Epidemiology Publish Ahead of Print

DOI: 10.1097/EDE.0000000000001410

The Case Time Series Design

Antonio Gasparrini[1*,2]

[1]Department of Public Health Environments and Society, London School of Hygiene & Tropical Medicine, London UK

[2]Centre for Statistical Methodology, London School of Hygiene & Tropical Medicine, London UK

*Correspondence to: Antonio Gasparrini, London School of Hygiene & Tropical Medicine, 15-17 Tavistock Place, London WC1H 9SH, UK. Telephone: 0044 (0)20 79272406. E-mail: antonio.gasparrini@lshtm.ac.uk.

Running title: The case time series design

Funding: This work was supported by the Medical Research Council-UK [Grant ID: MR/R013349/1].

Competing financial interests: The author declares he has no actual or potential conflict of interest.

Data and code: Online supplemental material includes documents for simulating data with the same features of the datasets used in the two case studies, and for reproducing the steps and results of the analyses presented in the article. An updated version complemented with scripts of the R statistical software is available at https://github.com/gasparrini/CaseTimeSeries.

Acknowledgments: The author is thankful to Dr Charlotte Warren-Gash, and Dr Fay Johnston and Mr Iain Koolhof for providing data access and information for the two case studies used as illustrative examples. The author is also grateful to colleagues who provided comments on various drafts of the manuscript and analyses, in particular Mr Francesco Sera,

Dr Ana Maria Vicedo-Cabrera, and Prof Ben Armstrong. Finally, the author is indebted to Prof Paddy Farrington for offering critical insights on asymptotic biases of maximum likelihood estimators in self-controlled case series. The study on influenza and AMI was originally approved by the Independent Scientific Advisory Committee (ISAC) of the Clinical Practice Research Datalink (Ref: 09_034), the Cardiovascular Disease Research Using Linked Bespoke Studies and Electronic Records (CALIBER) Scientific oversight committee and Myocardial Ischaemia National Audit Project (MINAP) Academic Group (ref: 09_08), and the UCL Research Ethics committee (Ref: 2219/001). This study, which used the analysis dataset only, was approved through a minor ISAC amendment (granted on 12/01/2016) and a MINAP Academic Group amendment (granted on 11/01/2016). More information about AirRater are available at https://airrater.org.

ABSTRACT

Modern data linkage and technologies provide a way to reconstruct detailed longitudinal profiles of health outcomes and predictors at the individual or small-area level. While these rich data resources offer the possibility to address epidemiologic questions that could not be feasibly examined using traditional studies, they require innovative analytical approaches. Here we present a new study design, called case time series, for epidemiologic investigations of transient health risks associated with time-varying exposures. This design combines a longitudinal structure and flexible control of time-varying confounders, typical of aggregated time series, with individual-level analysis and control-by-design of time-invariant between-subject differences, typical of self-matched methods such as case–crossover and self-controlled case series. The modeling framework is highly adaptable to various outcome and exposure definitions, and it is based on efficient estimation and computational methods that make it suitable for the analysis of highly informative longitudinal data resources. We assess the methodology in a simulation study that demonstrates its validity under defined assumptions in a wide range of data settings. We then illustrate the design in real-data examples: a first case study replicates an analysis on influenza infections and the risk of myocardial infarction using linked clinical datasets, while a second case study assesses the association between environmental exposures and respiratory symptoms using real-time measurements from a smartphone study. The case time series design represents a general and flexible tool, applicable in different epidemiologic areas for investigating transient associations with environmental factors, clinical conditions, or medications.

Keywords: study design; self-matched; self-controlled; case-only; time series; epidemiological methods; longitudinal data; AirRater.

BACKGROUND

Observational studies aim to discover and understand causal relationships between exposures and health outcomes through the analysis of epidemiologic data.[1] Paramount to this objective is removing biases due to the non-experimental setting, in the first place confounding. It is, therefore, no surprise that traditional approaches based on cohort and case–control methods have been complemented with, and extended by, alternative study designs and statistical techniques applicable in specific contexts. An active area of research is so-called self-matched studies, which investigate acute effects of intermittent exposures by comparing observations sampled at different times within the same unit. These include individual-level designs such as the case–crossover,[2] the case-only,[3] the case–time–control,[4] the exposure–crossover,[5] and the self-controlled case series,[6] among others. An alternative but related epidemiologic method for aggregated data is the time series design, applied in particular in environmental studies.[7] A thorough overview of self-matched methods is provided in a recent publication by Mostofsky and colleagues.[8]

This landscape is likely to be transformed further by ongoing technologic and methodologic developments in data science, which offers unique opportunities for epidemiologic investigations, for instance through electronic health records linkage,[9] exposure modeling,[10] and real-time measurements technologies.[11,12] Ultimately, these data resources can be used to reconstruct detailed longitudinal profiles with repeated measures of health outcomes and various risk factors, offering the chance to investigate complex aetiological mechanisms and to test elaborate causal hypotheses. However, existing self-matched methods present limitations in this context, and new analytical techniques must be developed for epidemiologic investigations in these intensive longitudinal and big data settings.[13]

In this contribution, we present the *case time series design*, a novel self-matched method for the analysis of transient changes in risk of acute outcomes associated with time-varying

exposures. This innovative design combines the longitudinal modeling structure of time series analysis with the individual-level setting of other self-matched methods, offering a flexible and generally applicable tool for modern epidemiologic studies. First, we introduce the case time series design and its features, including the design structure, modeling framework, estimation methods, and key assumptions. Later, we assess the methodology in a simulation study that evaluates its performance under various data generating scenarios. Then, we demonstrate its application through two real-data epidemiologic analyses. In a final discussion section, we describe the epidemiologic context, advantages, and limitations, and areas of further development. We add documents for reproducing real-data examples and the simulation study as eAppendix 1-3 in the online supplementary material; http://links.lww.com/EDE/B841, with an updated version complemented with and R scripts available at the personal website and GitHub webpage of the author (see 'Data and Code').

A NOVEL SELF-MATCHED DESIGN

The study design proposed here, called case time series, is a generally applicable tool for the analysis of transient health associations with time-varying risk factors. This novel design considers multiple observational units, defined as cases, for which data are longitudinally collected over a pre-defined follow-up period. The main design feature that defines the case time series methodology is the split of the follow-up period in equally spaced time intervals, which results in a set of multiple case-level time series. Data forming the series can originate from actual sequential observations or be reconstructed by aggregating or averaging longitudinal measurements, but, eventually, they are assumed to represent a continuous temporal frame. A graphical representation is provided in Figure 1, showing case-specific time series data with various types of measurements of outcome and exposure collected for multiple subjects.

The case time series data setting provides a flexible framework that can be adapted for studying a wide range of epidemiologic associations. For instance, outcomes, exposures, and other predictors can be represented by either indicators for events, episodes, or continuous measurements that vary across units and times, as in Figure 1. The time intervals can be of any length (from seconds to years), depending on the temporal association between outcome and exposures and on practical design considerations. A case is a general definition, and it can represent a subject or other entities such as a geographic area to which observations are assigned, thus allowing analyses to be conducted either at individual level or with aggregated data. Eventually, the case time series structure combines characteristics of various other study designs: it allows individual-level analyses of transient risk associations as in traditional self-matched methods, but it retains the longitudinal temporal frame typical of time series data, with ordered repeated measures of outcomes, exposures, and other predictors. As discussed below, this flexible design setting offers important advantages.

Modeling Framework

A case time series model can be written in a regression form by defining the expectation of a given health outcome y_{it} for case i at time t in relation to a series of predictor terms. Algebraically, the model can be written as:

$$g[\mathrm{E}(y_{it})] = \xi_{i(k)} + f(x_{it}, \ell) + \sum_{j=1}^{J} s_j(t) + \sum_{p=1}^{P} h_p(z_{ipt}) \qquad (1)$$

The definition in Eq. (1) resembles a classic time series regression model traditionally used in environmental epidemiology, where the ordered and sequential nature of the data allows the application of cutting-edge analytical techniques.[7] Specifically, the function $f(x, \ell)$ specifies the association with the exposure of interest x, defined either as a binary episode indicator or as a continuous variable, optionally allowing for non-linearity and complex temporal

dependencies along the lag dimension ℓ. These complex relationships can be modeled through distributed lag linear and non-linear models, which can flexibly define cumulative effects of multiple exposure episodes.[14] The term(s) s_j represent functions expressed at different timescales to model temporal variations in risk associated to underlying trends or seasonality, among others.[15] Other measurable time-varying confounders z_p can be modeled through functions h_p, and these can include for instance age or time since a specific intervention. The two sets of terms s_j and h_p ensure a strict control of temporal variation in risks over multiple time axes. The outcome y can represent binary indicators, counts of rare or frequent events, or continuous measures. The analysis can be performed on multiple cases $i = 1, \dots, n$, with intercepts $\xi_{i(k)}$ expressing baseline risks for different risk sets, optionally stratified further in time strata $k = 1, \dots, K_i$ nested within them, allowing an additional within-case control for temporal variations in risk.

Estimation

The estimation procedures in case time series analyses rely on estimators and efficient computational algorithms provided by the general framework of fixed-effects models.[16] These were developed in econometrics and often applied in panel studies with repeated observations.[10,17] Fixed-effects methods allow the estimation of coefficients for the various functions in Eq. (1), without including the potentially high number of case/stratum-specific intercepts $\xi_{i(k)}$, treated as nuisance (or incidental) parameters.[16]

Fixed-effects estimators are available for the three main types of outcomes and distributions within the extended exponential family of generalized linear models (GLMs). Specifically, for continuous outcomes with a Gaussian distribution, the estimation procedure involves mean-centring and a simple correction of the degrees of freedom. For event-type indicator or count outcomes following a Bernoulli and Poisson distribution, respectively, estimators for fixed-effects models with canonical logit and log links can be defined through conditional

likelihoods for logistic and Poisson regression.[18,19] These are forms of partial likelihoods that are derived by defining reduced sufficient statistics for $\xi_{i(k)}$, obtained by conditioning on the total number of events within each of the n cases or $n \times K$ strata.

The main advantage of fixed-effects models is that the effect of any unmeasured predictor that does not vary within each risk set is absorbed by the intercept $\xi_{i(k)}$, and therefore the related confounding effect is controlled for implicitly by design, as in other self-matched methods.[8] In addition, the within-case design offers important computational advantages, especially from a big data perspective. First, the analysis is restricted to informative strata, *i.e.* cases and risk sets with variation in both outcome and exposure. Second, the estimators are based on efficient computational schemes, where the conditional or fixed-effect likelihood is defined by the sum of parts related to multiple risk sets, and the corresponding nuisance parameters $\xi_{i(k)}$ are not directly estimated.

Key assumptions and threats to validity

As discussed above, the case time series framework has interesting design and modeling features that offer important advantages. On the other hand, its self-controlled structure, while appealing, only operates within an elementary causal framework and requires relatively strict assumptions to protect against key threats to validity. Specifically, the main requirements are the following:

1. *Distributional assumptions on the outcome*. The outcome y_{it} must represent conditionally independent observations originating from one of the standard family distributions, for instance, Poisson counts, Bernoulli binary indicators, or Gaussian continuous measures.
2. *Outcome-independent follow-up period*. The period of observation for each case i must be independent of a given outcome, meaning that the follow-up period cannot be defined or modified by the outcome itself.

3. *Outcome-independent exposure distribution*. The probability of the exposure x_t must be independent of the outcome history prior to t, meaning that the occurrence of a given outcome must not modify the exposure distribution in the following period.
4. *Constant baseline risk conditionally on measured time-varying predictors*. The baseline risk along the (strata of) follow-up period of each case i must be constant, meaning that variations in risks must be fully explained by model covariates.

These requirements enable valid conditional comparison of observations at different times within the follow-up of each case. Departures from these assumptions can produce imbalances in the temporal distribution of the outcome, the exposure, or unmeasured risk factors, thus determining spurious associations.

Some of these assumptions have been separately described in the literature of self-matched designs and fixed-effects models.[20-23] Specifically, Assumption 1 dictates that outcomes must occur independently, and in particular that the occurrence of a given outcome level or event must not modify the risk of following outcomes.[24] This assumption indirectly implies that outcomes are recurrent, and non-recurrent events can only be analysed if rare in the population of interest.[25,26] Assumptions 2 and 3 are those posing more limitations to the application of self-matched methods, as for many associations of interest an outcome can modify both the follow-up period and exposure distribution.[27,28] These requirements often restrict the case time series designs to the analysis of exogenous exposures, which are by definition outcome-independent, and for which the observation period can be extended even beyond a terminal event, as in bi-directional case–crossover schemes.[29] Assumption 4 requires a constant baseline risk to ensure conditional exchangeability between observations within each risk sets,[20,30,31] requiring that relevant time-varying confounders are included and all the terms in Eq. (1) are correctly specified.

Importantly, the design setting described above is not suited to represent complex causal scenarios characterised by dynamic mechanisms between time-varying terms. Specifically, feedback between outcomes and between outcome and exposure are forbidden by Assumptions 1 and 3, respectively, while more generally exposure–confounder feedback cannot be validly handled through traditional regression-based methods for longitudinal data.[32]

SIMULATION STUDY

We evaluated the performance of the case time series design in a set of simulated scenarios that involved various data-generating processes and assumptions (Table). Detailed information on the simulation settings, definitions, and additional results are provided in eAppendix 3 (online supplementary material; http://links.lww.com/EDE/B841). Briefly, we simulated and analysed data for 500 subjects followed up for one year, testing the method in terms of relative bias, coverage, and relative root mean square error (RMSE) in 50,000 replications. The basic scenario involves an outcome represented by repeated event counts and binary indicators of exposure episodes associated with a constant increase in risk in the next 10 days.

The first part of the simulation study (Scenarios 1-10) evaluates the performance of the new design in recovering the true association under increasingly complex data settings. Specifically, the scenarios depict different outcome and exposure types, the presence of common or subject-specific trends, time-invariant and time-dependent confounders, and more complex lag structures. Results in the Table indicate that the case time series design provides correct point estimates and confidence intervals in almost all ten scenarios. The small underestimation in Scenario 2 is consistent with the asymptotic bias of maximum likelihood estimators originating from the extreme unbalance of expected events between risk and control periods, previously described and defined analytically in the self-controlled case

series literature.[33] eFigure 1 (online supplementary material; http://links.lww.com/EDE/B841) shows that the case time series models can correctly recover the true association, both in the basic Scenario 1 with constant risk and no confounding, and in the more complex Scenario 10 representing varying lag effects, strong temporal trends, and highly correlated confounders.

The second part of the simulation study (Scenarios 11-14) illustrates basic applications, but where each of the four assumptions, in turn, does not hold. Specifically, Scenario 11 describes the case where the occurrence of an outcome can change the risk status of a subject and temporally reduce their underlying risk. This can occur for instance when the event results in the prescription of drugs or therapies. This induces a form of dependency in the outcome series that violates Assumption 1 and, in this example, results in a negative bias (Table). Scenarios 12 simulates a different situation, namely when the outcome event carries a risk of censoring the follow-up, for instance, if it increases the probability of death. This contravenes Assumption 2 and generates a bias in the opposite direction. In Scenario 13, the outcome event reduces instead the probability of exposure episodes in the following two weeks, a situation that can occur for example if the event results in hospitalization or lifestyle changes. Here Assumption 3 does not hold, and the estimators are again biased upward. Finally, Scenario 14 illustrates the case of unobserved periods of lower baseline risk within the follow-up, for instance corresponding to holiday periods with a reduced probability of an outcome being reported. This undermines the conditional exchangeability requirements of Assumption 4 and induces a large positive bias.

ILLUSTRATIVE EXAMPLES

This section illustrates the application of the case time series design in two real-data examples. These case studies are described here only for illustrative purposes, and they are not meant to offer substantive epidemiological evidence on the associations under study.

Detailed information on the setting and sources of data can be found in the cited references. Documents in the online supplementary material; http://links.lww.com/EDE/B841 (eAppendix 1 and 2) provide notes and R code that reproduce the steps of these analyses using simulated data, and they offer details on the specific modeling choices.

Flu and Myocardial Infarction

The first example replicates a published analysis that assessed the role of influenza infection as a trigger for acute myocardial infarction (acute MI).[34] The data, retrieved by linking electronic health records from primary care and cohort databases for England and Wales, include 3,927 acute MI cases with at least one flu episode in the period 2003-2009. A representation of a sub-interval of the follow-up for six subjects is reported in eFigure 2 (online supplementary material; http://links.lww.com/EDE/B841). The original analysis relied on the self-controlled case series design to examine the association, using exposure windows in the 1-91 days after each flu episode and controlling for trends using 5-year age strata and trimester indicators. Limitations of this approach are the use of stratification to describe smooth continuous dependencies and the fact that multiple flu episodes experienced by some subjects resulted in the long exposure windows to overlap (see eFigure 2; http://links.lww.com/EDE/B841), requiring ad-hoc fixes that can generate biases.[35] Conversely, the rarity of the exposure, with most of the subjects experiencing a single flu episode, prevents the application of the case–crossover design, as most control sampling schemes would generate non-discordant case–referent sets.

We replicated the analysis with a case time series design, splitting the follow-up period of each subject into daily time series (see eAppendix 1, online supplementary material; http://links.lww.com/EDE/B841). We fitted a fixed-effects Poisson model to estimate the flu–acute MI association while controlling for underlying trends across multiple time scales. The model includes smooth functions to define the baseline risk, specifically using natural splines

(with two knots at the interquartile range) for age and cyclic splines (with three degrees of freedom) for seasonality. More importantly, we applied distributed lag models defined by either splines (with knots at 3, 10, and 29 lags) or step functions (with strata 1-3, 4-7, 8-14, 15-28, and 29-91 lags) to describe temporal effects along with the exposure window. Results are reported in Figure 2. The left and middle panels display the variation in risk of AMI by age and season, showing how the case time series design allows modeling baseline trends fluctuating smoothly across multiple time axes. The right panel illustrates the risk after a flu episode within the selected lag period, as estimated using a distributed lag model with spline functions. The graph indicates a high risk in the first days after a flu episode, which then attenuates and disappears after approximately one month. The same panel also includes the fit of the alternative DLM defined by step functions, which assumes a constant risk within exposure windows (see also eFigure 3 in the online supplementary material; http://links.lww.com/EDE/B841). This specification matches the stratification approach in the original self-controlled case series analysis,[34] although the case time series design with distributed lag models accounts for cumulative effects of potentially overlapping periods of flu episodes.

Environmental exposures and respiratory symptoms

The second example illustrates a preliminary analysis of the role of multiple environmental stressors in increasing the risk of respiratory symptoms using smartphone technology. Data were collected within AirRater, an integrated online platform operating in Tasmania that combines symptom surveillance, environmental monitoring, and real-time notifications.[12] A smartphone app allowed the self-reported recording of respiratory symptoms and the reconstruction of personalized exposure series by linking geo-located positions with high-resolution spatio-temporal maps derived from environmental monitors (see Figure 3). Standard cohort analyses based on between-subject comparisons are unsuitable in this

complex study setting, characterized by continuous recruitment, high dropout rates, and intermittent participation (see eFigure4 in the online supplementary material; http://links.lww.com/EDE/B841). Similarly, the frequent and highly seasonal outcome pose problems in adopting a case–crossover design, with issues in selecting control times and about the assumption of constant within-stratum risk. Finally, the presence of multiple continuous exposures prevents the application of the self-controlled case series design, either in its standard or extended forms.[36,37]

We, therefore, applied a case time series design (see eAppendix 2, online supplementary material; http://links.lww.com/EDE/B841). The analysis included 1,601 subjects followed between October 2015 and November 2018, with a total of 364,384 person–days. The event-type outcome was defined as daily indicators of reported respiratory symptoms and associated with individual exposure to pollen (grains/m^3), fine particulate matter ($PM_{2.5}$, μg/m^3), and temperature (°C) (Figure 3). We modeled the relationships using a fixed-effects logistic regression over a lag period of 0-3 days, using an unconstrained DLM for the linear association with $PM_{2.5}$, and bi-dimensional spline distributed lag non-linear models for specifying non-linear dependencies with pollen and temperature.[14,38] A strict temporal control was enforced by using subject/month strata intercepts, natural splines of time (with 8 df/year), and indicators of the day of the week, thus modeling individually varying baseline risks on top of shared long-term, seasonal, and weekly trends.

Figure 4 shows the preliminary results, with estimated associations reported as odds ratios (ORs) from the model that includes simultaneously the three environmental stressors. The graphs display the overall cumulative exposure-response relationships (top panels), interpreted as the net effects across lags, and the full bi-dimensional exposure-lag-response associations (bottom panels)[14,38]. The lefthand panels indicate a positive association between risk of allergic symptoms and pollen, with a step increase in risk that flattens out at high

exposures, and a lagged effect up to 2 days. The middle panels suggest an independent association with $PM_{2.5}$, where the risk is entirely limited to the same-day exposure. Finally, results in the righthand panels show a positive association with high ambient temperature, with the OR increasing above 1 beyond daily averages of 15°C.

DISCUSSION

The novel case time series methodology offers a general modeling framework for the analysis of epidemiologic associations with time-varying exposures. The design is adaptable to various data settings for the analysis of highly informative longitudinal measurements, and it is particularly well-suited in applications with modern data resources such as individual-level exposure models and real-time technologies.

The main feature of methodology is a flexible scheme that embeds a longitudinal time series structure in a within-subject design, providing unique modeling advantages. For instance, the sequential order of observations offers the opportunity to assess complex temporal relationships with multiple exposures, where patterns of cumulative effects for linear or non-linear exposure-response dependencies can be easily modeled. Furthermore, the time series and self-controlled features offer a structure that enables strict control for confounding: time-invariant and time-varying factors can be adjusted for by stratifying the baseline risk between and within subjects, respectively, while residual temporal variations can be directly modeled through time-varying predictors that represent confounders or shared trends across multiple time axes.

The new design complements and extends the already rich set of self-matched methods for observational studies described in the epidemiological literature.[8] Previous methodological contributions have highlighted links and similarities between various designs,[18,21,29,30,39-41] and ultimately these can be seen as alternative approaches to model the same risk associations. However, each method relies on different sets of assumptions and modeling

choices, which explain in part their separate areas of application. The case time series methodology, nevertheless, offers a general framework that combines and extends features of existing designs, with important advantages. For example, it borrows flexible modeling tools from aggregated-data time series design, but it implements them in individual-level analyses that allow a finer reconstruction of outcomes, exposures, and other risk factors. It is applicable to assess associations with multiple continuous predictors as the case–crossover design, and it can model recurrent events, either common or rare, as the self-controlled case series analyses, but it can be extended to the analysis of outcomes represented by binary indicators or continuous measures, simply assuming different distributions. Finally, its time series structure allows the application of sophisticated techniques such as smoothing methods and distributed lag models, characterized by well-defined parameterizations, computational efficiency, and standard software implementations. A thorough and critical comparison of the case time series methodology with alternative approaches will be provided in future contributions.

Together with other self-matched methods, the new case time series design is based on strict assumptions to protect against key threats to validity. However, these conditions are not always met in practice, and their violations can lead to important biases. Specifically, the requirement that both exposures and follow-up periods are independent of the outcome poses severe limitations to the application of the method, in particular in clinical and pharmaco-epidemiologic studies. In fact, the temporal distribution of endogenous predictors such as behaviours, clinical therapies, or drug prescriptions are often modified by an outcome event. In contrast, the case time series and other self-controlled designs are well suited for the analysis of exogenous exposures such as environmental factors, as discussed before. Extension to test and relax these strong assumptions have been developed for the self-controlled case series design,[27,28] but further research is needed to implement and assess their

validity in case time series models. Conversely, the new design is well suited to control for temporal confounding that can invalidate the assumption of constant baseline risk, through the stratification of the follow-up period and the inclusion of lagged and smooth continuous terms in the model.

Other limitations and areas of current research must be discussed. First, as a method based on a within-subject comparison, the case time series design is ideal for investigating phenomena with short-term changes in risk relative to the study period, while it is less suitable for the analysis of long-term effects and chronic exposures. In fact, while it is in theory possible to extend indefinitely the lag period within the follow-up interval, there is a limit to which the model can disentangle long-lagged effects from seasonal and other trends.[42] In addition, the splitting of the follow-up period in individual-level time series produces a substantial data expansion, with considerable computational demand especially in the presence of a high number of subjects or long study periods. Schemes based on risk-set sampling, previously proposed for cohort and nested case–control studies,[43-45] are currently under development to address this issue. Finally, the simulation study and the two real-data examples presented basic epidemiological relationships between time-varying variables. However, more complex causal dependencies, involving, for instance, dynamic feedback or multiple pathways, explicitly violate the strict assumptions underpinning the case time series design, and cannot be modeled in the proposed framework. The definition, limitations, and potential extensions of fixed-effects models and related designs within a general causal inference setting is an area of current research.[23]

In conclusion, the case time series design represents a novel epidemiologic method for the analysis of transient health associations with time-varying exposures. Its flexible modeling framework can be adapted to various contexts and research areas, for instance in clinical,

environmental, and pharmaco-epidemiology, and it is suitable for the analysis of intensive longitudinal data provided by modern data technologies.

REFERENCES

1. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. Philadelphia: Lipcott Williams \& Wilkins, 2008.
2. 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.
3. Armstrong BG. Fixed factors that modify the effects of time-varying factors: applying the case-only approach. *Epidemiology*. 2003;14(4):467-472.
4. Suissa S. The case–time–control design. *Epidemiology*. 1995;6(3):248-253.
5. Redelmeier DA. The exposure–crossover design is a new method for studying sustained changes in recurrent events. *Journal of Clinical Epidemiology*. 2013;66(9):955-963.
6. Farrington CP. Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics*. 1995;51(1):228-235.
7. Bhaskaran K, Gasparrini A, Hajat S, Smeeth L, Armstrong B. Time series regression studies in environmental epidemiology. *International Journal of Epidemiology*. 2013;42(4):1187-1195.
8. Mostofsky E, Coull BA, Mittleman MA. Analysis of observational self-matched data to examine acute triggers of outcome events with abrupt onset. *Epidemiology*. 2018;29(6):804-816.
9. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, Smeeth L. Data resource profile: clinical practice research datalink (CPRD). *International Journal of Epidemiology*. 2015;44(3):827-836.
10. Janes H, Sheppard L, Shepherd K. Statistical analysis of air pollution panel studies: an illustration. *Annals of Epidemiology*. 2008;18(10):792-802.

11. Dixon WG, Beukenhorst AL, Yimer BB, Cook L, Gasparrini A, El-Hay T, Hellman B, James B, Vicedo-Cabrera AM, Maclure M, Silva R, Ainsworth J, Pisaniello HL, House T, Lunt M, Gamble C, Sanders C, Schultz DM, Sergeant JC, McBeth J. How the weather affects the pain of citizen scientists using a smartphone app. *NPJ Digital Medicine*. 2019;2(1):1-9.

12. Johnston FH, Wheeler AJ, Williamson GJ, Campbell SL, Jones PJ, Koolhof IS, Lucani C, Cooling NB, Bowman DMJS. Using smartphone technology to reduce health impacts from atmospheric environmental hazards. *Environmental Research Letters*. 2018;13(4):044019.

13. Walls TA, Schafer JL. *Models for Intensive Longitudinal Data* Oxford University Press, 2006.

14. Gasparrini A. Modeling exposure-lag-response associations with distributed lag non-linear models. *Statistics in Medicine*. 2014;33(5):881-899.

15. Touloumi G, Atkinson R, Le Tertre A, Samoli E, Schwartz J, Schindler C, Vonk JM, Rossi G, Saez M, Rabszenko D. Analysis of health outcome time series data in epidemiological studies. *EnvironMetrics*. 2004;15(2):101-117.

16. Gunasekara FI, Richardson K, Carter K, Blakely T. Fixed effects analysis of repeated measures data. *International Journal of Epidemiology*. 2013;43(1):264-269.

17. Arellano M, Honoré B. Panel Data Models: Some Recent Developments. Handbook of Econometrics. Vol. 5 Elsevier, 2001;3229-3296.

18. Xu S, Zeng C, Newcomer S, Nelson J, Glanz J. Use of fixed effects models to analyze self-controlled case series data in vaccine safety studies. *Journal of Biometrics & Biostatistics*. 2012;Suppl 7:006.

19. Allison PD. *Fixed Effects Regression Methods for Longitudinal Data Using SAS*. Cary, NC, USA: Sas Institute Inc, 2005.

20. 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(6):717-726.

21. Lumley T, Levy D. Bias in the case–crossover design: implications for studies of air pollution. *EnvironMetrics*. 2000;11(6):689-704.

22. Whitaker HJ, Ghebremichael-Weldeselassie Y, Douglas IJ, Smeeth L, Farrington CP. Investigating the assumptions of the self-controlled case series method. *Statistics in Medicine*. 2018;37(4):643-658.

23. Imai K, Kim IS. When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science*. 2019;63(2):467-490.

24. Farrington CP, Hocine MN. Within-individual dependence in self-controlled case series models for recurrent events. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2010;59(3):457-475.

25. Whitaker HJ, Steer CD, Farrington CP. Self-controlled case series studies: Just how rare does a rare non-recurrent outcome need to be? *Biometrical Journal*. 2018;60(6):1110-1120.

26. Ghebremichael-Weldeselassie Y, Whitaker HJ. Self-controlled case series methodology. *Annual Review of Statistics and Its Application*. 2019;6:241-261.

27. Farrington CP, Anaya-Izquierdo K, Whitaker HJ, Hocine MN, Douglas I, Smeeth L. Self-controlled case series analysis with event-dependent observation periods. *Journal of the American Statistical Association*. 2011;106(494):417-426.

28. Farrington CP, Whitaker HJ, Hocine MN. Case series analysis for censored, perturbed, or curtailed post-event exposures. *Biostatistics*. 2009;10(1):3-16.

29. Navidi W. Bidirectional case–crossover designs for exposures with time trends. *Biometrics*. 1998;54(2):596-605.

30. Lu Y, Zeger SL. On the equivalence of case–crossover and time series methods in environmental epidemiology. *Biostatistics*. 2007;8(2):337-344.

31. Mittleman MA, Mostofsky E. Exchangeability in the case–crossover design. *International Journal of Epidemiology*. 2014;43(5):1645-1655.

32. Mansournia MA, Etminan M, Danaei G, Kaufman JS, Collins G. Handling time varying confounding in observational research. *British Medical Journal*. 2017;359:j4587.

33. Musonda P, Hocine MN, Whitaker HJ, Farrington CP. Self-controlled case series analyses: small-sample performance. *Computational Statistics & Data Analysis*. 2008;52(4):1942-1957.

34. Warren-Gash C, Hayward AC, Hemingway H, Denaxas S, Thomas SL, Timmis AD, Whitaker H, Smeeth L. Influenza infection and risk of acute myocardial infarction in England and Wales: a CALIBER self-controlled case series study. *Journal of Infectious Diseases*. 2012;2006(11):1652-1659.

35. Whitaker HJ, Farrington CP, Spiessens B, Musonda P. Tutorial in biostatistics: the self-controlled case series method. *Statistics in Medicine*. 2006;25(10):1768-1797.

36. Farrington CP, Whitaker HJ. Semiparametric analysis of case series data. *Journal of the Royal Statistical Society: Series C*. 2006;55(5):553-594.

37. Ghebremichael-Weldeselassie Y, Whitaker HJ, Farrington CP. Spline-based self-controlled case series method. *Statistics in Medicine*. 2017;36(19):3022-3038.

38. Gasparrini A, Armstrong B, Kenward MG. Distributed lag non-linear models. *Statistics in Medicine*. 2010;29(21):2224-2234.

39. Armstrong BG, Gasparrini A, Tobias A. Conditional Poisson models: a flexible alternative to conditional logistic case cross-over analysis. *BMC Medical Research Methodology*. 2014;14(1):122.

40. Greenland S. A unified approach to the analysis of case-distribution (case-only) studies. *Statistics in Medicine*. 1999;18(1):1-15.

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

42. Schwartz J. The distributed lag between air pollution and daily deaths. *Epidemiology*. 2000;11(3):320-326.

43. Langholz B, Goldstein L. Risk set sampling in epidemiologic cohort studies. *Statistical Science*. 1996;11(1):35-53.

44. Borgan O, Goldstein L, Langholz B. Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *The Annals of Statistics*. 1995;23(5):1749-1778.

45. Langholz B, Goldstein L. Conditional logistic analysis of case–control studies with complex sampling. *Biostatistics*. 2001;2(1):63-84.

FIGURES

FIGURE 1. Graphical representation of data configurations for the case time series design applied in the analysis of transient health risks of time-varying exposures. The figure represents three examples of data for three subjects (cases) followed for a period of time, with equally spaced measures of outcome and exposure that form case-level time series. This setting allows the definition of predictors and time axes as unique and sequential observations. The three examples illustrate different measures of outcome and exposure. The former is represented as counts (top), a binary indicator (middle), or a continuous measure (bottom). Similarly, exposure can be represented by a simple binary episode indicator (top), or continuous term (middle and bottom). Continuous variables are represented by shaded colours. The graphical representation demonstrates the potential of the case time series design to be applied in various research areas for modeling associations defined by different types of measurements.

FIGURE 2. Results of the analysis on the association between influenza infection and myocardial infarction (AMI), as relative risk (RR) and 95% confidence intervals. The three panels show the AMI risk by age (left) and by season (middle), and the lag-response curve representing the risk in the 1-91 days after a flu episode (right). The latter is estimated in the main model using natural splines (continuous red line), with superimposed the results from an alternative model using step functions (dashed grey line).

FIGURE 3. Graphical representation of the individual time series of a subject participating in the AirRater study on the association between environmental exposures and respiratory symptoms. The four panels (from top to bottom) display the daily series of counts of allergic events and levels of the three environmental stressors, represented by pollen (grains/m^3), $PM_{2.5}$ ($\mu g/m^3$), and temperature (°C).

FIGURE 4. Results of the analysis on the association between environmental exposures and respiratory symptoms, as odds ratio (OR) and 95% confidence intervals. The three columns of panels show estimated associations with pollen (left, grains/m^3), $PM_{2.5}$ (middle, $\mu g/m^3$), and temperature (right, °C). The top row of panels displays the net risk cumulated in the lag period 0-3 days as overall cumulative exposure-response associations, assumed linear for $PM_{2.5}$ and non-linear for pollen and temperature. The bottom row of panels shows instead the full exposure-lag-response associations, represented as the bi-dimensional risk surface for pollen and temperature or the lag-specific risks for a 10 $\mu g/m^3$ increase in $PM_{2.5}$.

TABLES

TABLE. Results of the simulation study, with nine scenarios representing increasingly complex data settings (Scenarios 1-10), and four additional scenarios simulating data where the key design assumptions are violated (Scenarios 11-14). The table reports empirical figures of relative bias (%), coverage, and relative root mean square error (RMSE, %) in 50,000 replications. A detailed description of the scenarios, definitions, and additional results and graphs are provided in the supplementary material [Appendix A].

SCENARIO	RELATIVE BIAS	COVERAGE	RELATIVE RMSE
SCENARIO 1: BASIC	0.0%	0.951	8.8%
SCENARIO 2: RARE OUTCOME/EXPOSURE	-4.5%	0.951	86.0%
SCENARIO 3: CONTINUOUS EXPOSURE	-0.1%	0.950	15.2%
SCENARIO 4: BINARY OUTCOME	0.3%	0.949	9.1%
SCENARIO 5: CONTINUOUS OUTCOME	0.0%	0.950	14.7%
SCENARIO 6: COMMON TREND	-0.1%	0.950	28.8%
SCENARIO 7: SUBJECT-SPECIFIC TREND	0.1%	0.948	35.2%
SCENARIO 8: UNOBSERVED BASELINE CONFOUNDER	0.2%	0.951	25.8%
SCENARIO 9: TIME-VARYING CONFOUNDER	-0.2%	0.949	35.1%
SCENARIO 10: COMPLEX LAG STRUCTURE	0.0%	0.950	29.2%
SCENARIO 11: OUTCOME-DEPENDENT RISK	-18.9%	0.738	24.7%
SCENARIO 12: OUTCOME-DEPENDENT FOLLOW-UP	16.8%	0.797	22.7%
SCENARIO 13: OUTCOME-DEPENDENT EXPOSURE	11.1%	0.744	14.4%
SCENARIO 14: VARIATION IN BASELINE RISK	40.7%	0.222	43.3%

Figure 1

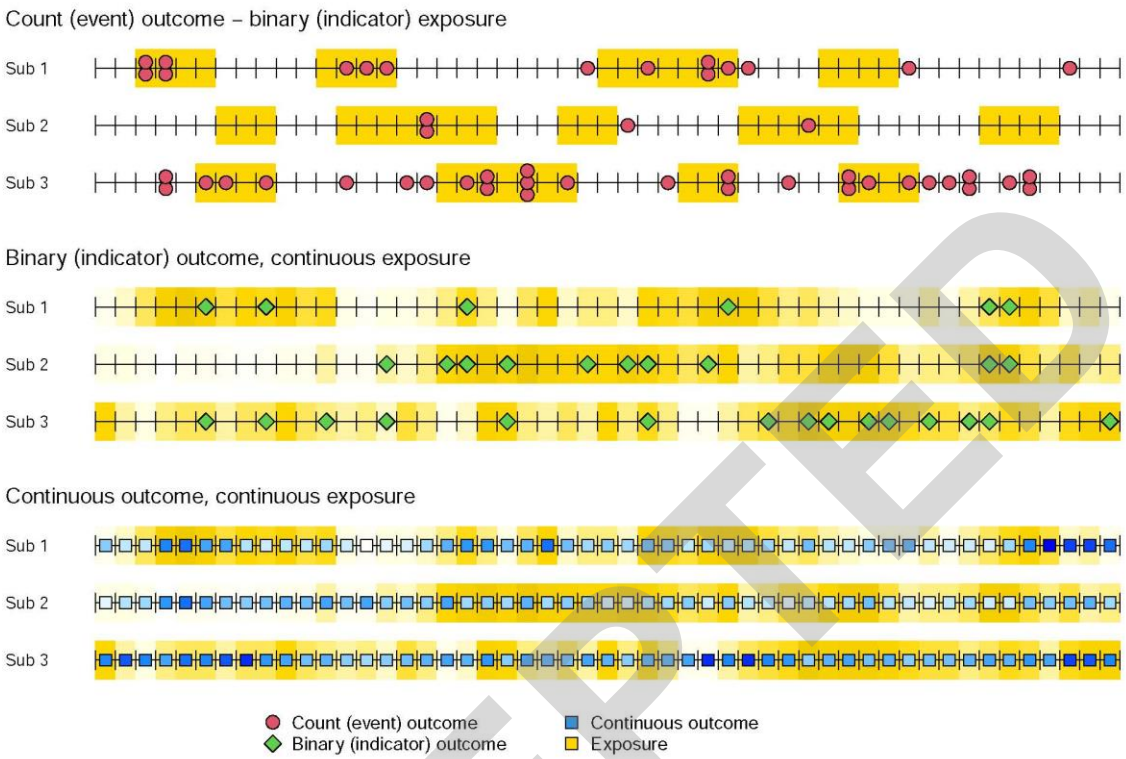

Figure 2

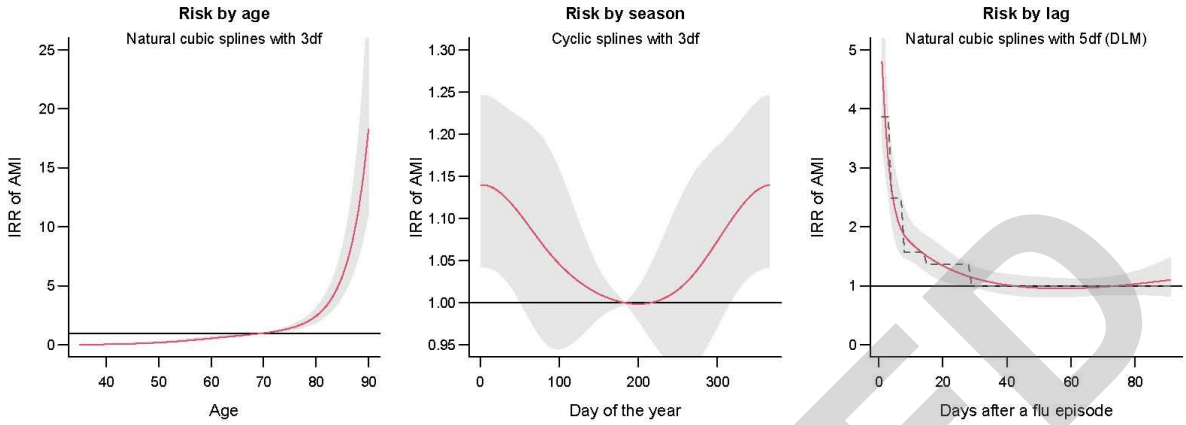

Figure 3

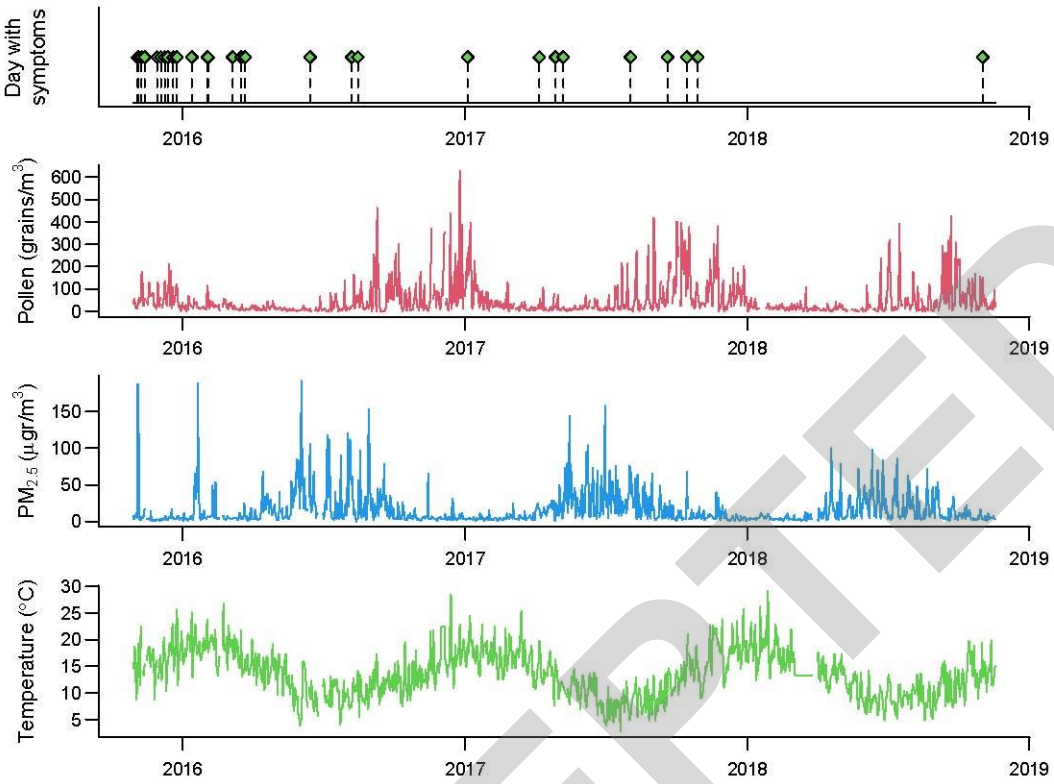

Figure 4