Novel time series methods in epidemiology and public health

Antonio Gasparrini

London School of Hygiene & Tropical Medicine, UK

PhD course in Translational Specialistic Medicine 'G.B. Morgagni', University of Padua

Virtual seminar – 30 September 2022

Why time series?

Time series analysis consists of study designs and analytical techniques that have been historically used in specific research areas

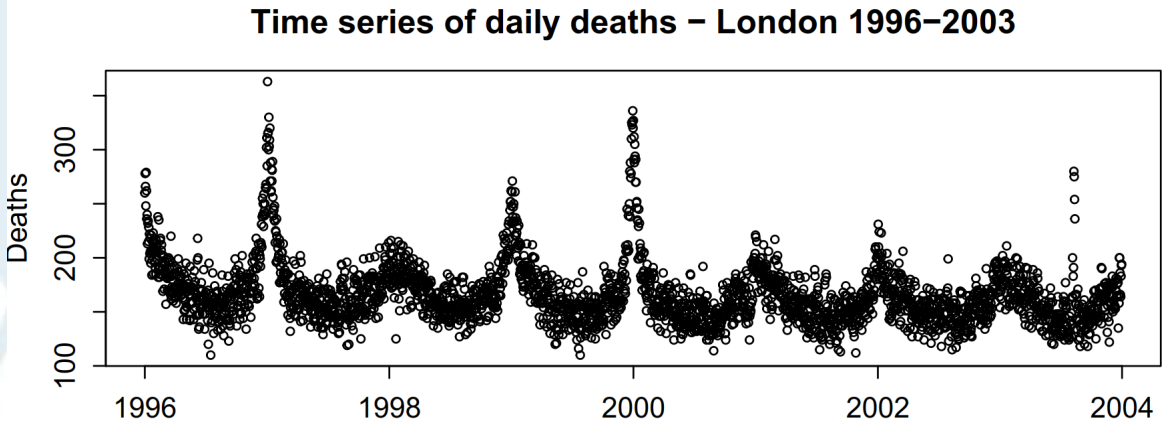
Analytical methods for time series analysis have been **initially developed in econometrics**, where the availability of such type of data were widespread since decades ago

Intense methodological development, but tailored to data and research context of econometrics

Time series in health research

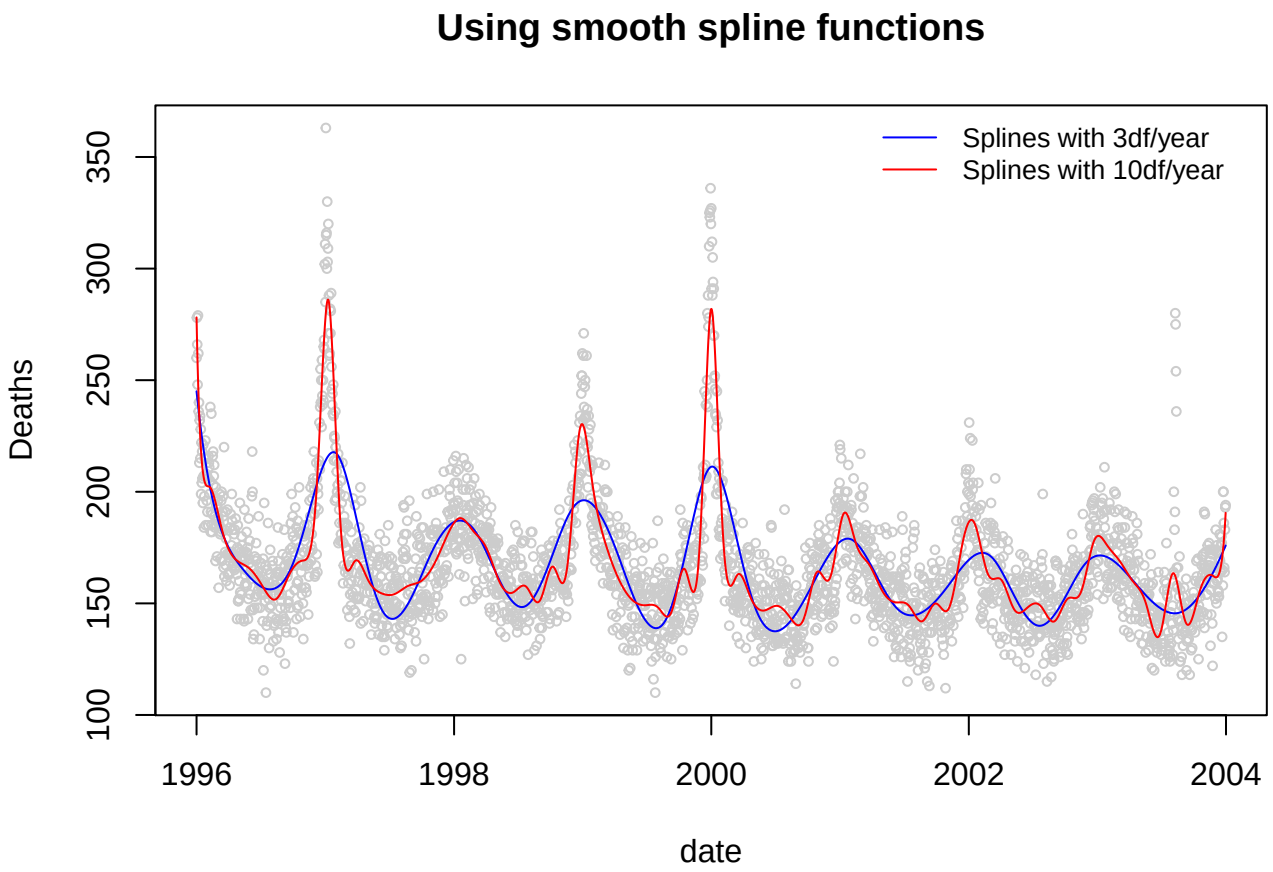
In recent times, however, time series methods are slowly but progressively becoming **more popular** also in health research

One of the reasons is the increasing availability of measures of health outcomes and risk factors routinely collected at **equally-spaced times**

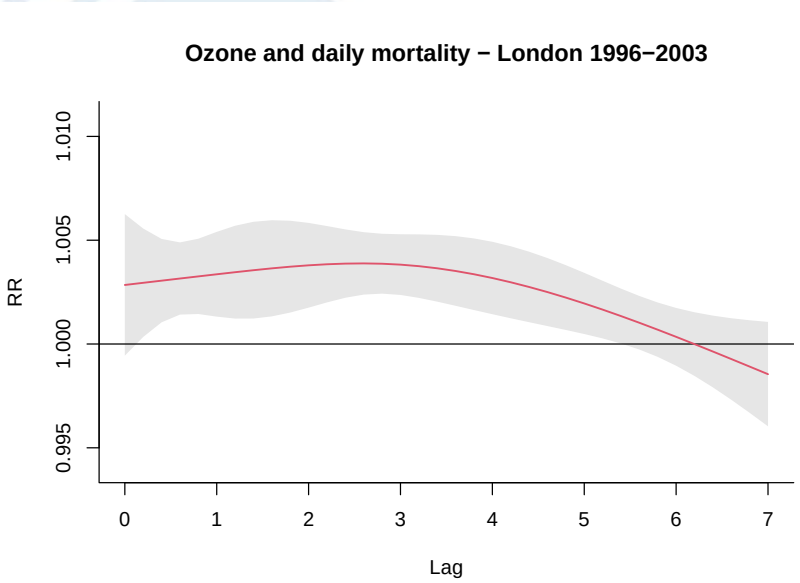


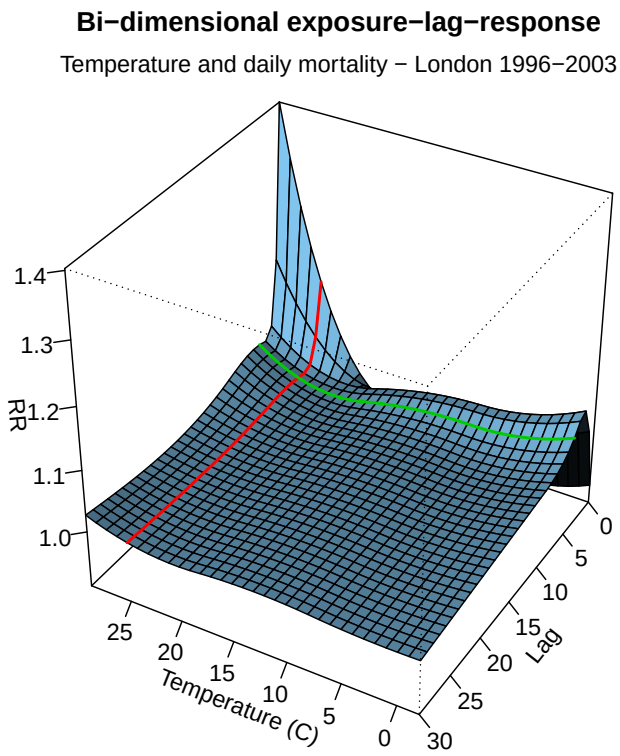
LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Flexible TS modelling framework



Exposure-lag-response relationships





New opportunities and challenges

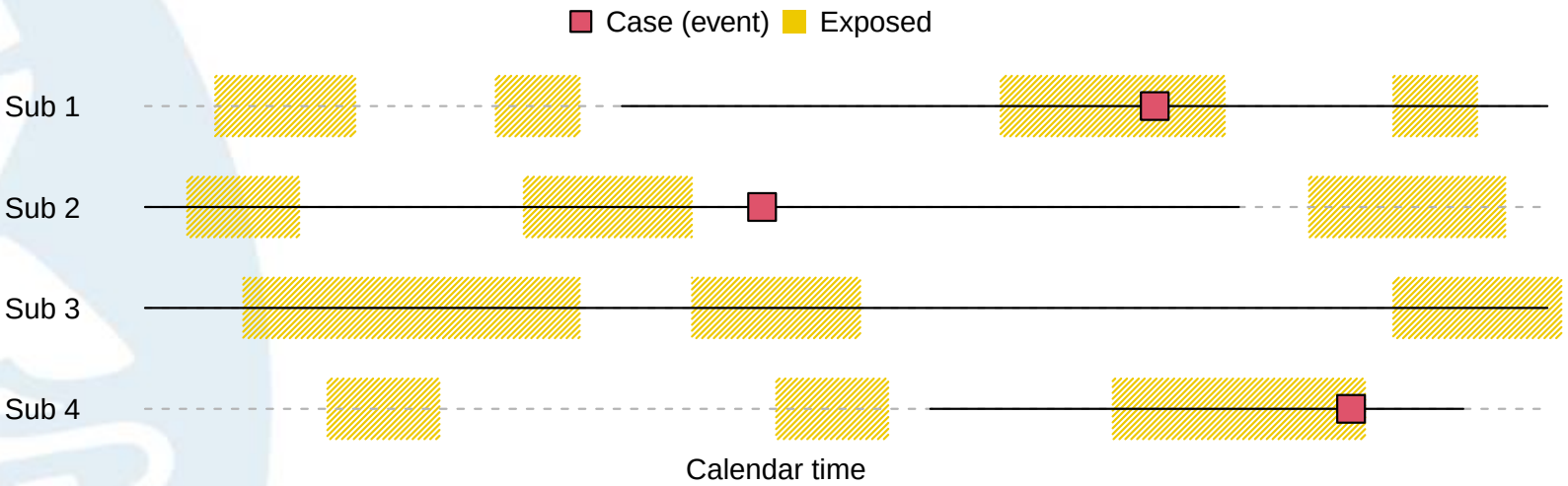
Novel **big data technologies** (e.g., wearables, remote sensing, electronic health records) offers new opportunities for health research

Potential of collection of **large population-based datasets** with measurements of individual-level risk factors and personal characteristics

Ideally, time series methods are well suited for analyses of **longitudinal repeated measures** of time-varying health outcomes and predictors

However, **big limitation**: traditional time series methods only developed for aggregated data

A methodological case study



Examples of short-term (transient) associations:

- Physical exercise and myocardial infarction (clinical study)
- Air pollution and asthma exacerbations (environmental research)
- Paracetamol use and gastrointestinal bleeding (pharmaco-vigilance)

Self-matched designs

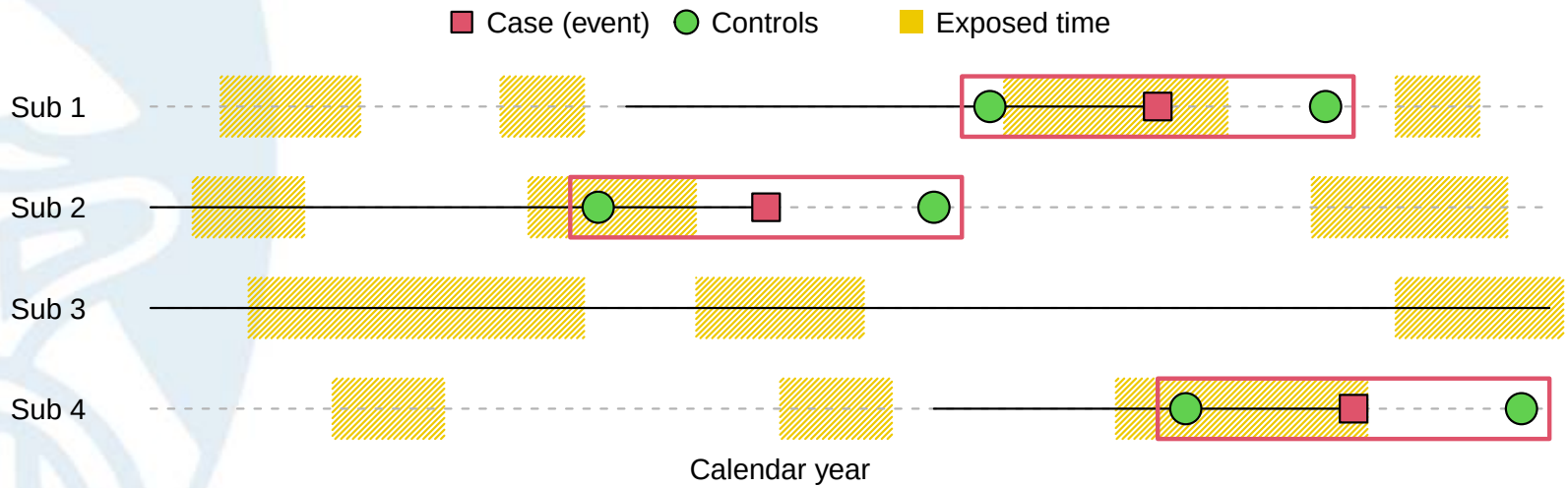
Recent development of **self-matched methods** for the analysis of transient (short-term) effects associated with intermittent or generally time-varying exposures

Main advantages:

- **Control by design** for time-invariant risk factors, reducing the set of potential confounders if compared to studies requiring separate controls
- **Computational efficiency** related to the analysis being restricted to cases, and the specific form of estimators

The case-crossover design

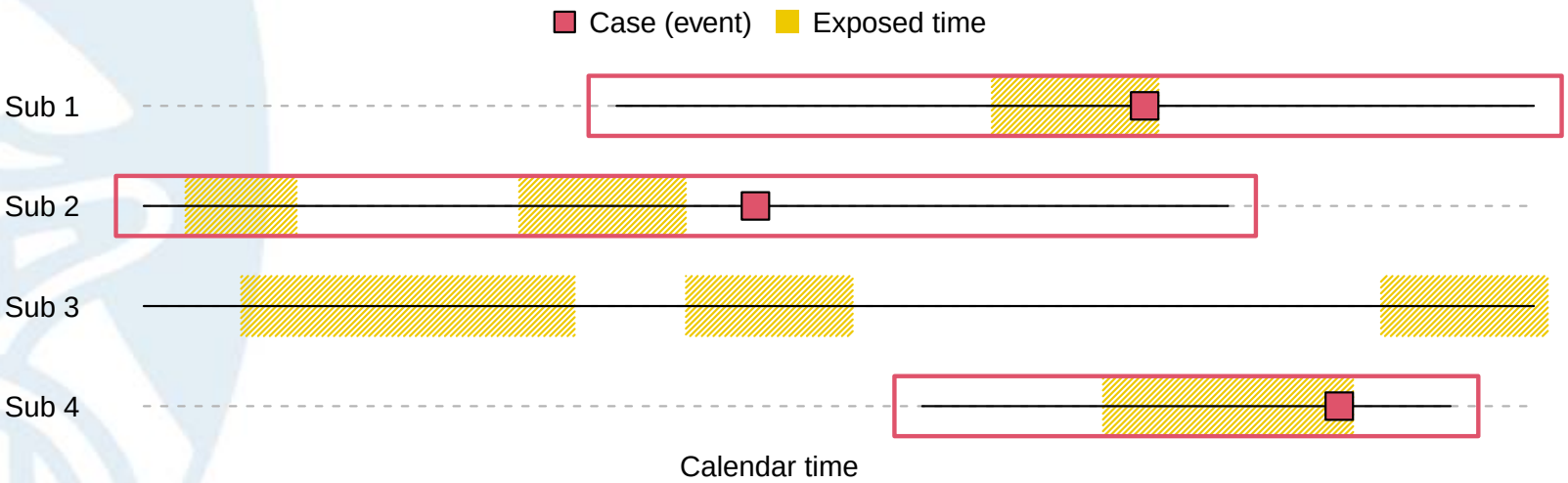
Case-only design with **within-subject matched risk sets** base on a **case-control logic**



Intense methodological work on **control sampling schemes**

The self-controlled case series design

Originally developed in vaccine safety evaluation, it is a case-only design based on a **cohort logic**



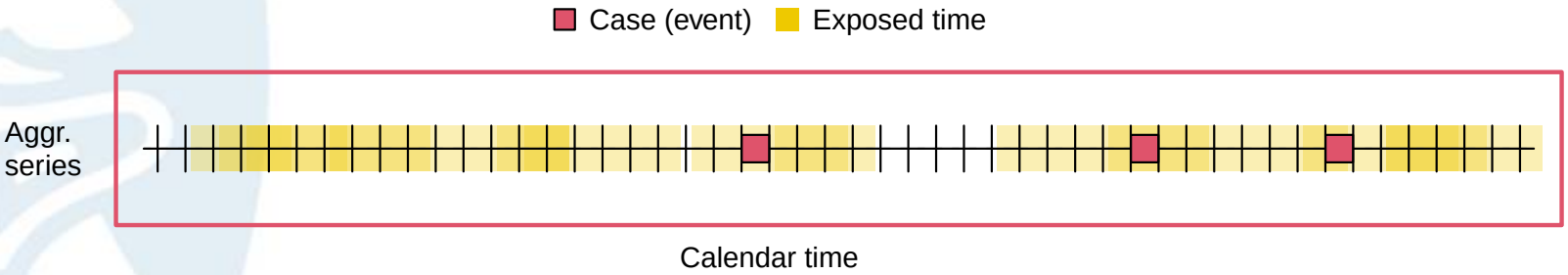
Elegant framework supported by a set of **well-defined assumptions**

Limitations

- Only applicable for **event-type** outcomes (SCCS and CC)
- Not applicable with **continuous** (SCCS) or **rare exposures** (CC)
- Difficult to control for **time-varying confounders** (SCCS)
- Difficult to model **temporal dependencies** (SCCS and CC)
- Complexity of **control sampling schemes** (CC)
- Lack of **longitudinal structure** (SCCS and CC)

The time series design

Developed in econometrics and more recently proposed for epidemiological analysis, applied by **aggregating the data** in a single series



Flexible design framework based on **advanced analytical techniques**

An idea...

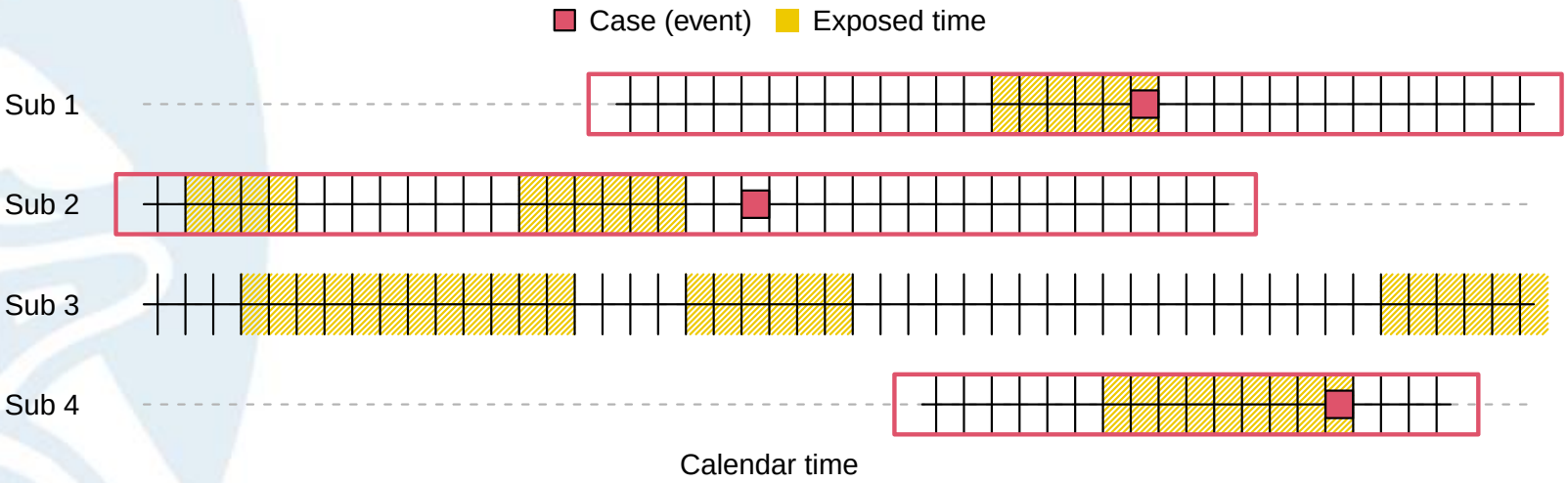
Each of these study designs has its own advantages and limitations

What about combining their features, keeping:

- the **individual-level setting** and **self-matched contrasts** of CC/SCCS
- the **temporal structure** and **modelling flexibility** of TS

The case time series design

Combining CC/SCCS **individual-level setting** with TS **temporal structure**



The design is based on the reconstruction of **longitudinal profiles** of health outcomes and time-varying predictors in **subject-specific series**

Modelling framework

Regression model for case time series analyses:

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

Not surprisingly, this resembles regression models for time series analysis, with:

- **Temporal relationships** modelling $f(x, \ell)$ with DLMs/DLNMs
- **Smoothing methods** for controlling for trends in $s(t)$
- **Time-varying confounders** easily modelled through $h(z_p)$

However:

- Analysis of multiple **individual series** (index i)
- Subject-specific **strata terms** $\xi_{i(k)}$

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Estimation and computational aspects

While the modelling framework resembles time series, the **estimation methods** are borrowed from other case-only designs

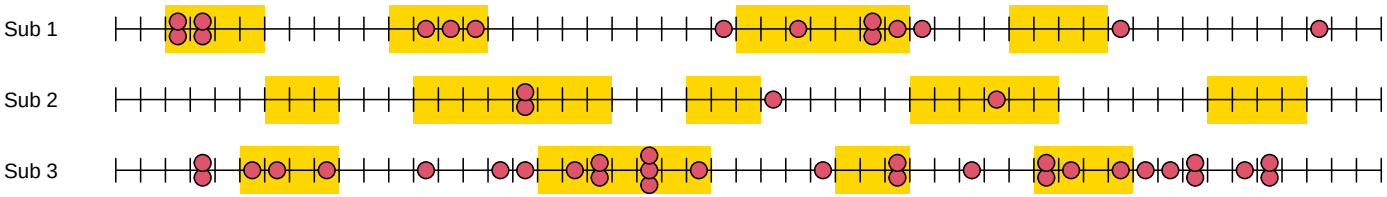
Estimators based on conditional likelihood expressed as within-subject comparisons: subject-specific terms ξ_i treated as *nuisance parameters* and **conditioned out**

Similar conditional statements used for *fixed-effects* models with normally-distributed responses in **panel data analysis**

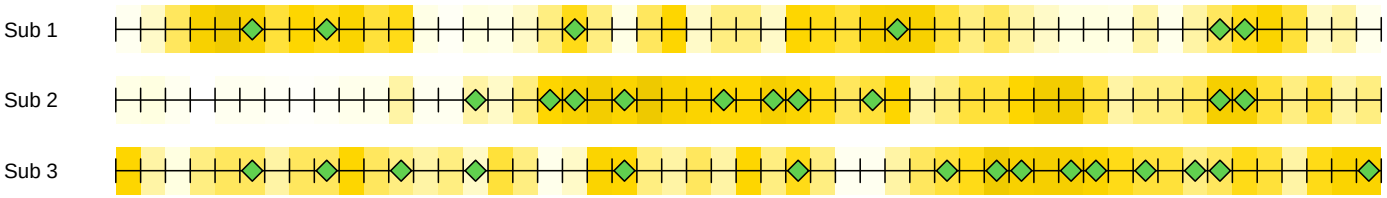
Efficient computational scheme, with case-only data and conditional likelihood written as a sum of subject-specific components

Flexible analytical framework

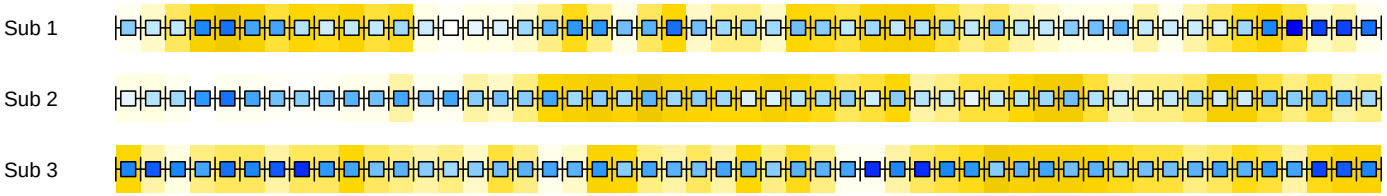
Count (event) outcome – binary (indicator) exposure

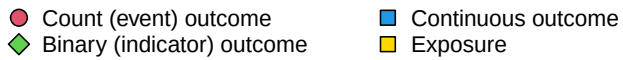


Binary (indicator) outcome, continuous exposure



Continuous outcome, continuous exposure





LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Assumptions

- **Distributional assumptions on the outcome**: conditionally independent observations originating from one of the standard family distributions (Poisson, Bernoulli/bionomial, Gaussian)
- **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
- **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
- **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

Case Study 1: flu and myocardial infarction

Background: hypothesis that acute respiratory infections act as a trigger acute myocardial infarction (AMI)

Data: Linkage between EHR (MINAP and GPRD) to retrieve data on 3,927 patients who experienced a first AMI and had at least one flu consultation in 2003–2009

Analysis: Application of smoothing techniques to control for trends (age and calendar time), and DLMs to investigate temporal patterns

Objective: Demonstrate application with EHR, highlight the flexible modelling framework of case time series, comparison with self-controlled case series

Original results and and limitations

Application of a standard self-controlled case series analysis, defining **temporal windows** within 1–91 days after a flu episode and controlling for age and season with **strata indicators**

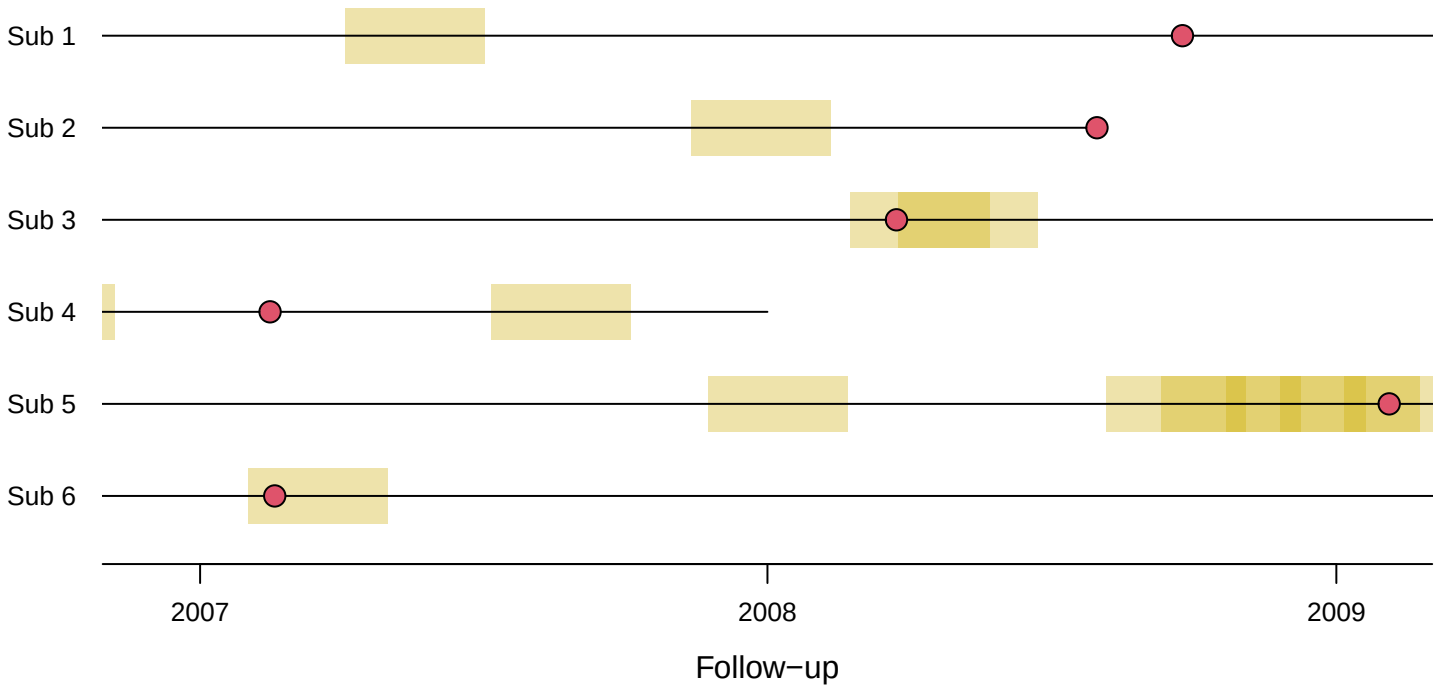
Results (IRR with 95% C):

- Days 1–3: 4.19 (3.18-–5.53)
- Days 4–7: 2.69 (1.99—3.63)
- Days 8–14: 1.66 (1.24—2.23)
- Days 15–28: 1.41 (1.12-–1.77)
- Days 29–91: 1.05 (0.92-–1.21)

However, issues with **overlapping windows** and control for **time-varying confounders**

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Subject-specific profiles



Main analysis

Analysis of 3,927 subjects followed up between 01/01/2003 and 31/07/2009

Conditional Poisson GLM using the function `gnm()` in R

Outcome: binary indicator of first event of acute myocardial infarction

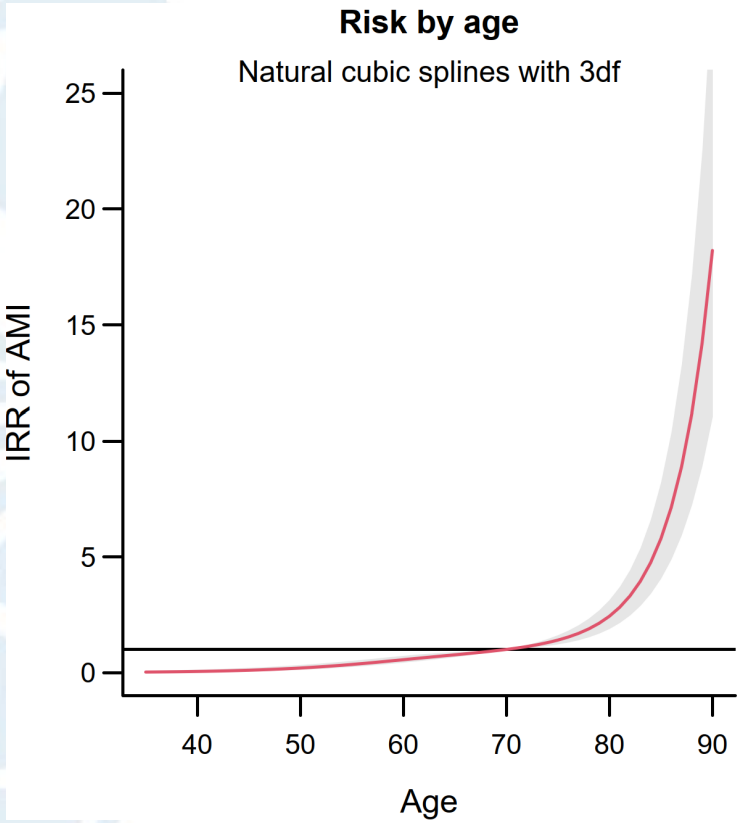
Exposures: flu episodes(s) with a lag period of 1–91 days, modelled with natural cubic splines or with step functions with 5df

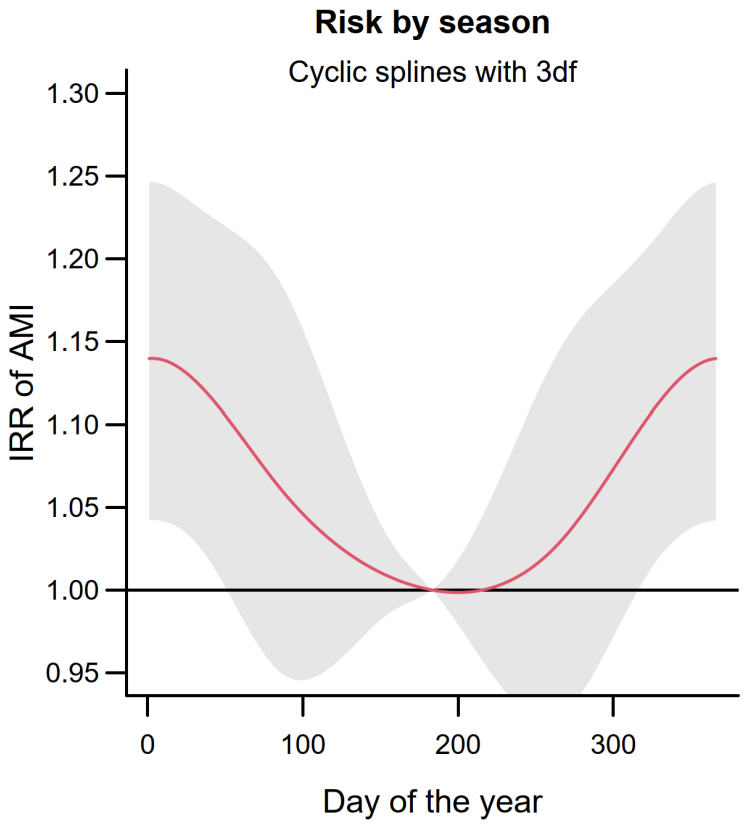
Stratification: subject-specific

Additional temporal control: natural cubic splines with 4df for age and cyclic splines with 3df for seasonality

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

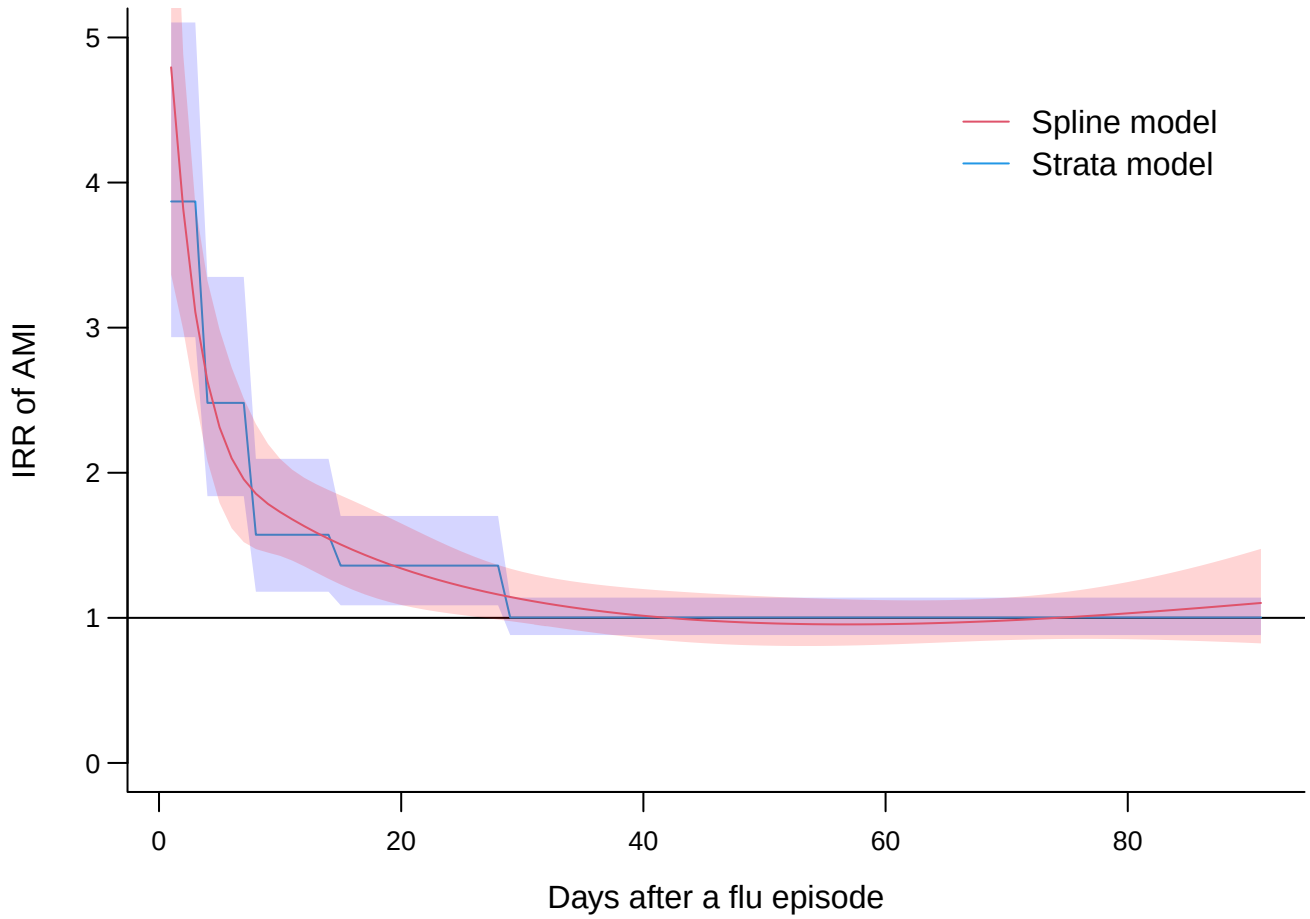
Control for time-varying confounders





LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Lag-response relationship



Case Study 2: Antipsychotic drugs and AMI

Background: hypothesis that use of antipsychotic drugs increases the risk of acute myocardial infarction (AMI)

Data: Linkage between EHR (MINAP and CPRD) to retrieve data on 1,546 patients prescribed with antipsychotic who experienced an AMI in 2003–2009

Analysis: Application of DLMs to investigate temporal patterns and account for overlapping periods exposure

Objective: Demonstrate application to investigate side effects of drugs in pharmaco-epidemiological studies

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Original study results

Table 2 **Results self-controlled case series**

Type of anti-psychotic	Exposure	Patient years	n MIs	Crude rate-ratio for MI [95% confidence interval (CI)]	Age-adjusted rate ratio for MI (95% CI) corrected for censoring
First generation	Unexposed	11 748	1021	Baseline	Baseline
	Exposed periods first 1–30 days of exposure	94	35	2.85 (2.02–4.02)	2.82 (2.0–3.99)
	Exposed periods 1–30 days for subsequent episodes of exposure[a]	97	17	2.04 (1.25–3.34)	1.95 (1.19–3.21)
	Exposed periods 31–90 days	330	49	1.44 (1.07–1.94)	1.41 (1.04–1.9)
	Exposed periods >90 days	789	104	1.57 (1.21–2.06)	1.47 (1.12–1.93)
	Post-exposure period 1–60 days	282	31	1.17 (0.81–1.68)	1.15 (0.8–1.66)
	Post-exposure period 61–120 days	239	34	1.53 (1.08–2.17)	1.52 (1.07–2.16)
	Post-exposure period 121–180 days	215	15	0.76 (0.46–1.28)	0.76 (0.45–1.27)

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Main analysis

Analysis of 1,546 subjects followed up between 01/01/2003 and 31/07/2009

Conditional Poisson GLM using the function `gnm()` in R

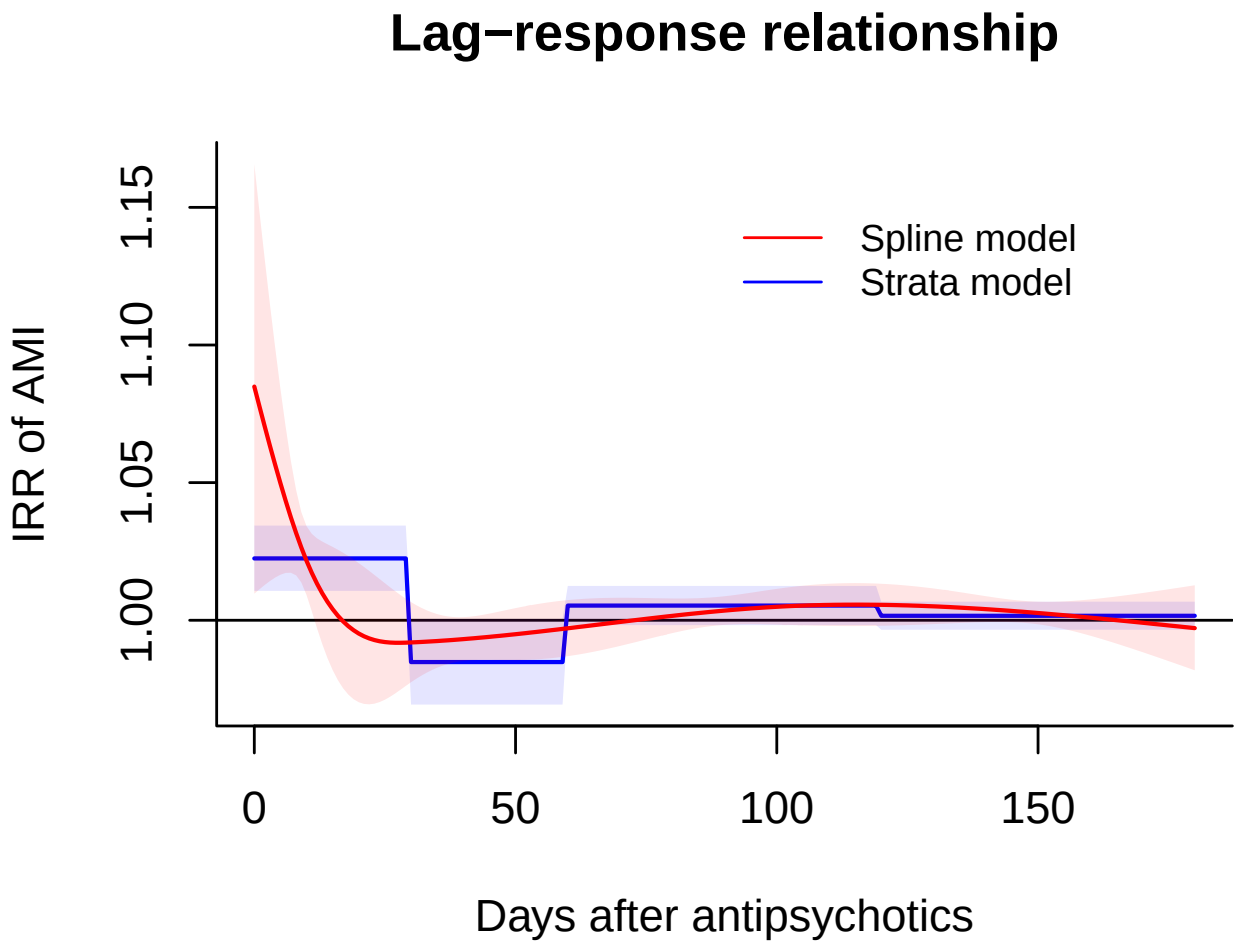
Outcome: binary indicator of first event of acute myocardial infarction

Exposures: days under prescription with a lag period of 0–180 days, modelled with natural cubic splines or with step functions with 4df

Stratification: subject-specific

Additional temporal control: none

Preliminary results



Case Study 3: environmental factors and allergy

Background: hypothesis that multiple environmental stressors exacerbate the risk allergic symptoms

Data: 1,601 subjects in Tasmania during the period Oct 2015–Nov 2018, with daily questionnaires obtained through a smartphone app and linked with environmental measurements through geo-location

Analysis: Complex temporal relationships between continuous exposures and repeated measurements of outcomes

Objective: Demonstrate application real-time mobile technologies, describe the setting where other case-only designs cannot be applied

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

The AirRater app

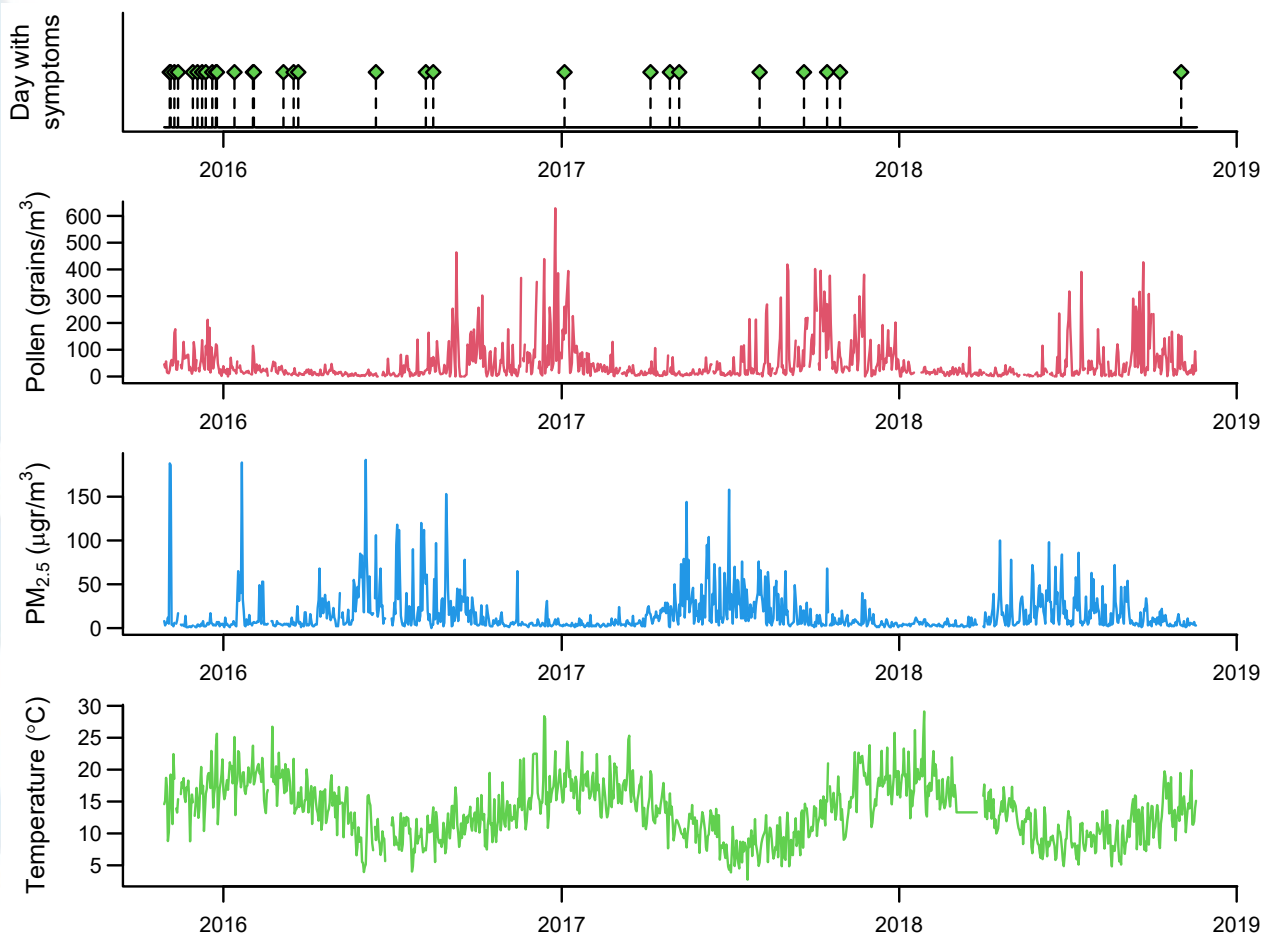
Aim: association between multiple environmental exposures and allergic symptoms

Mobile app developed to gather information on reported allergic symptoms indices and potentially related aspects (sleep, mood, activity) with a daily questionnaire

The smartphones also track individuals, offering **geo-located coordinates** that can be linked with spatio-temporal exposure maps from other sources

[https://airrater.org/]

Subject-specific profiles



Main analysis

Analysis of 1,601 subjects in Tasmania during the period Oct 2015–Nov 2018

Binomial GLM using the function `gnm()` in R

Outcome: daily binary indicator of occurrence of allergic symptoms

Exposures: pollen, PM2.5, and temperature, each modelled with DLMs/DLNMs

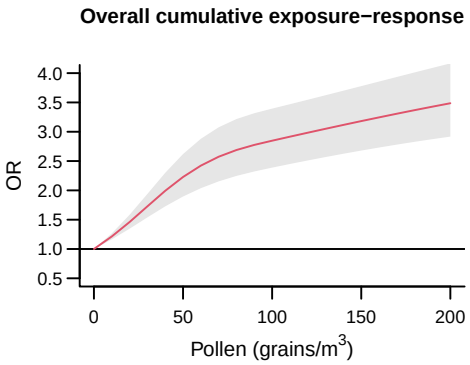
Stratification: subject/period risk sets

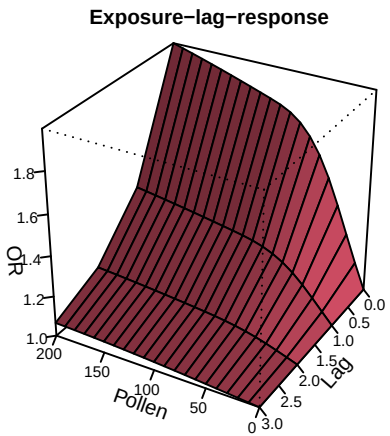
Additional temporal control: natural cubic spline of time with 8df/year

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

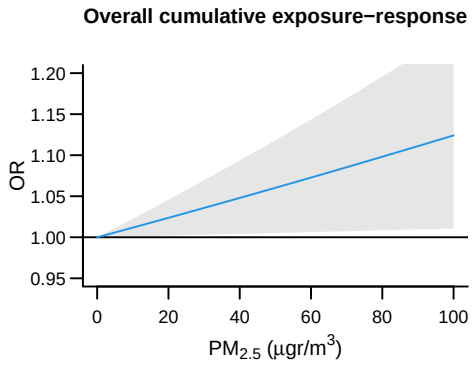
Results

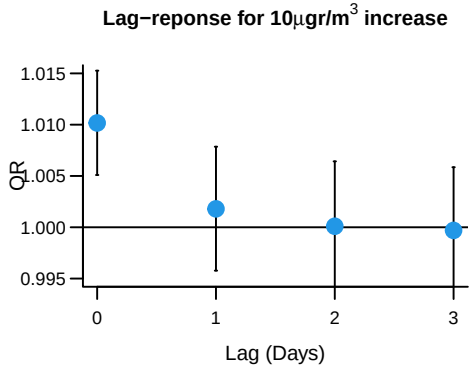
Pollen



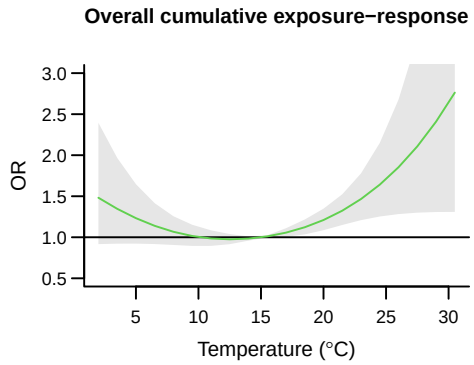


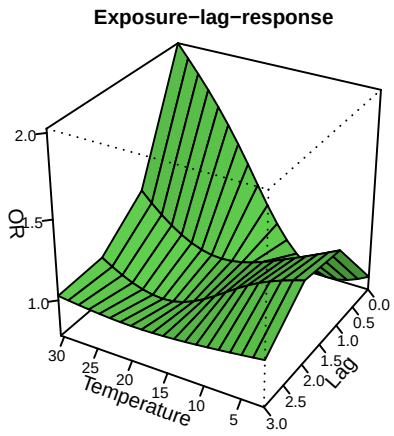
$PM_{2.5}$





Temperature





Computational issues

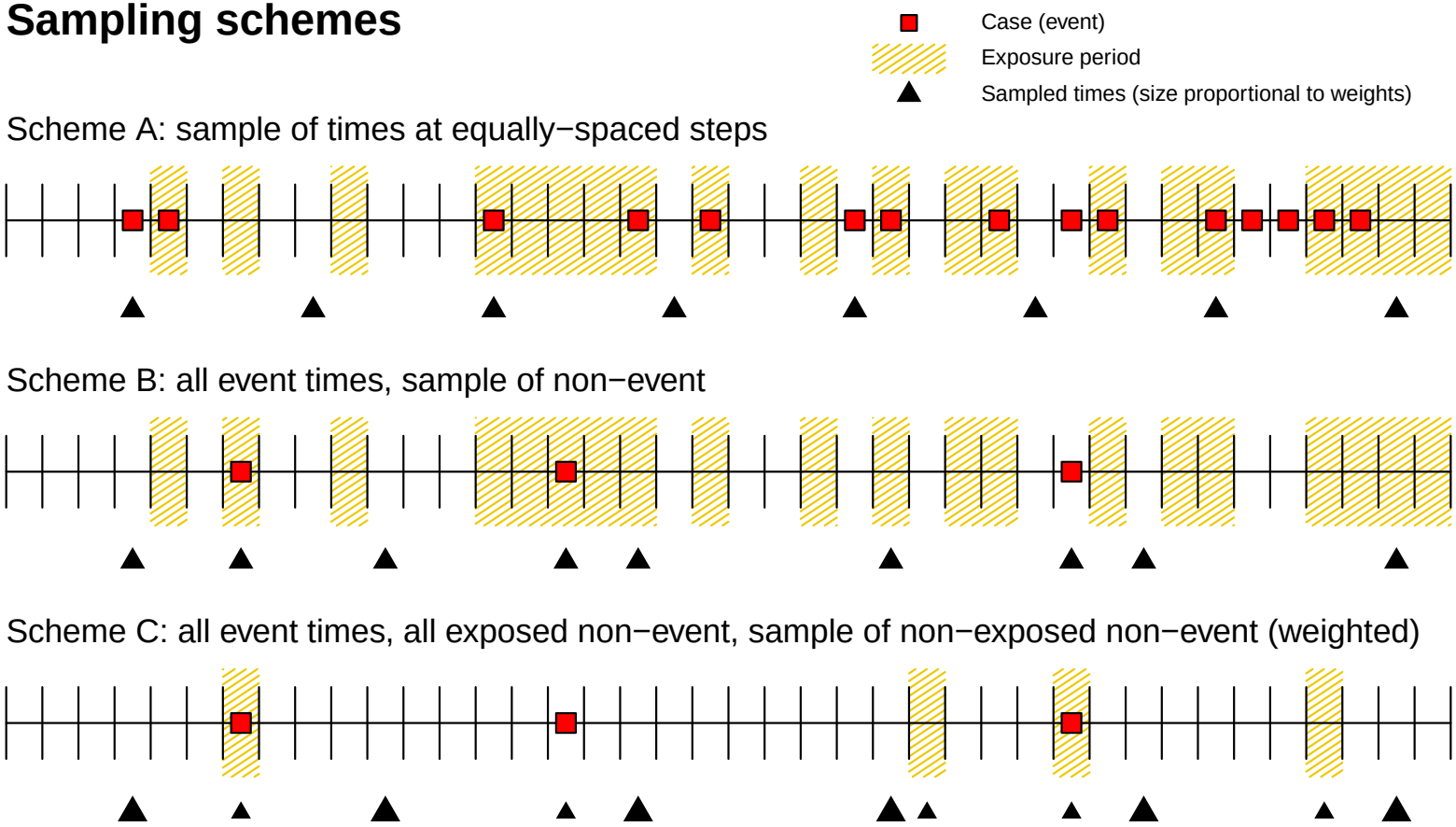
One of the problem of the case time series design is related to the significant **data expansion** due to the longitudinal splitting

In the first case study, data from 3,927 patients is expanded in individual daily series totalling **8,067,949 observations**

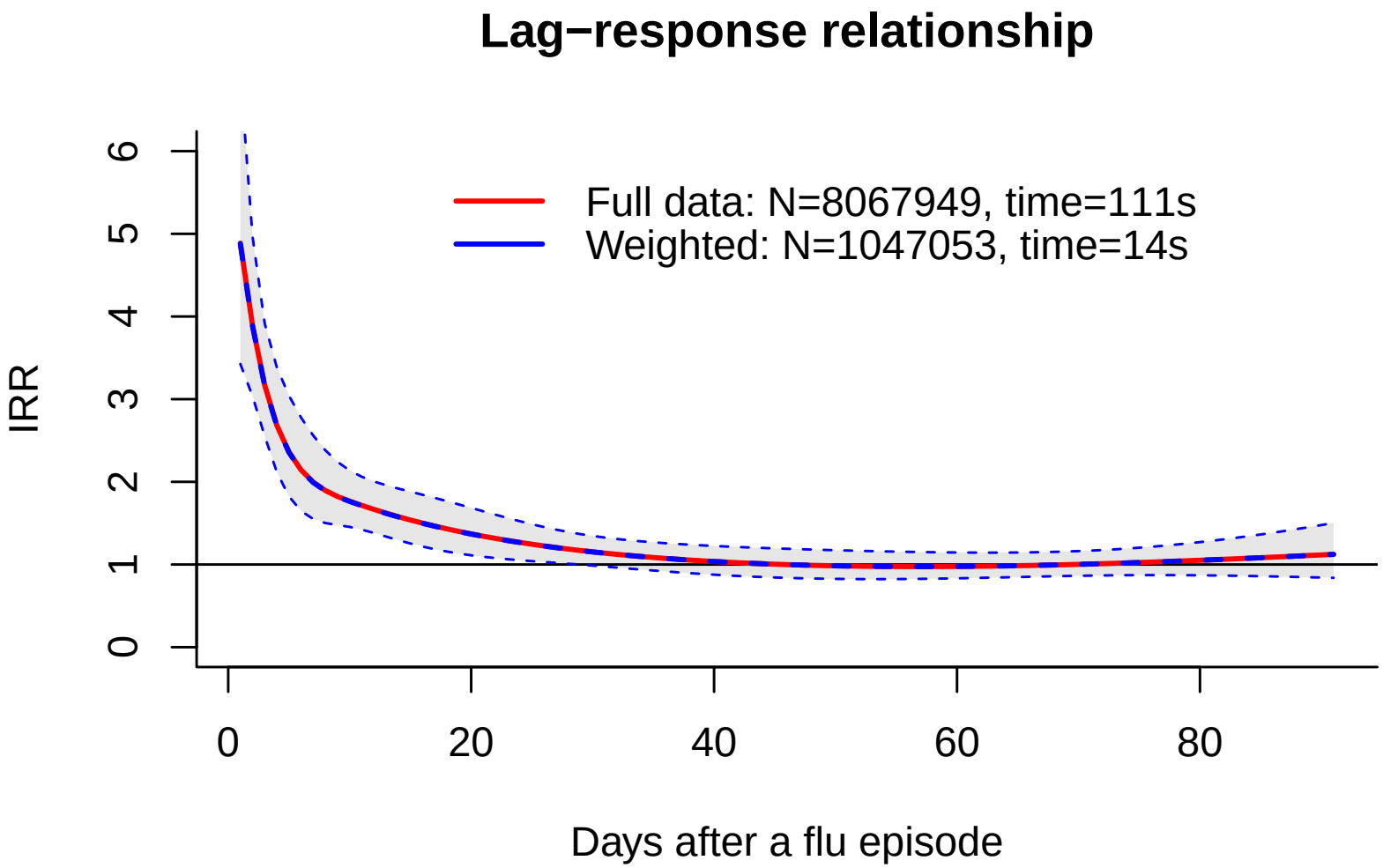
One solution is to apply **sampling schemes**: the aim is to maximize the reduction of the computational burden while minimizing the loss of statistical power

Sampling schemes

Sampling schemes



Comparison


Lag−response relationship
Full data: N=8067949, time=111s
Weighted: N=1047053, time=14s
IRR
0
1
2
3
4
5
6
0
20
40
60
80
Days after a flu episode

In conclusion

- **Adaptable framework**: combination of the flexibility of time series and design features of individual-level case-only methods
- **Generality**: applicable with intermittent or continuous exposures, and with event-type or continuous outcomes
- **Flexibility**: longitudinal structure provides setting for modelling non-linear/delayed effects and controlling for time-varying confounders
- **Wide applicability**: potential for investigating a broad range of health associations in different areas

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Example of R code

```
library(dlnm) ; library(gnm) ; library(splines)

splage <- onebasis(data$age, "ns", knots=c(60,80))
splseas <- onebasis(data$doy, "pbs", df=3)

cb <- crossbasis(exphist, lag=c(1,91), argvar=list("strata", breaks=0.5),
  arglag=list("ns", knots=c(3,10,29)))

model <- gnm(y ~ cb + splage + splseas, data=data, family=poisson,
  eliminate=factor(id))

cp <- crosspred(cbspl, model, at=1)

plot(cpspl, var=1, col=2, ylab="IRR of AMI", xlab="Days after flu",
  ylim=c(0,5), main="Risk by lag")
```

Links & references

Article

Gasparrini A. (2021). The case time series design. *Epidemiology*. 2021;2021;32(6)829-837.

Personal website & GitHub

`http://www.ag-myresearch.com/2021_gasparrini_epidemiol.html`
`https://github.com/gasparrini/CaseTimeSeries`

Contacts

Email: `antonio.gasparrini@lshtm.ac.uk`
Twitter: `@AGasparrini75`

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE