suppl* 2003;**40**:81s–85s.
3. Janes H, Sheppard L, Lumley T. Case-crossover analyses of air pollution exposure data: referent selection strategies and their implications for bias. *Epidemiology* 2005;**16**:717–26.
4. Bhaskaran K, Gasparrini A, Hajat S, Smeeth L, Armstrong B. Time series regression studies in environmental epidemiology. *Int J Epidemiol* 2013;**42**:1187–95.
5. Iniguez C, Ballester F, Tobias A. Data supporting the short-term health effects of temperature and air pollution in Valencia, Spain. *Data Brief* 2022;**44**:108518.
6. Armstrong BG, Gasparrini A, Tobias A. Conditional Poisson models: a flexible alternative to conditional logistic case cross-over analysis. *BMC Med Res Methodol* 2014;**14**:122.
7. Buckley JP, Samet JM, Richardson DB. Commentary: Does air pollution confound studies of temperature? *Epidemiology* 2014;**25**:242–45.
8. Liu C, Chen R, Sera F *et al*. Ambient Particulate Air Pollution and Daily Mortality in 652 Cities. *N Engl J Med* 2019;**381**:705–15.
9. Stafoggia M, Forastiere F, Agostini D *et al*. Vulnerability to heat-related mortality: a multicity, population-based, case-crossover analysis. *Epidemiology* 2006;**17**:315–23.
10. Wu Y, Li S, Guo Y. Space-Time-Stratified Case-Crossover Design in Environmental Epidemiology Study. *Health Data Sci* 2021;**2021**:1–3.
11. Barrera-Gomez J, Puig X, Ginebra J, Basagana X. Conditional Poisson Regression with Random Effects for the Analysis of Multi-site Time Series Studies. *Epidemiology* 2023;**34**:873–78.
12. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. *Am J Epidemiol* 1991;**133**:144–53.
13. Lu Y, Zeger SL. On the equivalence of case-crossover and time series methods in environmental epidemiology. *Biostatistics* 2007;**8**:337–44.
14. Sera F, Gasparrini A. Extended two-stage designs for environmental research. *Environ Health* 2022;**21**:41.
15. Kontopantelis E, Sperrin M, Mamas MA, Buchan IE. Investigating heterogeneity of effects and associations using interaction terms. *J Clin Epidemiol* 2018;**93**:79–83.
16. Gasparrini A. The Case Time Series Design. *Epidemiology* 2021;**32**:829–37.
17. Gasparrini A. A tutorial on the case time series design for small-area analysis. *BMC Med Res Methodol* 2022;**22**:129.