Online supplementary material of the article "The Case Time Series Design"

Antonio Gasparrini[a,b]

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

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

Content:

The case time series design – eAppendix 1

A case study for applications in clinical epidemiology

Antonio Gasparrini

09 August 2021

Contents

This document was originally presented as eAppendix 1 of the article "*The case time series design*," accepted for publication in Epidemiology (Gasparrini 2021), and it reproduces the analysis presented as the first case study. An updated version of this document and related material are available at the GitHub page and at the personal website of the author. The material includes the Rmarkdown files to compile the document, plus scripts with the embedded R code. Note that the code is profiled for clarity, not for speed, with the aim of illustrating the steps of the analysis and the features of the design. It can (probably should) be modified when re-used in real analyses.

This case study illustrates the application of the case time series design in clinical studies. Specifically, the example describes an analysis of the association between acute respiratory infection (flu) and acute myocardial infarction (AMI) using a cohort reconstructed from linked electronic health records. The sample includes 3,927 subjects who experienced a (first) AMI event and had at least one primary care consultation for flu during a pre-determined follow-up period. The analysis illustrates an application of the case time series design with a non-repeated event outcome and binary indicators of exposure episodes. These data were originally presented and analysed with an alternative method in previous publications (Warren-Gash et al. 2012). The code shown below creates and uses simulated data to reproduce the features of the original dataset, which cannot be made publicly available, and the steps and (approximate) results of the application of the case time series design.

Preparation

The following packages are loaded in the session, and need to be installed to run the R code:

```
library(dlnm) ; library(gnm) ; library(mgcv) ; library(pbs)
library(data.table) ; library(scales)
```

We first set a seed to ensure the exact replicability of the results, as the code includes expressions with random number generation, and we also set the graphical parameter `las` for the plots:

```
set.seed(13041975)
par(las=1)
```

Simulating the original data

The data used in this case study are simulated directly in this section. The user can skip it if not of interest, and start with the following section for the data analysis. First, we set the parameters, namely the number of subjects `n` and the date of start and end of follow-up. Note that we reduce the follow-up period to one year, in order to obtain a more manageable dataset. The code:

```
n <- 3927
dstart <- as.Date("2007-01-01")
dend <- as.Date("2007-12-31")
```

Then we generate the time variables across the follow-up period, namely `date` (calendar days), `time` (a sequence of integers starting from 1), `month` (months in numbers), and `doy` (days of the year). In addition, we randomly generate `dob` (date of birth) for each subject, with age at start between 35 and 100 years old.

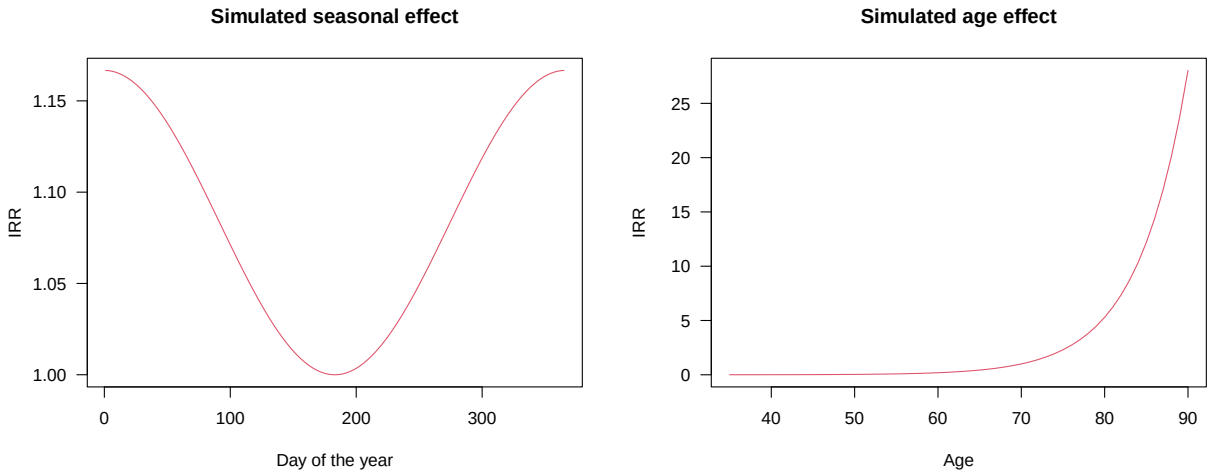
```
date <- seq(dstart, dend, by=1)
times <- seq(length(date))
month <- month(date)
doy <- yday(date)
dob <- sample(seq(dstart-round(100*365.25), dstart-round(35*365.25), by=1), n)
```

These variables are used for simulating the temporal variation in the underlying risk of AMI, with a cyclic seasonal trend and a long-term change by age modelled by a cosine function and polynomials, respectively. These effects are defined as a incident rate ratio (IRR), and created by the following code:

```
frrseas <- function(doy) (cos(doy*2*pi / 366) + 1) / 12 + 1
frrage <- function(age) exp((age - 70) / 6)
```

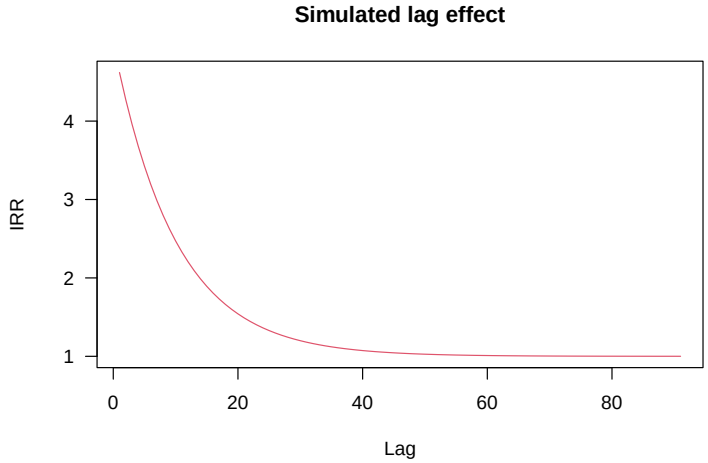
These temporal variations in risk along day of the year and age are represented in the graphs below:

```
plot(1:365, frrseas(1:365), type="l", col=2, ylab="IRR", xlab="Day of the year",
  main="Simulated seasonal effect")
plot(35:90, frrage(35:90), type="l", col=2, ylab="IRR", xlab="Age",
  main="Simulated age effect")
```



Now we create a function to define the IRR along the *lag dimension*. In this case, this dimension represents the risk after a flu episode, with the lag unit defined by day. Similarly, we illustrate the phenomenon graphically:

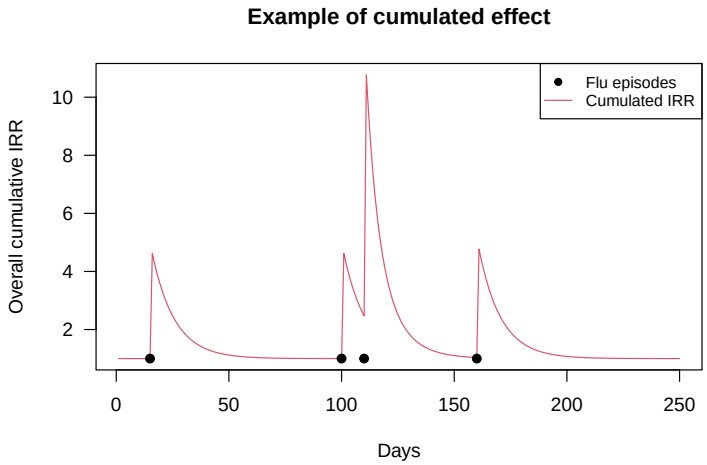
```
frrlag <- function(lag) exp(-(lag/10)) * 4 + 1
plot(1:91, frrlag(1:91), type="l", col=2, ylab="IRR", xlab="Lag",
  main="Simulated lag effect")
```



The graph indicates that, within a lag period of 3 months (1 to 91 days of lag) as in the original analyses, the risk is much increased in the first days after the flu episode, but then it attenuates and tends to null after approximately one month.

In the presence of multiple exposure episodes, lagged effects can cumulate in time, depending on the *exposure profile* of an individual. In this case, the risk at a given day is determined by the *exposure history* to flu, with potentially multiple flu episodes contributing at different lags for the same day. For instance, the code below shows an example with the risk associated with four flu episodes in a 250-day period, with the cumulated risk being the product of lag-specific contributions:

```r
expprof <- as.numeric(seq(250) %in% c(15,100,110,160))
exphist <- exphist(expprof, lag=c(1,91), fill=0)
rrflu <- apply(exphist, 1, function(x) prod(frrlag(1:91)[x==1]))
plot(seq(250), rrflu, type="l", col=2, ylab="Overall cumulative IRR", xlab="Days",
  main="Example of cumulated effect")
points(c(15,100,110,160), rep(1,4), pch=19)
legend("topright", c("Flu episodes","Cumulated IRR"), pch=c(19,NA), lty=c(NA,1),
  col=1:2, cex=0.8)
```



We have now all the information required for simulating the original data. These will consist of individual records with the following variables, with age measured in days:

- `id`: the identifier of the subject
- `dob`: date of birth
- `start`: the age of the subject at the start of follow-up
- `end`: the age of the subject at the end of follow-up
- `event`: the age of the subject at the occurrence of the AMI event
- `flu*`: multiple variables defining the age(s) of the subject at each flu episode

The data are simulated by looping in a list, producing the observations for each subject, and then binding them in a dataframe. Each of the blocks of code in the loop performs the following steps for each subject:

1. Sample the number of flu episode(s); define the risk of having a flu episode in each day; sample the flu episodes and create an exposure profile
2. Create the exposure history of flu for each day for a given lag period; compute the overall cumulative AMI risk due to flu for each day
3. Define the total AMI risk for each day, dependent on age, season, and flu; sample the unique AMI event
4. Put the information together in a dataframe; add the flu episodes, setting them to NA if less than the sampled maximum of 10

Here is the R code (it takes less than a minute):

```
dlist <- lapply(seq(n), function(i) {

  nflu <- rpois(1,1) + 1
  expprof <- drop(rmultinom(1, nflu, frrseas(doy))) > 0 + 0

  exphist <- exphist(expprof, lag=c(1,91), fill=0)
  rrflu <- apply(exphist, 1, function(x) prod(frrlag(1:91)[x==1]))

  rrtot <- frrage(as.numeric((date-dob[i])/365.25)) * frrseas(doy) * rrflu
  devent <-  date[drop(rmultinom(1, 1, rrtot))==1]

  data <- data.frame(id = paste0("sub", sprintf("%03d", i)), dob = dob[i],
    start = as.numeric(dstart - dob[i]), end = as.numeric(dend - dob[i]),
    event = as.numeric(devent - dob[i]))
  flu <- as.numeric(date[expprof == 1] - dob[i])
  for(j in seq(10)) data[paste0("flu", j)] <- if(j>nflu) NA else flu[j]

  return(data)
})
dataorig <- do.call(rbind, dlist)
```

Specifically, the total number of flu episodes `nflu` are sampled from a Poisson distribution with mean of 1 (plus one to ensure at least one episode). The occurrence of these flu episodes is sampled at random from a multinomial distribution, with probabilities varying by day of the year, thus determining a confounding effect by season. The AMI event for each subject in `devent` is then sampled from a multinomial distribution, with risk for each day `rrtot` defined by flu episodes (with lag), age, and season. Note that that in `rmultinom` probabilities are determined from IRRs by rescaling them internally.

The final line of code binds together all the records. This dataset has a simple form with one record per subject, but it contains all the information for conducting the case time series analysis in the next sections.

Data expansion

Now that we have the data, we can start our analysis using the case time series design. The first step is to expand the data to recover the individual series. You can appreciate that this leads back to the same data structure used to simulate the original dataset in the section above. We start by showing the process for a given subject (number 3), with data:

```
(sub <- dataorig[3,])
```

```
##       id        dob start   end event  flu1  flu2  flu3 flu4 flu5 flu6 flu7
## 3 sub003 1916-08-18 33008 33372 33274 33105 33273 33276   NA   NA   NA   NA
##   flu8 flu9 flu10
## 3   NA   NA    NA
```

Specifically, we reconstruct the daily series of outcome `y` (AMI event), `flu` (flu indicator), and all the time variables, including them in a new dataframe:

```
date <- as.Date(sub$start:sub$end, origin=sub$dob)
datasub <- data.frame(
  id = sub$id,
```

```
  date = date,
  times = seq(length(date)),
  age = as.numeric(date-sub$dob)/365.25,
  y = as.numeric(date-sub$dob) %in% sub$event + 0,
  flu = as.numeric(date-sub$dob) %in% na.omit(as.numeric(sub[6:15])) + 0,
  month = month(date),
  doy = yday(date)
)
```

These expanded data correspond to an individual series of outcome and predictors (therefore the name *case time series* for this design). We can have a look at the first observations for subject 3:

```
head(datasub)
```

```
##       id       date times      age y flu month doy
## 1 sub003 2007-01-01     1 90.37098 0   0     1   1
## 2 sub003 2007-01-02     2 90.37372 0   0     1   2
## 3 sub003 2007-01-03     3 90.37645 0   0     1   3
## 4 sub003 2007-01-04     4 90.37919 0   0     1   4
## 5 sub003 2007-01-05     5 90.38193 0   0     1   5
## 6 sub003 2007-01-06     6 90.38467 0   0     1   6
```

In addition, we create the exposure history for each observation within the follow-up period of the same subject, applying the function `exphist()` on the exposure series over the lag period 1-91:

```
exphistsub <- exphist(datasub$flu, lag=c(1,91), fill=0)
```

You can notice from above that subject 3 had the first flu episode at day 33,105, corresponding to time 98 of the series. We can check the exposure history matrix around those times:

```
timeflu1 <- sub$flu1-sub$start+1
exphistsub[timeflu1 + 0:5, 1:10]
```

```
##     lag1 lag2 lag3 lag4 lag5 lag6 lag7 lag8 lag9 lag10
## 98     0    0    0    0    0    0    0    0    0     0
## 99     1    0    0    0    0    0    0    0    0     0
## 100    0    1    0    0    0    0    0    0    0     0
## 101    0    0    1    0    0    0    0    0    0     0
## 102    0    0    0    1    0    0    0    0    0     0
## 103    0    0    0    0    1    0    0    0    0     0
```

The diagonal pattern of 1's identifies days corresponding to lags after this specific exposure episode.

We can now apply this expansion to all the subjects by repeating the steps above, and obtain the final data, including the exposure histories. Here is the code (it takes less than a minute):

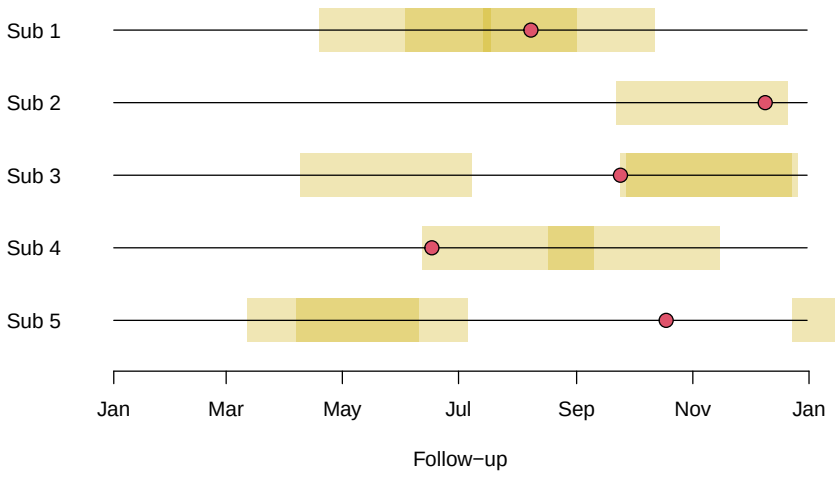
```
dlist <- lapply(seq(n), function(i) {

  sub <- dataorig[i,]

  date <- as.Date(sub$start:sub$end, origin=sub$dob)
```

```
  data <- data.frame(
    id = sub$id,
    date = date,
    times = seq(length(date)),
    age = as.numeric(date-sub$dob)/365.25,
    y = as.numeric(date-sub$dob) %in% sub$event + 0,
    flu = as.numeric(date-sub$dob) %in% na.omit(as.numeric(sub[6:15])) + 0,
    month = month(date),
    doy = yday(date)
  )

  exphist <- exphist(data$flu, lag=c(1,91), fill=0)

  return(data.table(cbind(data, exphist)))
})
data <- do.call(rbind, dlist)
```

Analysis

Now that we have obtained the final dataset, we can start the data analysis. First, we have a look at the follow-up of the first five subjects, including the time of AMI event (red circle) and exposure periods in the 1-91 days after flu episodes:

```
plot(unique(data$date), unique(data$date), ylim=c(0.5,5+0.5), yaxt="n",
  ylab="", xlab="Follow-up", frame.plot=F)
axis(2, at=5:1, labels=paste("Sub",1:5), lwd=0, las=1)
for(i in 5:1) {
  sub <- subset(data, id==unique(data$id)[i])
  flu <- sub$date[sub$flu==1]
  rect(flu+1, rep(i-0.3,length(flu)), flu+91, rep(i+0.3,length(flu)), border=NA,
    col=alpha("gold3",0.3))
  lines(sub$date, rep(i, nrow(sub)))
  points(sub$date[sub$y==1], i, pch=21, bg=2, cex=1.5)
}
```



While many of the subjects only have a single flu episode, four of them in this sample have multiple ones, and these episodes are so close that they generate overlapping exposure windows. This could not be dealt with in the original self-controlled case series analysis (Warren-Gash et al. 2012), while as shown below the case time series design can appropriately account for cumulative effects.

Now, we replicate the main case time series analysis illustrated in the original article (Gasparrini 2021). We first derive the terms to control for age and season using natural cubic and cyclic splines, respectively. We use the wrapper function `onebasis()` that simplifies the prediction and plotting of these associations, to be performed later. We call the function `pbs` from the package with the same name to generate the basis transformations for the cyclic splines. The code:

```
splage <- onebasis(data$age, "ns", knots=quantile(data$age, c(1,3)*0.25))
splseas <- onebasis(data$doy, "pbs", df=3)
```

Then, we implement the distributed lag model (DLM), defining the cross-basis parameterisation for flu with a lag period from 1 to 91 days, using the exposure histories included as lagged terms in `data`:

```
exphist <- data[,-c(1:8)]
cbspl <- crossbasis(exphist, lag=c(1,91), argvar=list("strata",breaks=0.5),
  arglag=list("ns",knots=c(3,10,29)))
```

The function `crossbasis()` internally calls `strata()` with cut-off at 0.5 to parameterise the exposure-response for flu using a simple indicator, and `ns()` to produce the natural cubic splines with specific knots for the lag-response function (see `help(crossbasis)`).

We now have all the terms for fitting the fixed-effects Poisson regression using the function `gnm()`. The regression model includes all the predictors, and defines the conditional stratification through the argument `eliminate`. This is the code:

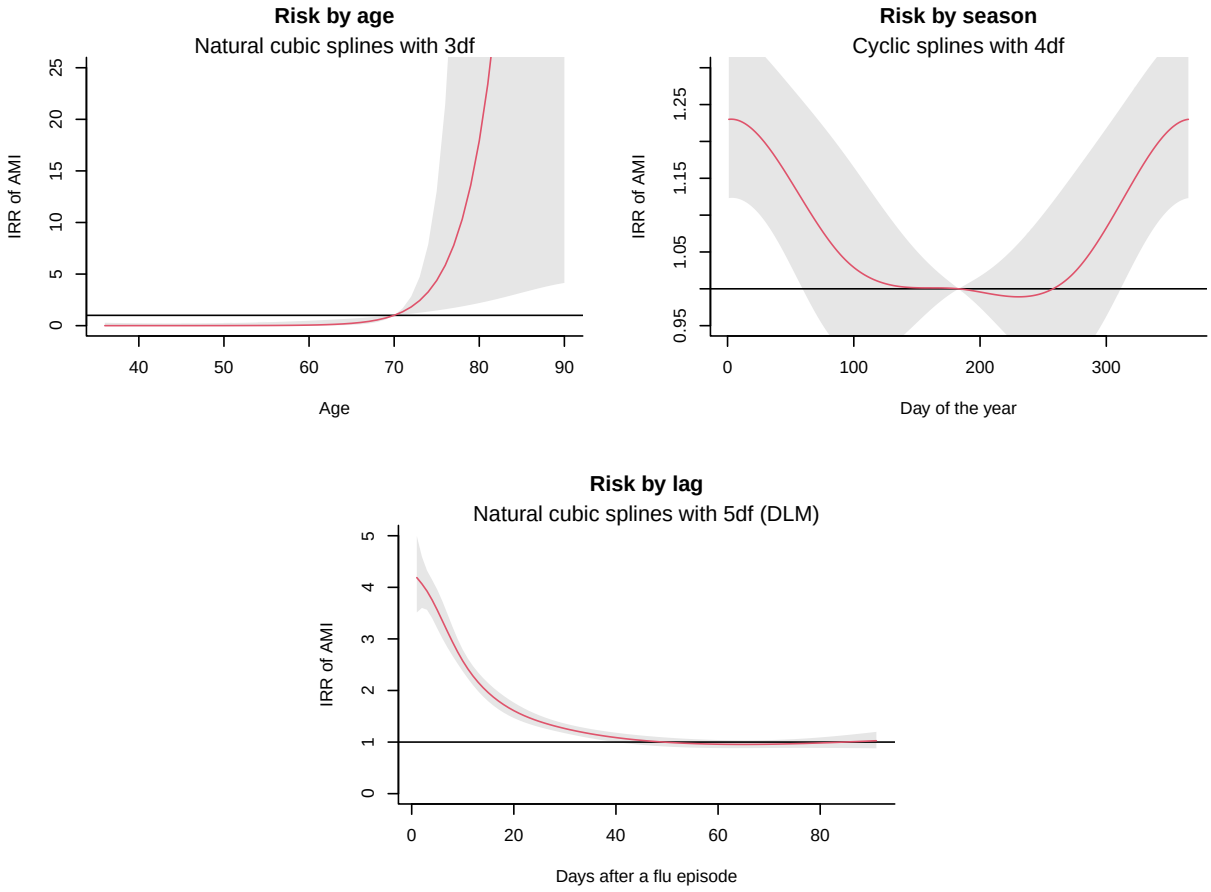
```
mspl <- gnm(y ~ cbspl+splage+splseas, data=data, family=poisson,
  eliminate=factor(id))
```

The estimated coefficients and associated (co)variance matrix of the model can now be used to predict the association of the various terms with the risk of AMI, using the function `crosspred()`:

```
cpspl <- crosspred(cbspl, mspl, at=1)
cpsplage <- crosspred(splage, mspl, cen=70, to=90)
cpsplseas <- crosspred(splseas, mspl, cen=366/2, at=1:365)
```

Finally, we can plot them:

```
plot(cpsplage, col=2, ylab="IRR of AMI", xlab="Age", main="Risk by age",
  ylim=c(0,25))
mtext("Natural cubic splines with 3df", cex=0.8)
plot(cpsplseas, col=2, ylab="IRR of AMI", xlab="Day of the year",
  main="Risk by season", ylim=c(0.95,1.30))
mtext("Cyclic splines with 4df", cex=0.8)
plot(cpspl, var=1, col=2, ylab="IRR of AMI", xlab="Days after a flu episode",
  ylim=c(0,5), main="Risk by lag")
mtext("Natural cubic splines with 5df (DLM)", cex=0.8)
```



Results are similar to those reported in the original article (Gasparrini 2021). Differences in estimated confidence intervals for the risk along with age and after a flu episodes are easily explained by the shorter follow-up period simulated here (one year), which reduces the within-subject age differences and increases the prevalence of exposure to flu.

An alternative parameterisation of cross-basis term can be used, specifically using strata functions to represent the risk along lags. The code:

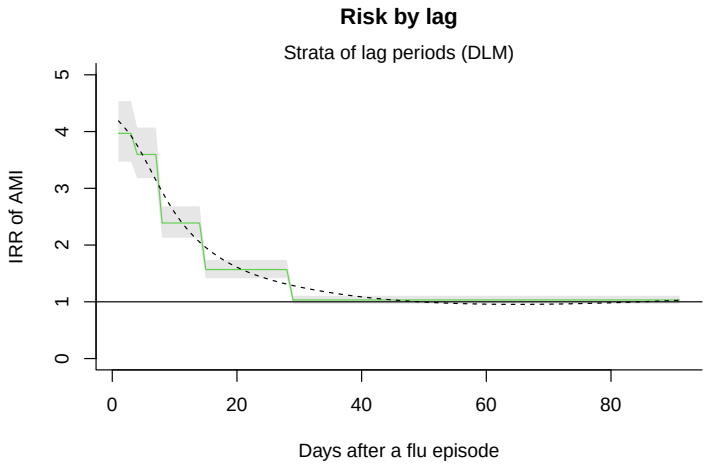
```
cbstr <- crossbasis(exphist, lag=c(1,91), argvar=list("strata",breaks=0.5),
  arglag=list("strata",breaks=c(4,8,15,29)))
```

We can now fit the alternative model:

```
mstr <- gnm(y ~ cbstr+splage+splseas, data=data, family=poisson,
  eliminate=factor(id))
cpstr <- crosspred(cbstr, mstr, at=1)
```

and create the related plot, including the previous fitted relationships as dashed lines:

```
plot(cpstr, var=1, col=3, ylab="IRR of AMI", xlab="Days after a flu episode",
  ylim=c(0,5), main="Risk by lag")
mtext("Strata of lag periods (DLM)", cex=1)
lines(cpspl, var=1, lty=2)
```



This parameterisation can be compared to the original analysis performed with the self-controlled case series design (Warren-Gash et al. 2012), where post-exposure periods consistent with the lag strata were used. However, here the case time series design and DLMs can appropriately handle potentially overlapping exposure periods. We can extract from the predictions the IRR (and 95% confidence intervals) corresponding to the five strata:

```
resstr <- round(with(cpstr, t(rbind(matRRfit,matRRlow,matRRhigh))), 2)
colnames(resstr) <- c("IRR","low95%CI","high95%CI")
resstr[paste0("lag", c(1,4,8,15,29)),]
```

```
##        IRR low95%CI high95%CI
## lag1  3.97     3.47      4.54
## lag4  3.60     3.18      4.07
## lag8  2.39     2.13      2.68
## lag15 1.57     1.42      1.74
## lag29 1.03     0.97      1.11
```

The results demonstrate the flexibility of the case time series design to investigate complex relationships using self-matched comparisons of individual-level data.

References

Gasparrini, A. 2021. “The case time series design.” *Epidemiology*, Accepted for publication.

Warren-Gash, Charlotte, Andrew C Hayward, Harry Hemingway, Spiros Denaxas, Sara L Thomas, Adam D Timmis, Heather Whitaker, and Liam Smeeth. 2012. “Influenza infection and risk of acute myocardial infarction in England and Wales: a CALIBER self-controlled case series study.” *Journal of Infectious Diseases* 2006 (11): 1652–59.

The case time series design – eAppendix 2

A case study for applications in environmental epidemiology

Antonio Gasparrini

09 August 2021

Contents

This document was originally presented as eAppendix 2 of the article "*The case time series design*," accepted for publication in Epidemiology (Gasparrini 2021), and it reproduces the analysis presented as the second case study. An updated version of this document and related material are available at the GitHub page and at the personal website of the author. The material includes the Rmarkdown files to compile the document, plus scripts with the embedded R code. Note that the code is profiled for clarity, not for speed, with the aim of illustrating the steps of the analysis and the features of the design. It can (probably should) be modified when re-used in real analyses.

This case study illustrates the application of the case time series design in environmental studies. Specifically, the example describes an analysis of the association between exposure to three different environmental stressors and the risk of respiratory symptoms using a cohort of participants to a smartphone study. The sample includes 1,601 subjects who reported daily the occurrence of respiratory symptoms such as asthma and allergic rhinitis in a smartphone app, and who were assigned exposure levels by linking their geo-located position with high-resolution spatio-temporal maps of pollen, air pollution, and temperature. The analysis illustrates an application of the case time series design with a binary outcome and multiple continuous exposures. The data were collected within the AirRater study, an integrated online platform that combines symptom surveillance, environmental monitoring, and real-time notifications operating in Tasmania (Johnston et al. 2018). The code shown below creates and uses simulated data to reproduce the features of the original dataset, which cannot be made publicly available, and the steps and (approximate) results of the application of the case time series design.

Preparation

The following packages are loaded in the session, and need to be installed to run the R code:

```
library(dlnm) ; library(gnm) ; library(data.table) ; library(splines)
```

We first set a seed to ensure the exact replicability of the results, as the code includes expressions with random number generation, and we also set the graphical parameter `las` for the plots:

```
set.seed(13041975)
par(las=1)
```

Simulating the original data

The data used in this case study are simulated directly in this section. The user can skip it if not of interest, and start with the following section for the data analysis. First, we set the parameters, namely the number of subjects `n` and the date of start and end of study period. Then we create a `date` and related time variables `year`, `month`, `doy` (day of the year), and `dow` (day of the week):

```
n <- 1601
dstart <- as.Date("2015-10-29")
dend <- as.Date("2018-11-19")
date <- seq(dstart, dend, by=1)
year <- year(date)
month <- month(date)
doy <- yday(date)
dow <- factor(wday(date))
```

Then we define follow-up periods for the 1601 subjects, randomly sampling starting dates and length of follow-up, with the constraints that the end of follow-up cannot be later than the end of the study period, and with a length of at least 10 days. The code:

```
fustart <- sample(seq(dstart, dend-10, by=1), n, replace=TRUE)
fuend <- fustart + pmax(pmin(round(exp(rnorm(n, 5.1, 2))), dend-fustart), 10)
sum(fuend-fustart+1)
```

```
## Time difference of 363901 days
```

While the follow-up distribution does not match perfectly the original study, the sampling parameters are set above to generate approximately the same number of total person-days, in this case, 363,901.

Finally, we define some variables used to simulate the distribution of the environmental exposures and, later, the seasonal baseline risk. These variables are the cosine transformation of `doy` and quadratic splines of `date` with 5 degrees of freedom per year. In addition, we simulate 20 random smoke days occurring in the (Australian) summer. The code:

```
cosdoy <- cos(doy*2*pi / 366)
spldate <- bs(date, degree=2, int=TRUE, df=round(length(date)/365.25)*5)
smokeday <- date %in% sample(date[month %in% c(1,2,12)], 20)
```

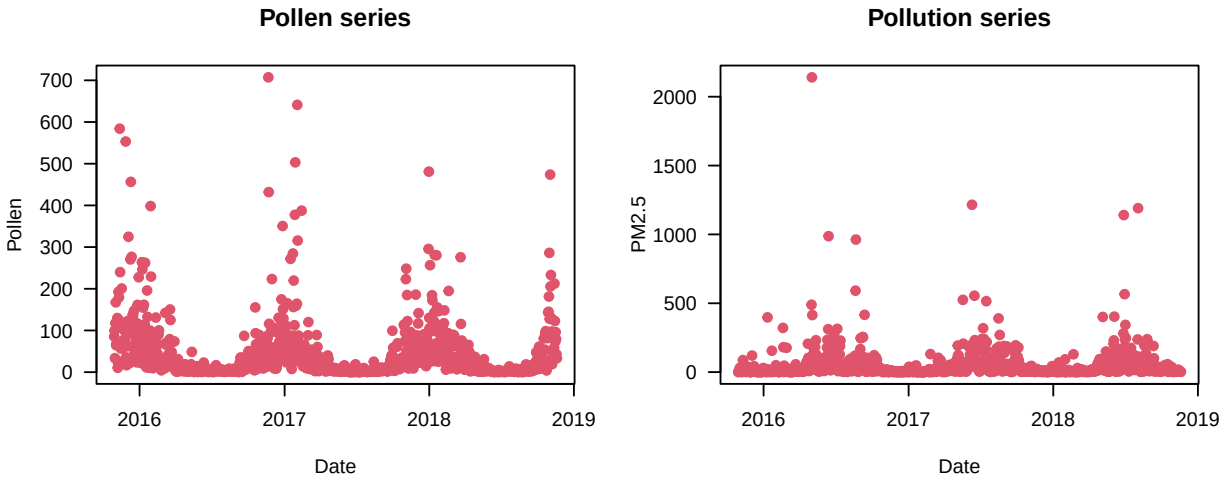
We are now ready to simulate the distribution of the three environmental stressors. In the original study, individual exposure series were reconstructed through the geo-location system of the smartphone by linkage with detailed spatio-temporal exposure maps. In order to simplify the simulation process, we derive here a single series for each stressor, assuming that all the 1601 subjects are exposed to the same levels on the same day. This does not affect the generality of the example, and in real-case settings, individual-level exposure series can nevertheless be used.

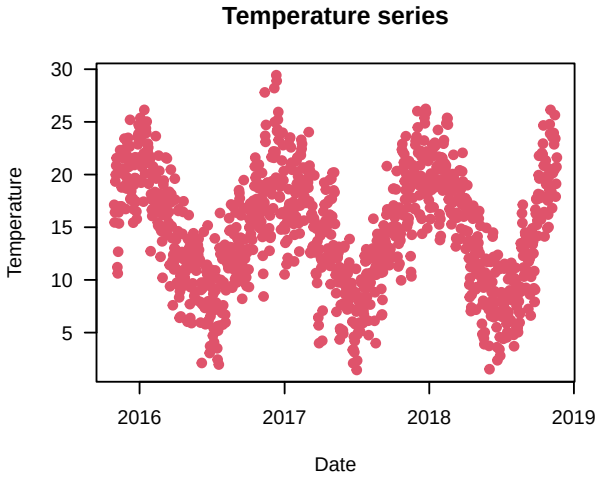
The environmental exposures are created by assuming an underlying seasonal trend, represented by the cosine variable above, plus auto-correlated random normal deviations. Exponentiation is used to produce non-negative values of pollen (grains/m^3) and pollution ($PM_{2.5}$, μgr/m^3), while temperature ($^\circ$C) is sampled directly. The code:

```
pollen <- exp(cosdoy*2+2.5 + arima.sim(list(ar=0.5), length(date), sd=0.8))
pm <- exp((-cosdoy)*1.6+2.5 + smokeday*3.2 +
  arima.sim(list(ar=0.6), length(cosdoy), sd=0.95))
tmean <- cosdoy*6+15 + arima.sim(list(ar=0.6), length(cosdoy), sd=2.6)
envdata <- data.frame(date, pollen, pm, tmean)
```

The variables are included in the dataframe `envdata`. The definitions above provide a realistic distribution of the three exposures, with pollen and temperature peaking in summer, while $PM_{2.5}$ shows higher wintertime levels but with isolated spikes in the summer corresponding to smoke days due to fires. A visual representation is offered by the plots obtained through:

```
plot(date, pollen, xlab="Date", ylab="Pollen", main="Pollen series", col=2,
  pch=19)
plot(date, pm, xlab="Date", ylab="PM2.5", main="Pollution series", col=2,
  pch=19)
plot(date, tmean, xlab="Date", ylab="Temperature", main="Temperature series",
  col=2, pch=19)
```



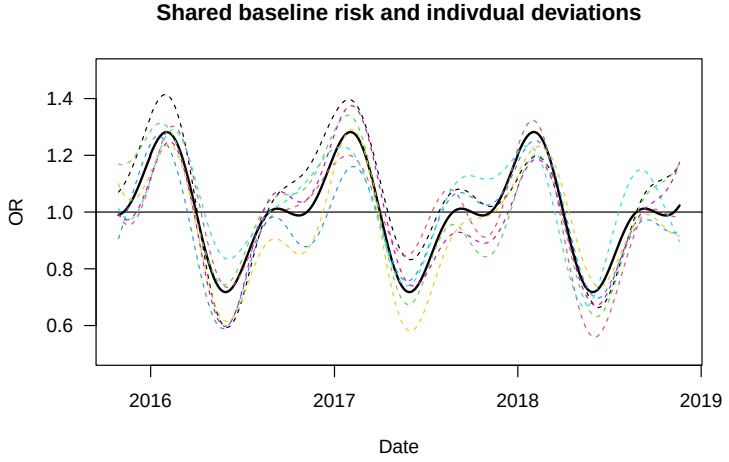


The variables created above can now be used to define individual risk profiles of experiencing allergic symptoms. These profiles will be simulated as risks associated to the three exposures on top of baseline trends. We first simulate the latter as a combination of shared underlying risks and individual-level deviations:

```
fortrend <- function(ind=TRUE) (cosdoy*1.6 + sin(doy*4*pi/366))/8+1 + if(ind)
  spldate %*% runif(ncol(spldate),-0.2,0.2) else 0
```

The function `fortrend()` includes harmonic terms at different periods to define the shared baseline risk common to all subjects, plus optionally individual deviations modelled using random coefficients for the spline of time. These trends are defined as odds ratio (OR). We can graphically represent them using the code below, with the bold black line representing the shared trend and the dashed coloured lines as individual profiles:

```
plot(date, fortrend(ind=F), type="l", lwd=2, ylim=c(0.5,1.5), xlab="Date",
  ylab="OR", main="Shared baseline risk and indivdual deviations")
abline(h=1)
for(i in 1:7) lines(date, fortrend(ind=T), type="l", lty=2, col=i)
```



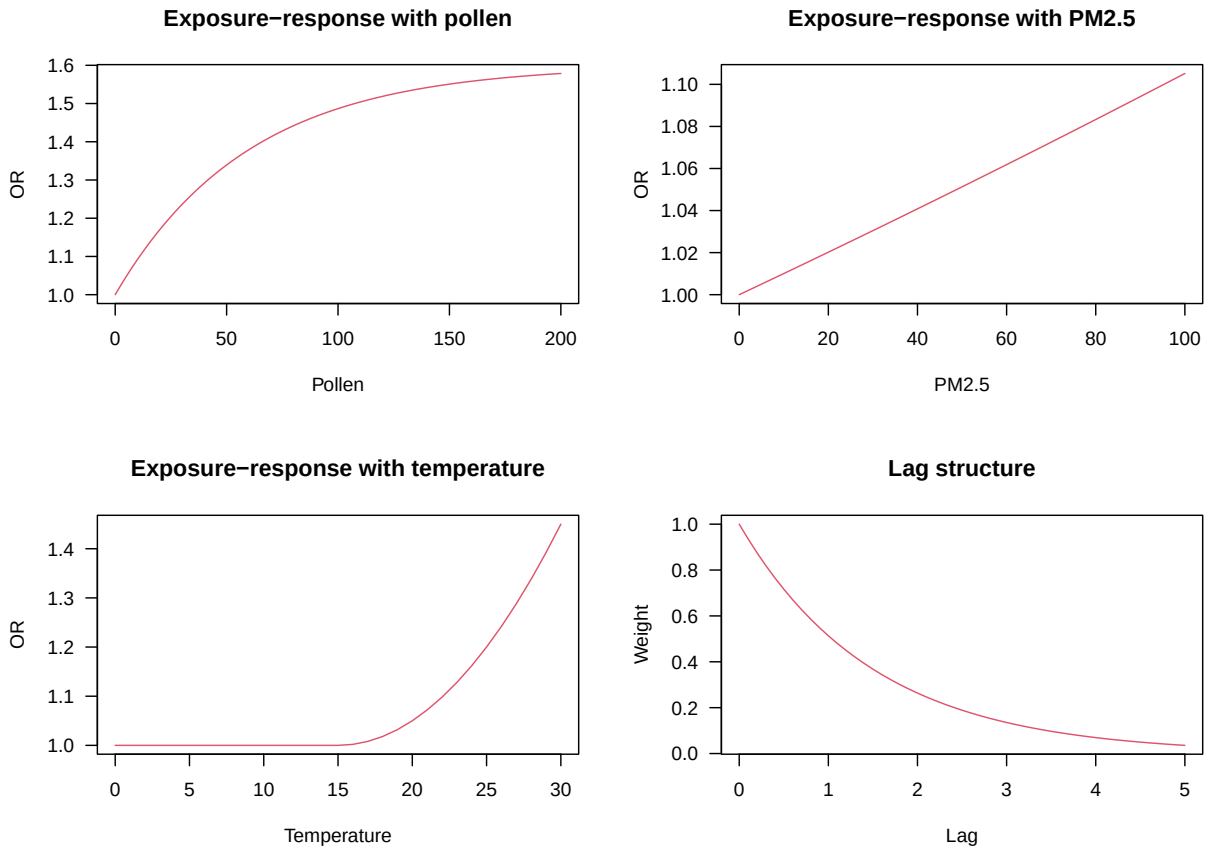
The next step is the definition of the increase in risk due to exposure to the three environmental stressors. Specifically, we define non-linear relationships for pollen and temperature, with effects lagged up to 3 days, and a linear and unlagged association with $PM_{2.5}$. First, we define the three functions to specify the three exposure-response risk shapes and the lag structure:

```
forpoll <- function(x) 1.6 - 0.6*exp(-x/60)
forpm <- function(x) exp(x/1000)
fortmean <- function(x) 1 + ifelse(x>15, 0.002*(x-15)^2, 0)
fwlag <- function(lag) exp(-lag/1.5)
```

These functions define relationships in the OR and lag scales, and can be represented graphically with:

```
plot(0:200, forpoll(0:200), type="l", xlab="Pollen", ylab="OR",
  main="Exposure-response with pollen", col=2)
plot(0:100, forpm(0:100), type="l", xlab="PM2.5", ylab="OR",
  main="Exposure-response with PM2.5", col=2)
plot(0:30, fortmean(0:30), type="l", xlab="Temperature", ylab="OR",
```

```
  main="Exposure-response with temperature", col=2)
plot(0:50/10, fwlag(0:50/10), type="l", xlab="Lag", ylab="Weight",
  main="Lag structure", col=2)
```



These shapes are similar to the associations estimated in the original study (Gasparrini 2021). The lag structure is defined as weights, and can be used to represent a decreasing OR proportionally to time after the exposure occurred. As an example, we used the functions above to calculate the net OR in a given day after exposures to pollen of 50, 9, 135, and 93 grains/m^3 in the same and past 3 days (lag 0–3):

```
exp(sum(log(forpoll(c(50,9, 135, 93))) * fwlag(0:3)))
```

```
## [1] 1.647004
```

Simply, the expression above computes the log-OR for each exposure occurrence, which are then weighted depending on the lag and then summed and exponentiated to obtain the overall cumulative OR .

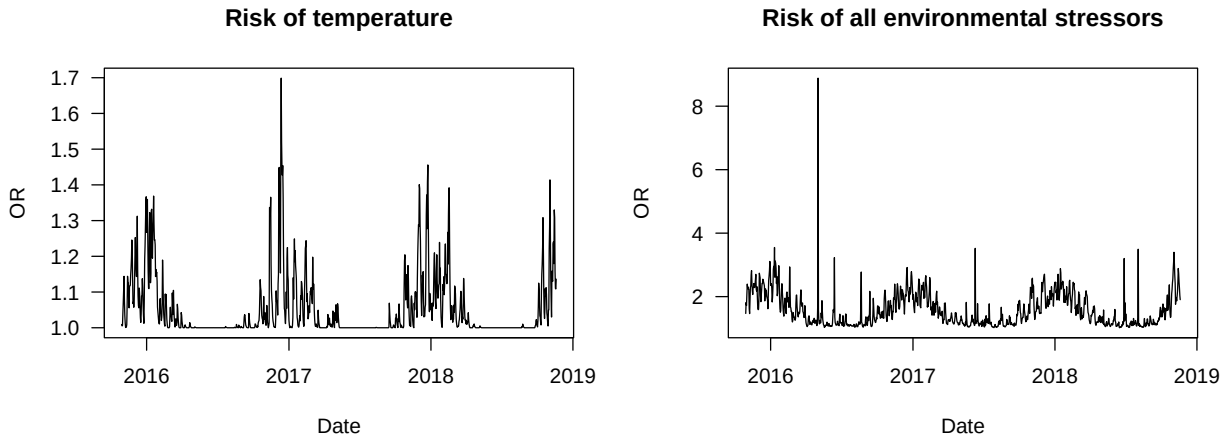
We can now apply the same computation to the whole series for the three exposures, using first the function `exphist` to generate the matrix of lagged exposures, and then applying the expression for each row:

```
orpoll <- apply(exphist(pollen, lag=3), 1, function(x)
  exp(sum(log(forpoll(x)) * fwlag(0:3))))
orpm <- forpm(pm)
```

```
ortmean <- apply(exphist(tmean, lag=3), 1, function(x)
  exp(sum(log(fortmean(x)) * fwlag(0:3))))
orenv <-  orpoll * orpm * ortmean
```

Note that we assume a lag 0–3 for pollen and temperature, while we simply define a same-day association with no lag for $PM_{2.5}$. The code above computes therefore the OR contribution for each exposure in each day, which are then multiplied to obtain the overall risk associated with all the three environmental stressors in the vector `orenv`. The series for temperature and all the exposures are graphically represented below:

```
plot(date, ortmean, type="l", xlab="Date", ylab="OR",
  main="Risk of temperature")
plot(date, orenv, type="l", xlab="Date", ylab="OR",
  main="Risk of all environmental stressors")
```



We have now all the information required for simulating the original data. These are created by looping in a list, producing the observations for each subject, and then binding them in a dataframe. Each of the blocks of code in the loop performs the following steps for each subject:

1. Define the follow-up period and identify the related subset of the study period
2. Create the total risk contribution in each follow-up day
3. Sample the occurrence of respiratory symptoms within the follow-up period
4. Put the information together in a dataframe, adding the series of environmental exposures

Here is the R code:

```
dlist <- lapply(seq(n), function(i) {

  fudate <- seq(fustart[i], fuend[i], by=1)
  sub <- date %in% fudate

  ortot <- fortrend(ind=T)[sub] * orenv[sub] * (1 + wday(fudate) %in% c(2:6)*0.4)

  pbase <- plogis(-3.3 + 14/length(fudate) - 0.0015*length(fudate))
  sympt <- rbinom(sum(sub), 1, plogis(qlogis(pbase) + log(ortot)))
```

```
  data <- cbind(data.frame(id=paste0("sub",sprintf("%04d", i)), date=fudate,
    year=year[sub], month=month[sub], dow=dow[sub], y=sympt), envdata[sub, -1])

  return(data.table(data))
})
data <- do.call(rbind, dlist)
```

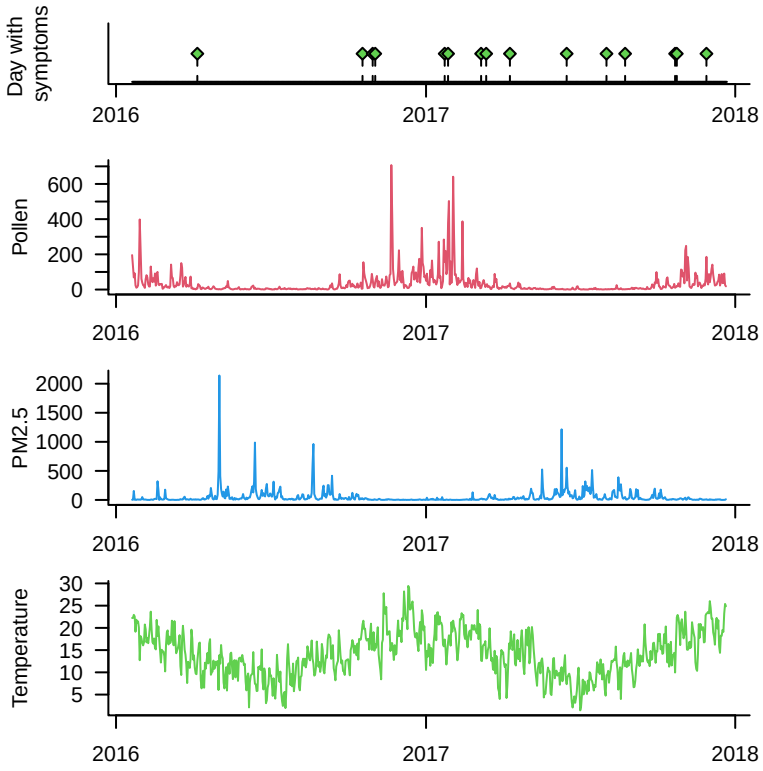
Specifically, the total OR in `ortot` is the product of underlying trends (as the sum of shared seasonal OR plus random individual deviations), the contribution of environmental factors, and a simulated OR of 1.4 for weekdays vs weekends. These are multiplied to a baseline risk in `base` to compute the day and subject-specific odds. The baseline risk varies across individuals, and it is inversely proportional to the follow up period, similary to the real-date example. The indicator of days with respiratory symptoms `sympt` is sampled then from a Bernoulli distribution (binomial with a single trial) with probabilities back-transformed from the logistic scale. Note that this method of sampling does not ensure that all the subjects have at least one day with reported symptoms, and these will be automatically discarded from the analysis and they do not contribute information to the conditional comparison.

The final line of code binds together all the data in a single dataframe. This dataset is already expanded to its case time series format, where the number of rows corresponds to the total person-days of follow-up (363,901). In some situations, it can be more convenient to store the data in multiple datasets, for instance separating individual information and environmental exposures, and then assemble them together for the final analysis.

Analysis

Now that we have obtained the final dataset, we can replicate the main case time series analysis. First, we have a look at the data for a given subject (identified as sub0036), represented as individual series of daily observations of outcome and predictors (therefore the name *case time series* for this design). The code:

```
dsub <- subset(data, id=="sub0036")
plot(y~date, data=dsub, type="h", lty=2, ylim=c(0,2), yaxt="n", bty="l", xlab="",
  ylab="Day with \nsymptoms", mgp=c(2.2,0.7,0), lab=c(5,3,7))
points(y~date, data=subset(dsub,y>=1), pch=23, bg=3)
plot(pollen~date, data=dsub, type="l", lty=1, bty="l", col=2, xlab="",
  ylab="Pollen")
plot(pm~date, data=dsub, type="l", lty=1, bty="l", col=4, xlab="Date",
  ylab="PM2.5")
plot(tmean~date, data=dsub, type="l", lty=1, bty="l", col=3, xlab="Date",
  ylab="Temperature")
```



We can now define the different terms to be included in the regression model. First, we define a set of splines of time with approximately 8 degrees of freedom per year, and subject/year/month strata indicators, to be used to model the shared seasonal trend and individual deviations, respectively.:

```
dftrend <- round(as.numeric(diff(range(data$date))/365.25 * 8))
btrend <- ns(data$date, knots=equalknots(data$date, dftrend-1))
data$stratum <- with(data, factor(paste(id, year, month, sep="-")))
```

We now apply the function `crossbasis()` to parameterise distributed lag linear and non-linear transformations of the environmental variables:

```
cbpoll <- crossbasis(data$pollen, lag=3, argvar=list(knots=c(40,100)),
  arglag=list(knots=1), group=data$id)
cbpm <- crossbasis(data$pm, lag=3, arglag=list("integer"), group=data$id)
cbtmean <- crossbasis(data$tmean, lag=3, argvar=list(knots=1:2*10),
  arglag=list(knots=1), group=data$id)
```

Specifically, the default `ns()` function is used in both the `argvar` and `arglag` arguments to specify natural cubic splines for the exposure-response and lag-response, respectively, of both pollen and temperature, using different knots placements. A default linear exposure-response is defined for $PM_{2.5}$, instead, while the lag-response is parameterised through an unconstrained distributed lag function, namely using indicators for each lag. The lag period is extended to 0–3 for all three exposures. A `group` argument is used to specify that the variables do not represent a unique and complete series, but multiple individual series. See `help(crossbasis)` for more information.

We now have all the terms for fitting the fixed-effects logistic regression using the function `gnm()`, with the strata indicators included in the argument `eliminate`:

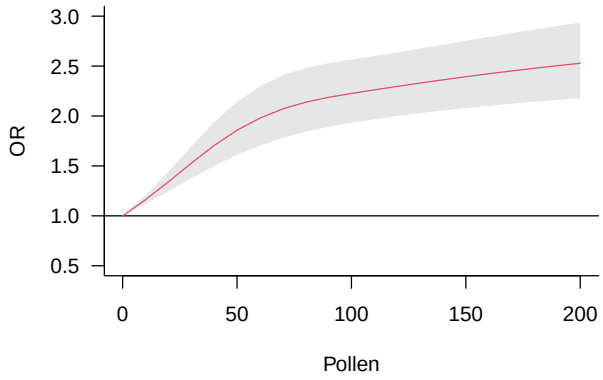
```
mod <- gnm(y ~ cbpoll + cbpm + cbtmean + btrend + dow, eliminate=stratum, data=data,
  family=binomial)
```

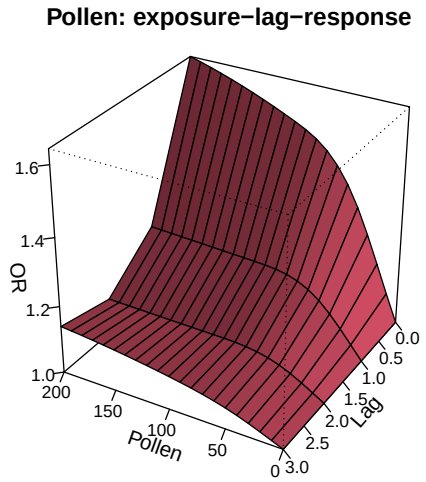
The estimated coefficients and associated (co)variance matrix of the model can now be used to predict the association of the various terms with the risk of respiratory symptoms, using the function `crosspred()`:

```
cppoll <- crosspred(cbpoll, mod, at=0:20*10, cen=0)
cppm <- crosspred(cbpm, mod, at=0:20*5, cen=0)
cptmean <- crosspred(cbtmean, mod, cen=15, by=1.5)
```

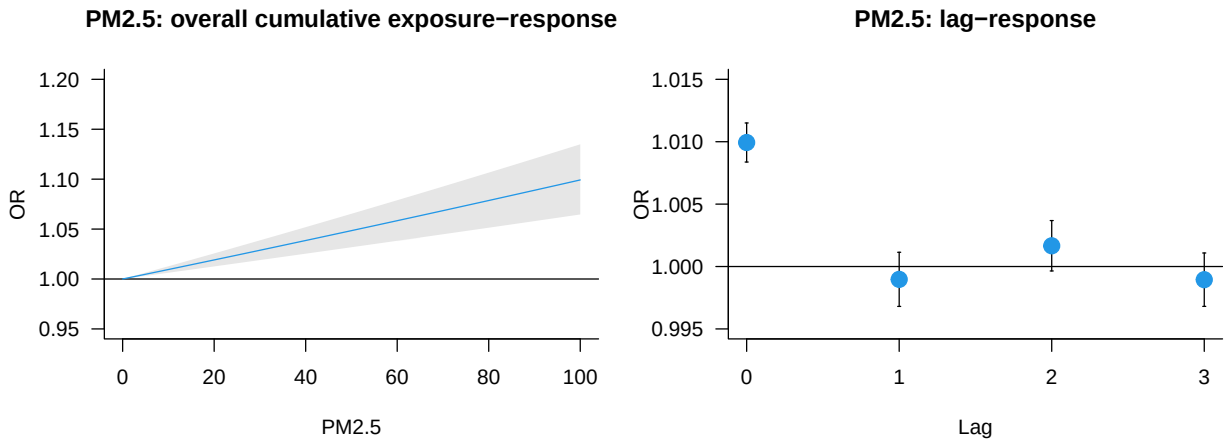
We can now represent graphically the association in both dimensions of exposure intensity and lag. Specifically, the plots below represent the overall cumulative exposure-responses (interpreted as the net associations accounting for the whole lag period), the full bi-dimensional exposure-lag-responses for non-linear relationships of pollen and temperature, and the lag-response corresponding to a 10μgr/m^3 increases in $PM_{2.5}$. The code:

```
plot(cppoll, "overall", xlab="Pollen", ylab="OR", col=2,
  main="Pollen: overall cumulative exposure-response", ylim=c(0.5,3))
plot(cppoll, xlab="Pollen", zlab="OR", main="Pollen: exposure-lag-response",
  cex.axis=0.8, col=2)
```

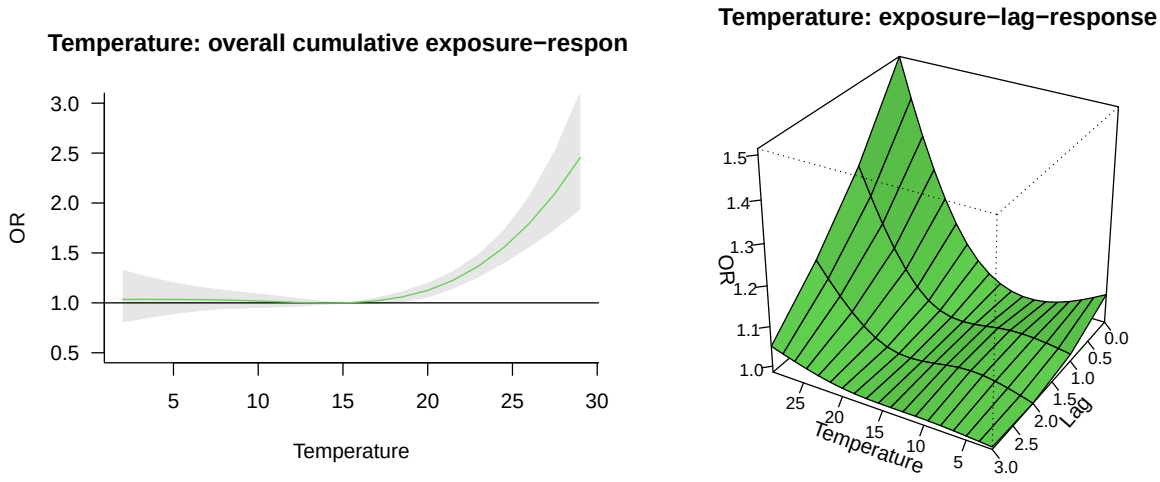





```
plot(cppm, var=10, "overall", xlab="PM2.5", ylab="OR", col=4,
  main="PM2.5: overall cumulative exposure-response", ylim=c(0.95,1.20))
plot(cppm, var=10, ci="b", type="p", ylab="OR", col=4, pch=19, cex=1.7,
  xlab="Lag", main="PM2.5: lag-response", lab=c(3,5,7), ylim=c(0.995,1.015))
```




```
plot(cptmean, "overall", xlab="Temperature", ylab="OR", col=3,
  main="Temperature: overall cumulative exposure-response", ylim=c(0.5,3))
plot(cptmean, xlab="Temperature", zlab="OR", ltheta=240, lphi=60, cex.axis=0.8,
  main="Temperature: exposure-lag-response", col=3)
```



The estimated associations are similar to those presented in the original analysis (Gasparrini 2021). The results demonstrate the flexibility of the case time series design to investigate complex relationships with multiple exposures using individual data in a complex cohort setting.

References

Gasparrini, A. 2021. “The case time series design.” *Epidemiology*, Accepted for publication.

Johnston, F H, A J Wheeler, G J Williamson, S L Campbell, P J Jones, I S Koolhof, C Lucani, N B Cooling, and D M J S Bowman. 2018. “Using smartphone technology to reduce health impacts from atmospheric environmental hazards.” *Environmental Research Letters* 13 (4): 044019.

The case time series design – eAppendix 3
A simulation study

This simulation study evaluates the inferential performance of regression models for the *case time series* design under various data-generating scenarios, through the assessment of bias, coverage of the confidence intervals, and root mean square error (RMSE) of the estimators. The study aims, first, at testing the ability of the model in recovering the true exposure-response association under increasingly complex data settings, and second, at evaluating the four key assumptions underpinning the case time series design.

All the simulated scenarios use a common setting with 500 subjects followed up for one year between 01/01/2019 and 31/12/2019. For each scenario, $m = 50{,}000$ datasets are simulated, each including (initially) $500 \cdot 365 = 182{,}500$ observations. The inference focuses on a risk summary β, whose definition is scenario-dependent. Specifically, all the simulated cases assume a risk period lasting 10 days following the exposure, which, using a time series terminology, corresponds to a *lag period* defined over days 0-10. For most scenarios, the risk summary β represents the constant effect in each day within the risk (lag) period. In contrast, Scenario 10 illustrates more complex lag structures where the effect varies within the risk period, and here β_c quantifies the net effect cumulated across lag 0-10. The performance is assessed in terms of relative bias (%), coverage, and relative RMSE (%), defined as:

$$\text{Bias} = \frac{\left|\sum_{i=1}^{m}\left(\hat{\beta}_i - \beta\right)/m\right|}{\beta_c}$$

$$\text{Coverage} = \sum_{i=1}^{m} I\left(\left|\hat{\beta}_i - \beta\right| \leq \Phi^{-1}(1-\alpha/2) \cdot \sqrt{V\left(\hat{\beta}_i\right)}\right)/m$$

$$\text{RMSE} = \frac{\sqrt{\sum_{i=1}^{m}\left(\hat{\beta}_i - \beta\right)^2/m}}{\beta}$$

where $\hat{\beta}_i$ is the estimate at each of the $i = 1, \ldots, m$ iterations, I is an indicator function, and $\Phi^{-1}(1-\alpha)$ is the quantile function of the cumulative normal distribution related to probability $1-\alpha$, with $\alpha = 0.05$.

Each scenario is described in detail in the sections below, with additional results that complement the figures reported in Table 1 of the manuscript. The R code to fully reproduce the simulations and results in each scenario is provided in the online supplemental material, with an updated version available at the personal website (http://www.ag-myresearch.com/) and GitHub webpage (https://github.com/gasparrini/) of the author.

Part I: assessment of modelling performance

The first part of the simulation study (Scenarios 1-10) applies the case time series methodology in increasingly complex data-generating settings, simulated under the four core assumptions. The expectation is that the case time series models will produce valid estimates of risk associations in all the cases. The study covers scenarios with various definitions of the outcome, exposure, underlying baseline risks, temporal associations, and both time-invariant and time-varying confounding. Specifically:

- The outcome is represented by different quantities, such as event counts, binary indicators, or continuous measures.
- The exposure, similarly, is represented either by binary indicators of episodes or by continuous measures.
- The time-varying baseline risk is optionally included, and in this case, simulated either as shared (common) trend or alternatively as subject-specific deviations from an average trend.

- Time-invariant and time-varying confounders are optionally included, and in this case, simulated as risk factors strongly correlated with the exposure.
- The temporal association is represented either as a simple constant risk period following an exposure or by more complex lag structures.

The scenarios, summarised in Table S1 below, depict combinations of the features above, from basic data setting to situations involving more complex definitions.

Table S1. Description of the simulation scenarios with combinations of the design features.

Scenario	Outcome	Exposure	Trend	Confounder	Lag structure
Scenario 1: Basic	Count	Episode	None	None	Simple
Scenario 2: Rare outcome/exposure	Count (rare)	Episode (rare)	None	None	Simple
Scenario 3: Continuous exposure	Count	Continuous	None	None	Simple
Scenario 4: Binary outcome	Binary indicator	Continuous	None	None	Simple
Scenario 5: Continuous outcome	Continuous	Continuous	None	None	Simple
Scenario 6: Common trend	Count	Continuous	Common	None	Simple
Scenario 7: Subject-specific trend	Count	Continuous	Subject-specific	None	Simple
Scenario 8: Unobserved baseline confounder	Count	Continuous	Subject-specific	Baseline	Simple
Scenario 9: Time-varying confounder	Count	Continuous	Subject-specific	Time-varying	Simple
Scenario 10: Complex lag structure	Count	Continuous	Subject-specific	Both	Complex

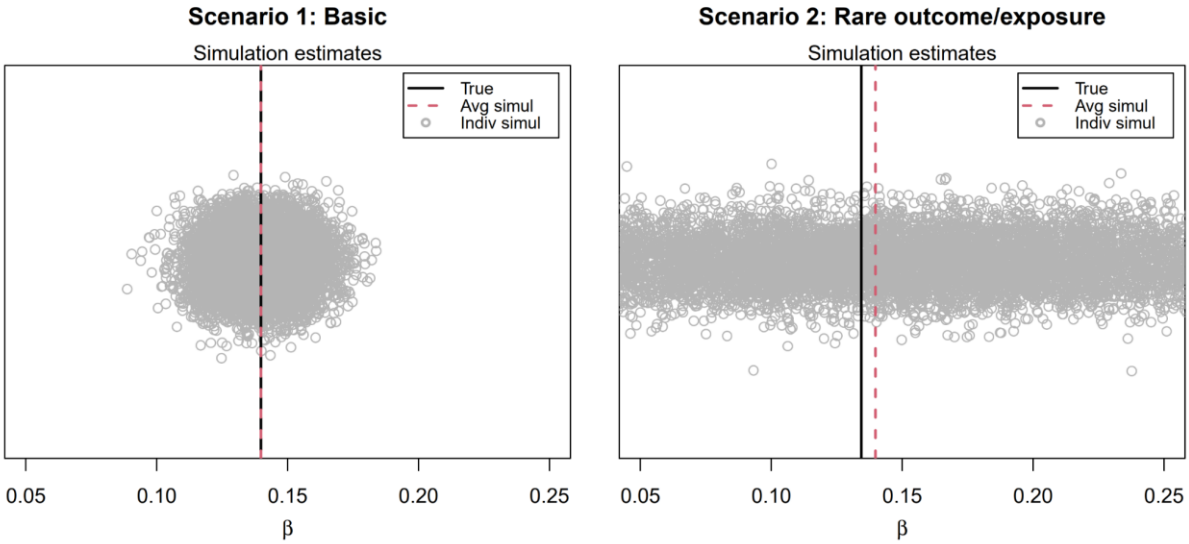
Scenario 1: Basic

In this first scenario, the exposure x is defined as a binary indicator of multiple episodes, randomly occurring in 10% of the 365 days of follow-up for each subject. Each exposure episode is associated with an increase in risk for an outcome event, which is assumed constant over the 0-10 risk period. The risk summary is represented by the relative risk (RR) of experiencing an outcome event in each of the 11 days within the risk period, with $\mathrm{RR} = \exp(\beta) = 1.15$. The outcome is represented as counts of repeated events y randomly sampled from a multinomial distribution, with the number of occurrences per subject varying randomly in the range 5-20. This feature simulates subject-specific constant baseline risks varying across the 500 subjects. No trend, either common or subject-specific, and time-varying confounders are included.

The case time series analysis is performed using a fixed-effects Poisson model, which corresponds to a conditional Poisson regression. The model includes a single term, defined by the cumulated exposure x_c, representing the sum of the exposure episodes in the same day and previous 10 days (lag 0-10). Subject-specific intercept terms ξ_i are included to model differential baseline risks. It is worth noting that, in these settings, the case time series design resembles a standard self-controlled case series (SCCS), although with the follow-up split into equally-spaced time intervals. However, the case time series data setting allows modelling multiple exposure episodes with potentially overlapping risk periods, through the computation of cumulative effects.

Results reported in the manuscript indicate no bias and perfect coverage for the estimate of the risk summary β. Figure S1 (left panel) confirms these findings, displaying the distribution of the 50,000 estimates $\hat{\beta}_i$ together with their average and the true effect β.

Figure S1. Estimates of $\log(\mathrm{RR}) = \beta$ from the case time series models applied in Scenario 1 (left panel) and Scenario 2 (right panel). The graphs report the true simulated association (black line), the average estimate of the 50,000 iterations (red line), and the estimates of the individual iterations (grey dots). A small bias in noticeable in Scenario 2. The individual estimates are scattered across the *y*-axis to show the distributions. The range of the x-axis in the right panel only includes a subset of the individual estimates.



Scenario 2: Rare outcome/exposure

Scenario 2 repeats the simulations from the previous scenario but in the case of rare outcome events and exposure episodes. Specifically, the same settings are used, but simulating only 1 to 5 exposure episodes and 1 to 3 outcome events per subject. The same fixed-effects Poisson regression model of Scenario 1 is used.

Results are reported in Figure S1 (right panel), with the distribution of the 50,000 estimates $\hat{\beta}_i$ within the same range as the previous scenarios. Note that the estimates cover a wider range when compared to Scenario 1, indicating the much lower precision due to the rare occurrence of exposure episodes and outcome events. More importantly, the plot confirms the small bias, with an underestimation of 4.5% (see Table 1 in the manuscript), which is consistent with the asymptotic bias of maximum likelihood estimators in this extreme scenario. This phenomenon was previously described in the SCCS literature and defined algebraically.[1] Specifically, the bias originates from the extreme unbalance between the expected events in the risk and control periods, and quickly reduces to negligible values when increasing the number of outcome events and/or the exposure episodes, as in Scenario 1. However, the bias is small even in this extreme scenario, and the case time series model maintains a nominal coverage (see Table 1 in the manuscript).

Scenario 3: Continuous exposure

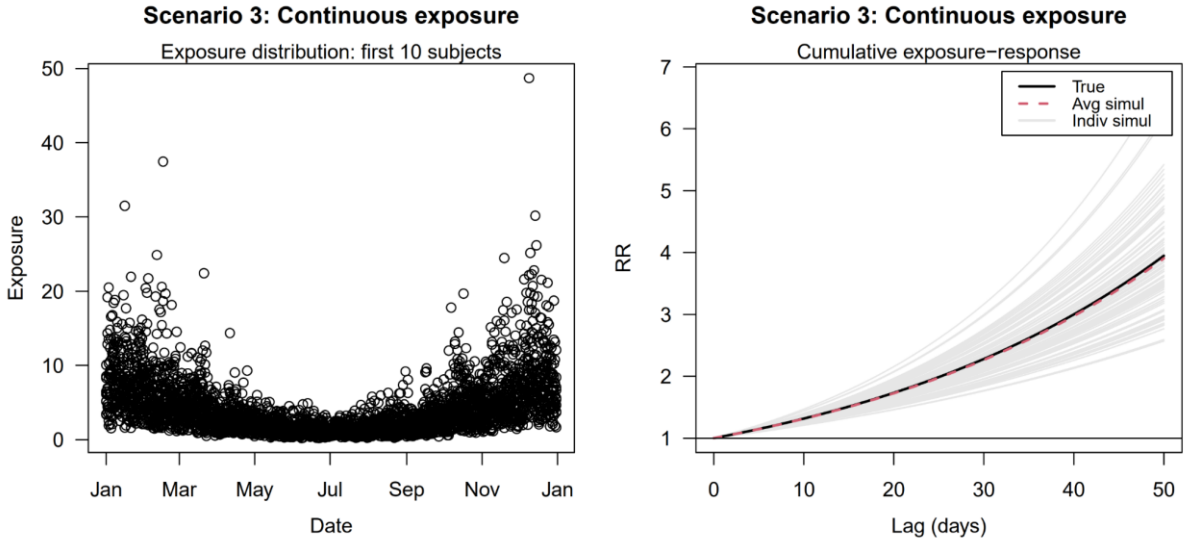
This scenario replicates Scenario 1, although using a continuous exposure instead of a binary indicator for exposure episodes. Specifically, a subject-specific exposure is simulated through the following function:

$$e(t) = \exp\big(\cos(2\pi/365 \cdot t) + 1 + v_{\rho\sigma}(t)\big)$$

where t is time defined as the day of the year (from 1 to 365), and $v_{\rho\sigma}$ is an auto-correlated random error simulated from a normal distribution with mean 0, standard deviation $\sigma = 0.5$, and correlation $\rho = 0.5$ (see also the related R script). This function defines an exposure variable x with strong seasonal distribution, as displayed in Figure S2 (left panel) with data simulated for the first 10 subjects. The exposure is associated with an increase in risk for an outcome event, with a constant relative risk (RR) of $\exp(\beta) = 1.0025$ for a unit increase in x in each day within the 0-10 risk period. Again, 5-20 occurrences of the repeated outcome events are sampled for each subject. The same fixed-effects Poisson regression model is used to estimate the association, although this time using as the single term the continuous measure of x_c cumulated across the 0-10 lag period.

Table 1 in the main manuscript confirms that the case time series models keep their optimal inferential properties in the analysis of continuous exposures. The flexible parameterization of the temporal effect of the exposure in the case time series design allows the derivation of more complex effect summaries. For instance, $\mathrm{RR} = \exp(\beta \cdot 11 \cdot x)$ represents the effect cumulated across the 11 days of the 0-10 lag period, and it can be

interpreted as the increase in risk at the end of the risk period associated to a single exposure episode x. This summary is displayed graphically in Figure S2 (right panel), which illustrates the cumulative exposure-response association across the exposure range. The graph confirms the absence of bias in the case time series model.



Scenario 4: Binary outcome

This scenario replicates Scenario 3, but by simulating an outcome represented by a binary indicator instead than by an event count. This means that the outcome measures presence/absence and not the number of events in each time unit. In most situations, the difference is subtle, but it implies different modelling choices, as described below. The scenario simulates the same continuous exposure with a seasonal distribution as in the previous case, and the same risk summary is represented by $\beta = \log(1.0025)$. However in this case $\exp(\beta)$ represents an odds ratio (OR) of a positive outcome and not a RR. The outcome is simulated from a Bernoulli distribution with probability $p = \exp(\alpha_b + \beta x_c)/\big(1 + \exp(\alpha_b + \beta x_c)\big)$, where $\alpha_b = \log(p_b/(1 - p_b))$, p_b is a baseline probability of 0.1, and $\exp(\beta x_c)$ is the OR associated with the number of exposure episodes x_c cumulated within the lag period.

Differently from all the other scenarios, the data are fitted using a fixed-effects logistic regression, simply

Figure S2. Left panel: distribution of the continuous exposure along the year for the first 10 subjects, simulated in one of the 50,000 iteration of Scenario 3, and then Scenarios 4-9 and 13. Right panel: Overall cumulative association representing the cumulative risk across the 0-10 risk period simulated in Scenario 3, with the true RR linear in the log scale (black line), the average estimate (red line), and the estimates of the first 100 iterations (grey lines).

replacing the Poisson with a binomial family. Results are reported in Table 1 of the main manuscript, and they demonstrate the ability of the case time series model in providing correct point estimates and coverage when analysing associations with binary outcome indicators that follow a Bernoulli distribution. It is worth noting that, in these data settings characterised by outcomes different from event counts, neither the SCCS nor the case-crossover (CC) designs are applicable.

Scenario 5: Continuous outcome

Similarly to the previous scenario, this simulation exercise replicates the settings of Scenario 3, with the only change being the outcome definition, this time represented by a continuous quantity instead of event counts. Specifically, a unit increase in exposure x is associated with an increase of $\beta = 0.01$ in a continuous outcome y, constant within the risk period 0-10. The outcome series y is simulated as the sum of three components: a subject-specific baseline randomly sampled from a uniform distribution between 50 and 150, the increase

associated with the exposure, and a random error simulated from a normal distribution with mean 0 and standard deviation of 20. The distribution of the outcome simulated for five subjects is displayed in Figure S3 (left panel).

Similarly to Scenario 4 and differently from the other previous scenarios, the case time series model is performed using a fixed-effects regression model that assumes a different distribution, specifically using a Gaussian family. This allows modelling additive relationships under the assumption of normally distributed errors. As in the case of binary outcomes in Scenario 4, it is worth noting that neither the SCCS nor the CC designs are applicable here. Results are reported in Figure S3 (right panel), which similarly to the same panel in Figure S2 displays the cumulative exposure-response association represented by $\beta \cdot 11 \cdot x$ across the exposure range. As above, the graph confirms the absence of bias in the case time series model for modelling relationships between a continuous exposure and a continuous outcome.

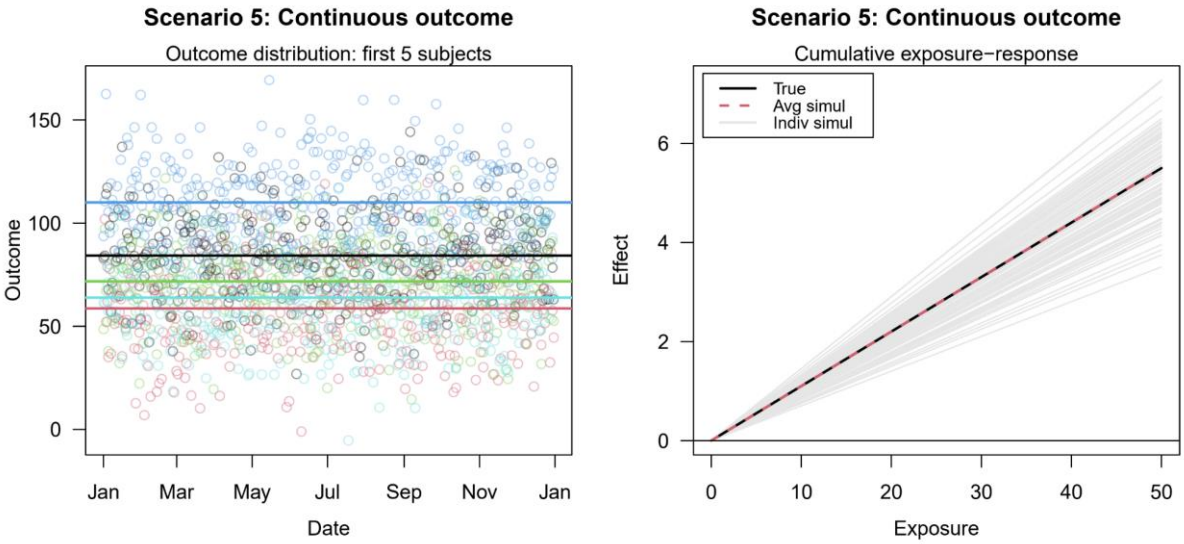


Figure S3. Left panel: distribution of the continuous outcome along the year for the first 5 subjects (in different colours), simulated in one of the 50,000 iteration of Scenario 5, together with the averages indicating differential baseline risks. Right panel: cumulative association representing the net effect across the 0-10 risk period simulated in Scenario 5, with the true linear relationship (black line), the average estimate (red line), and the estimates of the first 100 iterations (grey lines).

Scenario 6: Common trend

All the previous scenarios assume that the variation in risk within each individual is only due to the time-varying exposure and that the underlying baseline risk is in fact constant. This scenario depicts a more complex setting, with a common trend across the year that is shared by the 500 subjects. Given the strong seasonal distribution of the exposure, this trend needs to be adjusted for in order to obtain valid estimates of the association. The seasonal trend is simulated with the following function:

$$s(t) = \exp(\gamma_1 \sin(p_1 \pi/365 \cdot t) + \gamma_1 \cos(p_2 \pi/365 \cdot t))$$

where γ_1-γ_2 and p_1-p_2 are parameters of the sine and cosine terms, respectively, defining their amplitude and frequency. At each iteration, each parameter in the two pairs are sampled from a uniform distribution in the ranges -0.2 to 0.2, and -4 to 4, respectively, thus producing different common seasonal risk trends. Figure S4 (left panel) illustrates a random sample of seven iterations, showing shared trends with different peak/trough times, and either flat or strong.

The scenario replicates the repeated outcome events, continuous exposure, and constant risk period following exposure over lag 0-10 of Scenario 3. The same fixed-effects Poisson regression model is applied, but this time including a cyclic B-spline of the day of the year with 6 degrees of freedom (df) and a linear term for time to adjust for the seasonal and long-term trends. Results in the main manuscript (Table 1) indicate that the case

time series model is able to retrieve the true net risk with no bias and nominal confidence intervals, although with a higher root mean square error (RMSE), indicating a loss of precision due to the adjustment for the underlying trend.

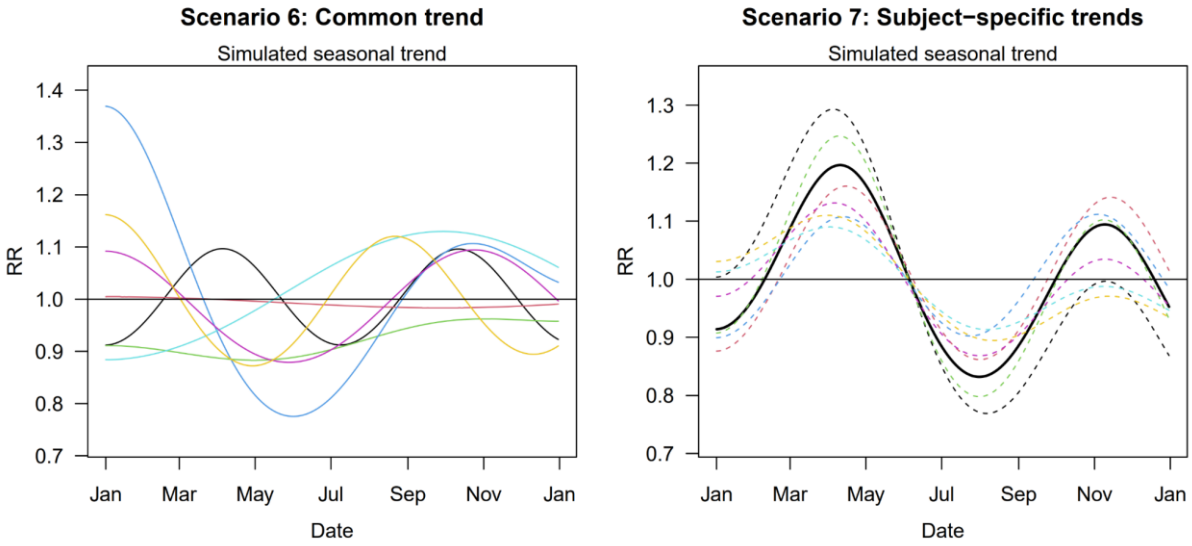


Figure S4. Left panel: common (shared) trend simulated in seven iterations (in different colours) in Scenarios 6-9, showing various shapes and strengths. Right panel: deviations (coloured dashed lines) from the common trend (continuous black line) for seven random subjects simulated in a single iteration in Scenarios 7-10, showing subject-specific trends.

Scenario 7: Subject-specific trend

This scenario makes the simulation setting even more complex by relaxing the assumptions of a common seasonal trend. An average baseline risk is first simulated as in Scenario 6 by randomly sampling the γ_1-γ_2 and p_1-p_2 parameters at each iteration. However, each parameter is then perturbed in each subject with a random amount independently sampled from a normal distribution with mean 0 and standard deviation 0.05, thus allowing subject-specific deviations. An example of a single iteration is depicted in Figure S4 (right panel), displaying the average and subject-specific trends.

The same fixed-effects Poisson regression of Scenario 6 is applied, but this time adding a stratification of the follow-up period, defining intercepts $\xi_{i(k)}$ at subject/month instead of subject-only level. These additional terms, not directly estimated but treated as nuisance parameters, allow subject-specific monthly deviations on top of the average trend captured by the cyclic B-splines. Again, the simulation results demonstrate that the case time series model can produce unbiased point estimates and confidence intervals. There is a further increase in RMSE due to the additional complexity of adjusting for the trends.

Scenario 8: Unobserved baseline confounder

This scenario introduces further complexities in the simulation setting by adding a risk factor z_f that varies across subjects but it is constant (fixed) in time. This is simulated independently for each subject by sampling a value from a uniform distribution between 0 and 100. A correlation with the continuous exposure x defined in Scenario 3 is then imposed by multiplying the latter by $z_f/50$, thus doubling the original exposure for a subject with $z_f = 100$. This creates a correlation between x and z_f, with a Pearson coefficient r of approximately 0.45. The risk factor z_f is assumed to be associated with a varying baseline risk, by setting the repeated events per subject as the rounded integer of $z_f/5 + 1$ instead of a random number between 5 and 20 as in the previous scenarios.

The case time series analysis is performed first using the same fixed-effects Poisson regression model of Scenario 7, without including the risk factor z_f. The results indicate no bias, thus demonstrating how the case time series, similarly to other self-matched methods, can control by design for unobserved baseline confounders that do not vary within the follow-up period.

Scenario 9: Time-varying confounder

This scenario follows the previous example by simulating an additional risk factor, which however is defined as a term z_v that varies both between and between subjects. This time-varying variable is simulated by perturbing the continuous exposure x defined in Scenario 3 with a random amount sampled from a normal distribution with mean 0 and standard deviation 3. This creates a strong correlation between the two terms x and z_v, with a Pearson coefficient r of approximately 0.80. The risk factor z_v is assumed to have an independent effect on the outcome, simulated as a same-day RR of $\exp(0.01) \cong 1.01$ for a unit increase.

The case time series analysis is performed first using the same fixed-effects Poisson regression model of Scenarios 7 and 8, although in this case adding z_v as a simple linear term with no lag. The results suggest no evidence bias, thus demonstrating how the case time series design provides a way to effectively control for confounding from measurable time-varying factors if their risk associations are appropriately specified in the regression model.

Scenario 10: Complex lag structure

The previous scenarios assume a simple temporal relationship with a constant risk across the pre-defined period of 10 days. This scenario uses a combination of the settings of Scenarios 7-9, but it describes a more complex temporal dependency by assigning different weights to each lag ℓ in the interval 0-10 through the function:

$$w(\ell) = \phi_{2,2}(\ell)$$

where $\phi_{m,s}$ is a normal density function with mean m and standard deviation s. This choice defines a lag structure with an initial increase in risk, a peak after 2 days, and then an attenuation until the effect disappears after about 8 days (see Figure S5). The weights are then re-calibrated to produce lag-specific effects β_ℓ, with $\sum \beta_\ell = \beta_c$ and $\mathrm{RR} = \exp(\beta_c)$. In order to produce comparable risk estimates as the previous scenarios with constant risk across lags, the cumulative risk is simulated as $\beta_c = \log(1.0025) \cdot 11$. This net risk summary is the focus of the inferential assessment using the measures or bias, coverage, and RMSE defined at the beginning of the document. In addition, in these most complex scenarios, both unobserved time-invariant and observed time-varying risk factors simulated following the same definition of Scenarios 8 and 9, respectively, plus the same subject-specific trends described in Scenario 7.

The case time series analysis is performed using a model similar to the fixed-effects Poisson regression of Scenario 9, but this time including a *distributed lag model* (DLM) to describe more flexibly the complex lag structure associated with the exposure.[2] This term is parameterized by a *cross-basis*, a bi-dimensional function expressed in the spaces of the exposure and lag. Specifically, the exposure-response is modelled using a simple linear function, while a natural cubic spline with three equally-spaced knots a lags 2.5, 5.0, and 7.5 is applied to model the lag-response association that describes the temporal structure.

Results reported in the main manuscript (Table 1) indicate unbiased point estimates and confidence intervals for the log cumulative risk β_c. Figure S5 confirms the findings across the lag-response space, showing that the case time series model is capable of retrieving complex lagged associations through the application of sophisticated time series techniques based on DLMs. It is worth noting that this complex temporal relationship is reliably estimated even in the presence of strong baseline and temporal confounding from time-invariant and time-varying risk factors, in addition to subject-specific trends, all of which are appropriately controlled for either by design or by including related terms in the regression model.

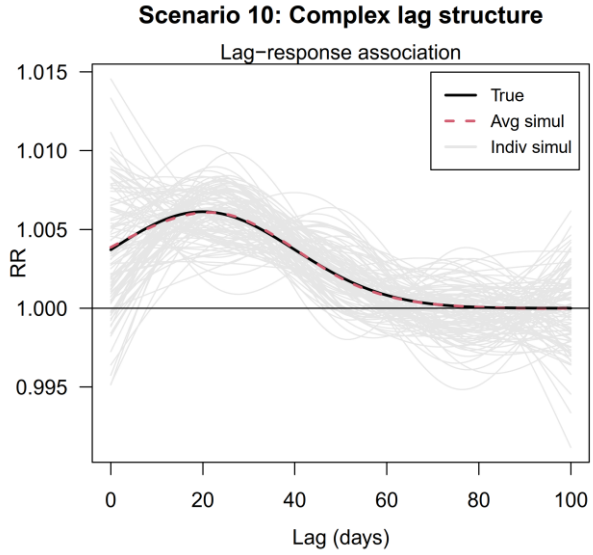


Figure S5. Complex lag structures simulated in Scenario 10. The graph shows the lag-response associations with the varying pattern of lag-specific RR, with the true simulated association (black line), the average estimate (red line), and the estimates of the first 100 iterations (grey lines). The estimates of the lag-response curve in this scenario is adjusted for additional time-invariant and time-varying confounder, in addition to underlying trends in risk.

Part II: assessment of underlying assumptions

In contrast to the previous nine scenarios, the second part of the simulation study (Scenarios 11-14) illustrates basic data settings, where however each of the four assumptions that underpin the case time series design is in turn violated (Table S2). It is expected that when data are simulated in scenarios where one of the assumptions does not hold, the inferential performance is affected, with the occurrence of biases in point estimates or wrong coverage of the confidence intervals.

Table S1. Description of simulation scenarios where each of the assumptions of the case time series design are violated.

Scenario	Outcome	Exposure	Trend	Lag structure	Confounder
Scenario 11: Outcome-dependent risk	Count	Episode	None	Simple	None
Scenario 12: Outcome-dependent follow-up	Count	Episode	None	Simple	None
Scenario 13: Outcome-dependent exposure	Count	Episode	None	Simple	None
Scenario 14: Variation in baseline risk	Count	Continuous	None	Simple	None

Scenario 11: Outcome-dependent risk

This is the first of four scenarios illustrating examples where one of the underlying assumptions of the case time series design does not hold. In particular, this scenario depicts a complex form of dependency within the series y, where the occurrence of an outcome event modifies the risk of future outcomes.[3] The example uses the same basic setting of Scenario 1, with the sampling of 5-20 *potential* outcome events per subject. However, each event carries a risk of 0.2 that future events are not occurring. This situation can arise, for instance, in the presence of a risk of death related to the outcome of interest, or because the subject changes status and his/her outcome cannot be recorded. An example is illustrated in Figure S6 (left panel), with 10 subjects for whom the risk of any future outcome can vanish after a given outcome event. Note that the subjects are still under follow-up, differently from the following scenario.

The same fixed-effects Poisson model is used to estimate the parameter β representing the risk associated with the exposure. As shown in Table 1 in the main manuscript, however, the estimates are affected by a noticeable negative bias. The mechanism can be explained by the fact that exposures episodes occurring after the change of status are not anymore associated with an increased risk. An extreme case of this situation is represented by the analysis of non-recurrent outcomes, which must be rare in the population of interest for avoiding the bias described here, as previously discussed in the literature of other self-matched designs.[4-6]

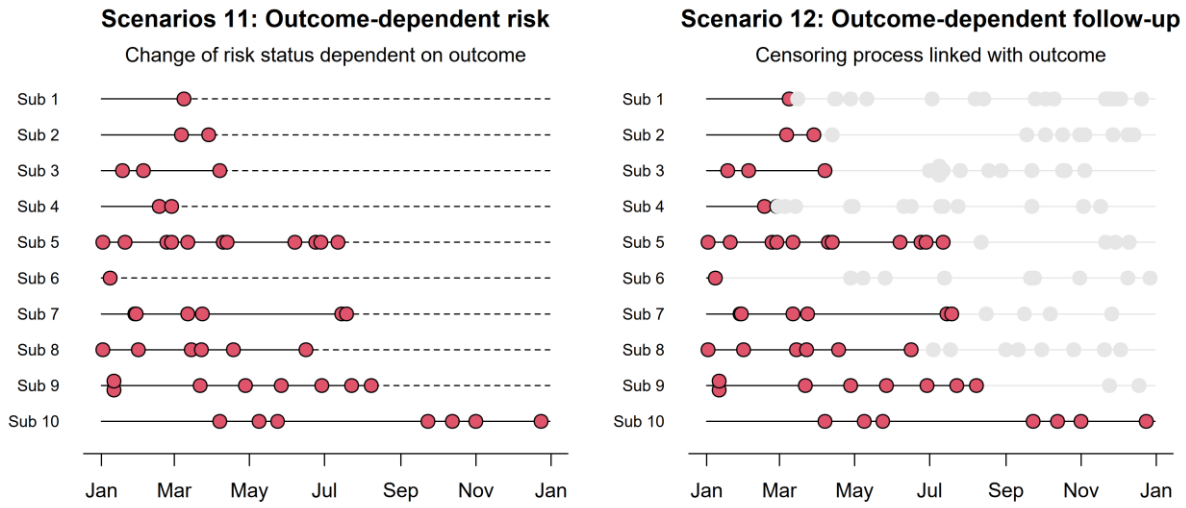


Figure S6. Graphical representation of data simulated in Scenarios 11 and 12, with the follow-up period (black lines) and outcome events (red circles) of 10 subjects. Left panel: an example of data with outcome-dependent risk simulated in Scenario 11, where each subject has a probability of 0.2 of switching to a no-risk status after each event. Right panel: the same example for Scenario 12, where the same process leads instead to censoring. Note that in the first example the follow-up continues (dashed lines) with no recorded outcome events, while in the second example the follow-up stops and the time (grey line) and potential outcomes (grey circles) do not occur.

Scenario 12: Outcome-dependent follow-up

The following assumption of the case time series design states that the follow-up period must be independent of the outcome. This assumption was previously described in the context of the self-controlled case series study.[7,8] In particular, for event-type outcomes, this means that the occurrence of an event must not modify the probability of censoring the follow-up. Similarly to the previous scenario, the same settings of Scenario 1 are used to generate the complete data for the 500 subjects. However, then an artificial outcome-dependent censoring mechanism is simulated, sampling the occurrence of a censoring event on the day after an outcome with a probability of 0.2. This means that subjects have a 20% risk of having their follow-up stopped after experiencing one outcome event. An example is shown in Figure S6 (right panel), with the follow-up periods of ten subjects.

Although no other modification is applied to the data, the estimate from the fixed-effects Poisson regression model is biased upward, as shown in Table 1 of the main manuscript. Compared to Scenario 11, the direction of the bias is reversed, as potential post-event times are not always included.

Scenario 13: Outcome-dependent exposure

The third assumption listed in the manuscript dictates that a given outcome must not modify the probability distribution of the exposure x in the following period. Similarly to the previous example, this assumption was previously described in the context of the self-controlled case series study.[7,9] This scenario drops this assumption by simulating an inverse temporal relationship between exposure and outcome. The simulation setting replicates again Scenario 1, with one modification. Specifically, in addition to the usual $\mathrm{RR} = \exp(\beta_\ell) = 1.15$ that defines the increase in risk in of the 0-10 lag day following the exposure, another relationship is defined over *lead times* 1:14, meaning the series of lags from -14 to -1. This inverse temporal relationship is defined as $\mathrm{RR} = 0.60$, thus generating data where the occurrence of an outcome event is associated with a decreased probability of an exposure episode in the following two weeks.

The data are fitted fixed-effects Poisson model with a single term x_c representing the exposure cumulated within the 0-10 lag period (corresponding to the same day and 10 days before the outcome). However, the presence of an independent but unaccounted inverse temporal relationship generates an imbalance in the temporal comparison defined within the self-matched data structure. This explains the noticeable bias in the estimates, with the underestimation reported in Table 1 of the main manuscript.

Scenario 14: Variation in baseline risk

The last scenario deals with the fourth assumption of the case time series design, which states that any variation in the baseline risk within the follow-up period (or within strata of it) must be fully explained by model covariates. This assumption is the same applied within the risk sets of case-crossover design.[5,6] There are various situations where this is not the case. This scenario simulates unobserved temporal changes in baseline risk, for example, due to holiday periods where a given outcome has fewer chances to be recorded. The simulation exercise uses the same settings of Scenario 3 but including a random period of one month within May-September where the subject is at lower risk, using an RR of 0.7.

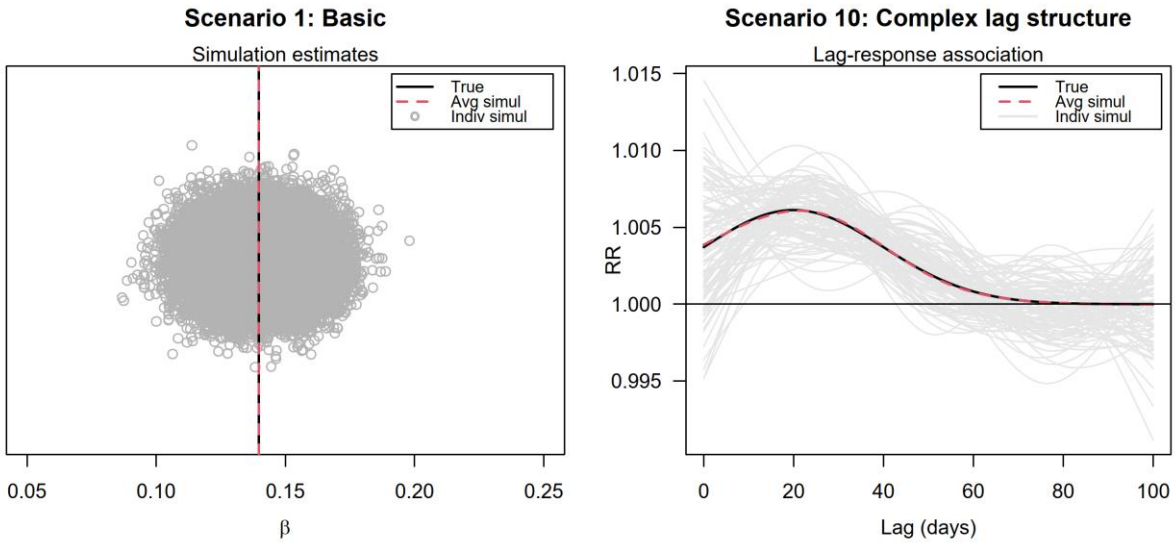
The data are fitted with the same fixed-effects Poisson model. As expected, the results in Table 1 of the main manuscript indicate a bias, with the overestimation due to unaccounted temporal differences that affect the conditional exchangeability required by the case time series design. Similar biases can arise in the presence of potentially measurable but unaccounted risk factors, for example with a modification of Scenario 9 when the time-varying variable z_v is not included in the model.

References

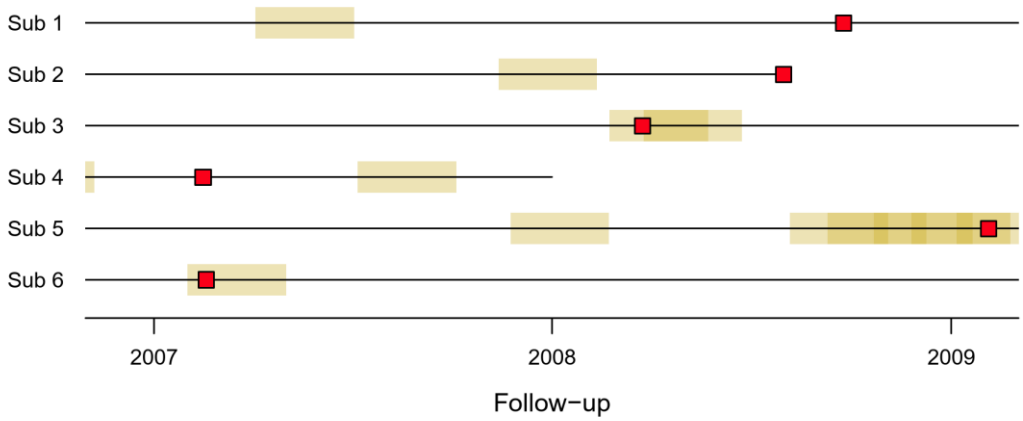
1. Musonda P, Hocine MN, Whitaker HJ, Farrington CP. Self-controlled case series analyses: small-sample performance. *Computational Statistics & Data Analysis*. 2008;52(4):1942-1957.
2. Gasparrini A. Modeling exposure-lag-response associations with distributed lag non-linear models. *Statistics in Medicine*. 2014;33(5):881-899.
3. 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.
4. 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.
5. 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.

6. Lu Y, Zeger SL. On the equivalence of case-crossover and time series methods in environmental epidemiology. *Biostatistics*. 2007;8(2):337-344.
7. 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.
8. 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.
9. Farrington CP, Whitaker HJ, Hocine MN. Case series analysis for censored, perturbed, or curtailed post-event exposures. *Biostatistics*. 2009;10(1):3-16.

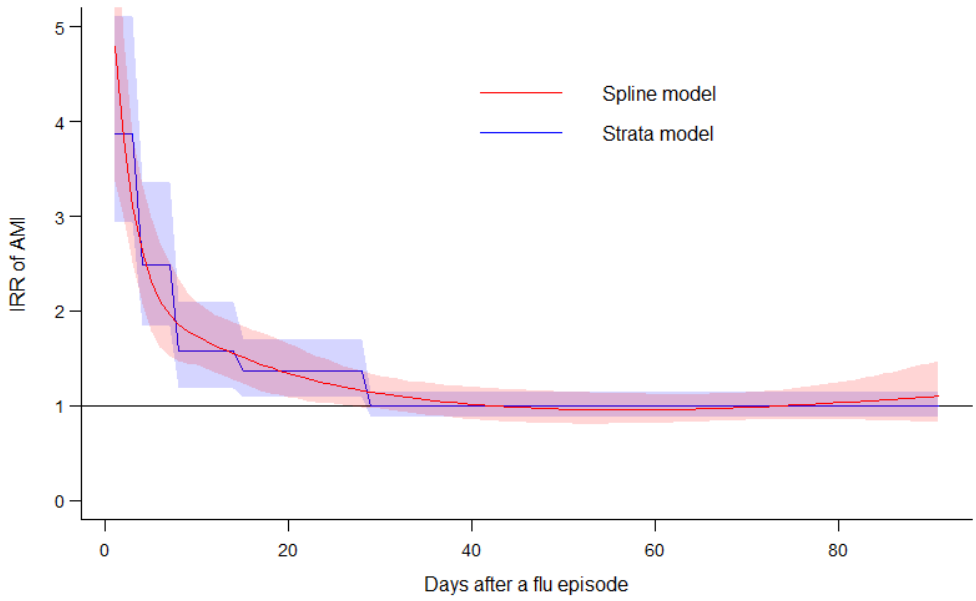
eFigures



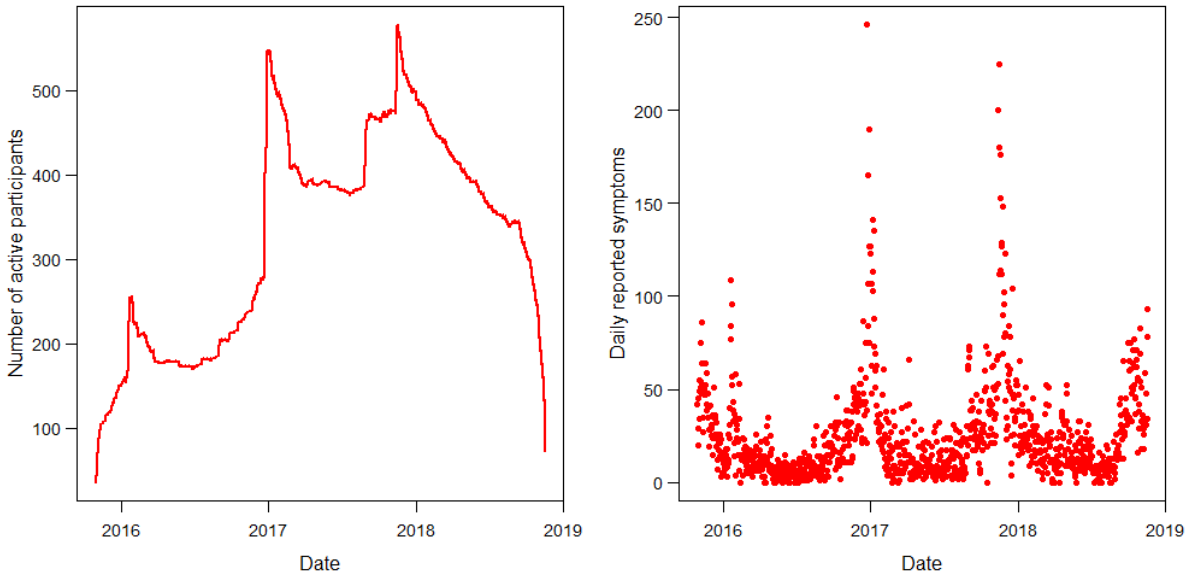
eFigure 1. Results of the simulation study in two scenarios. Left panel: results from the basic Scenario 1, with the true simulated association (black line), the average estimate of the 50,000 iterations (red line), and the estimates of the individual iterations (grey dots, scattered across the y-axis). Right panel: results from the more complex Scenario 10, with lag-response associations represented by the true simulated curve (black line), the average estimate (red line) of 50,000 iterations, and the estimates of the first 100 iterations (grey lines).



eFigure 2. Graphical representation of a sub-interval of the follow up period for six of the 3,927 subjects included in the study on the association between influenza infection and myocardial infarction (AMI). The red circles represent AMI events, while the yellow bands represent exposure period defined as 1-91 days after a flu episode. Note how some subjects have their follow-up censored before the end of the study period (subjects 2 and 4), or overlapping exposure periods for repeated flu episodes (subjects 3 and 5).



eFigure 3. Comparison of lag-response associations estimated by two alternative case time series (CTS) models, as relative risk (RR) and 95% confidence intervals. The curves represent the risk of acute myocardial infarction (AMI) in the 1-91 days following a flu episode, estimated using natural cubic splines (red) and step functions (blue).



eFigure 4. Number of participants (left panel) and daily respiratory symptoms (right panel) during the study period of the AirRater study. The graphs indicate a complex study setting, characterized by continuous recruitment, high dropout rates, intermittent participation, and a highly seasonal outcome with peaks of self-reported allergic symptoms in the Australian summer period.