Flexible modelling of the cumulative effects of time-varying exposures

Applications in environmental, cancer and pharmaco-epidemiology

Antonio Gasparrini

Department of Medical Statistics
London School of Hygiene and Tropical Medicine (LSHTM)

Centre for Statistical Methodology – LSHTM
28 November 2014

Outline

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Temporal aspects

The relationship between a risk factor and the associated health effect always implies a **temporal dependency**: a common problem in biomedical research

This issue encompasses study designs and statistical model:

- Tobacco smoke and CVD risk
- Occupational exposure and incidence of cancer
- Drug intake and beneficial or side effects
- Short-term temperature variation and mortality

A topic (somewhat) **neglected** in methodological research

Previous research

Standard statistical approaches do not directly characterize this temporal structure

Challenge: modelling (potentially complex) temporal patterns of risk due to time-varying exposures

Models previously proposed in **cancer epidemiology** (Thomas 1988, Hauptmann 2000, Richardson 2009) and **pharmaco-epidemiology** (Abrahamowicz 2012)

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Limitations

Incomplete statistical development: e.g. no measures of uncertainty

Poor software implementation: ad-hoc routines, computational issues, convergence problems

Lack of a **consistent conceptual and interpretational framework**

Distributed lag models

DLMs proposed by Almon (Econometrica 1965) in **econometrics** for time series data, then applied in **environmental epidemiology** by Schwartz (Epidemiology 2000).

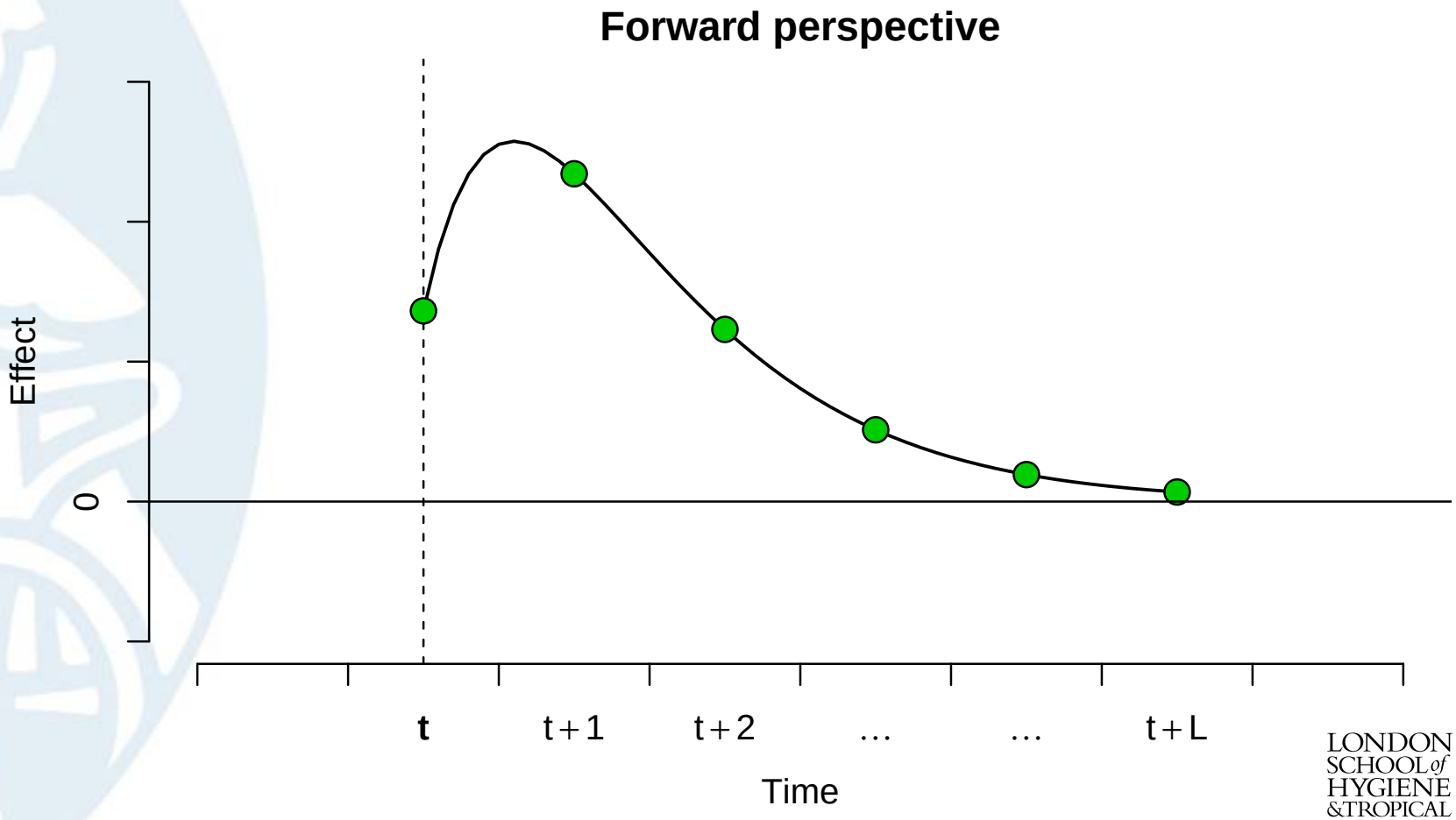
Armstrong (Epidemiology 2006) extended them to **distributed lag non-linear models** (DLNMs), applicable to non-linear exposure-response associations

A far more developed statistical framework, but **only applicable** to time series data

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Conceptual representation

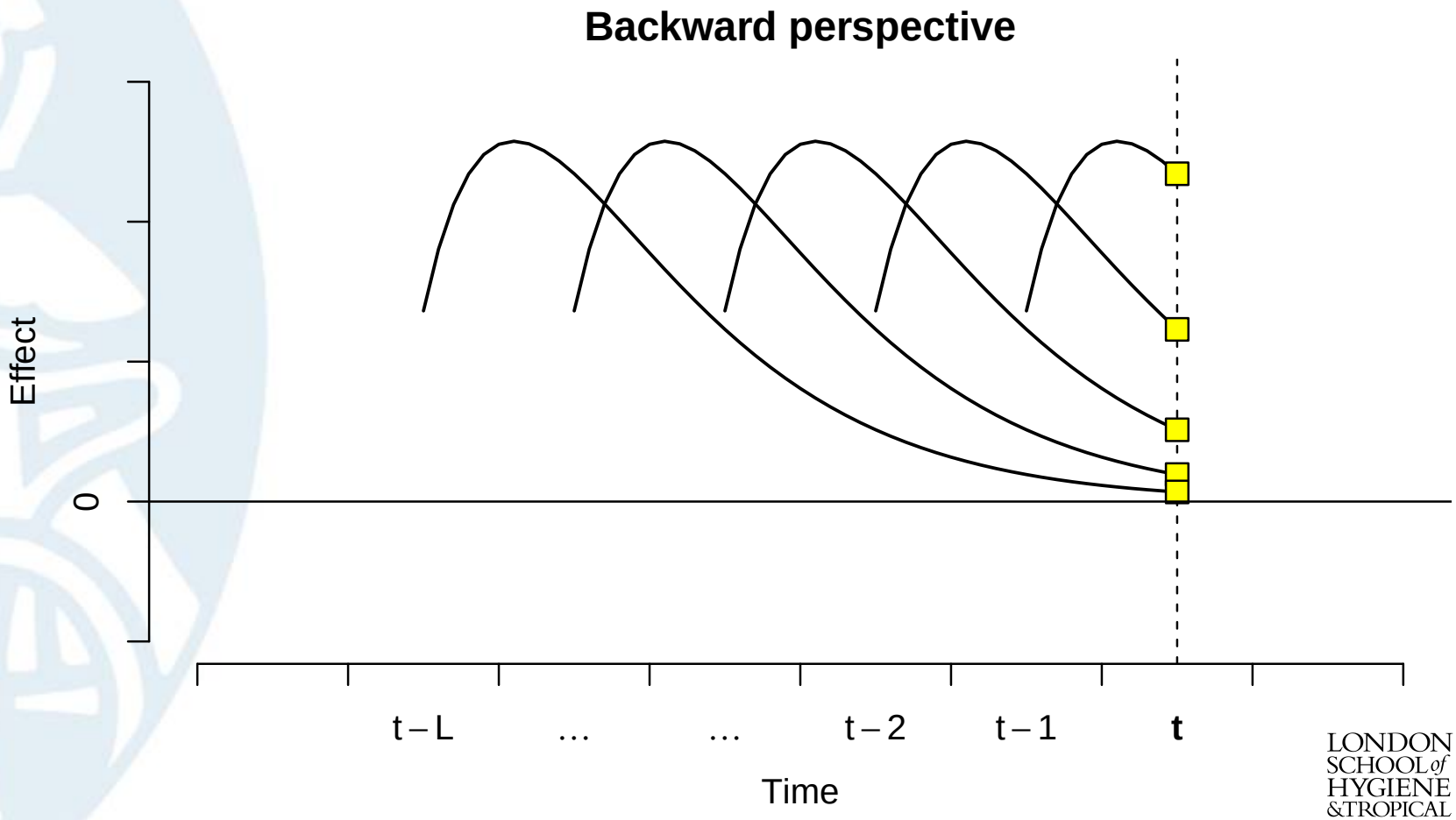
Single exposure event



LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Conceptual representation

Multiple exposure events



Assumptions

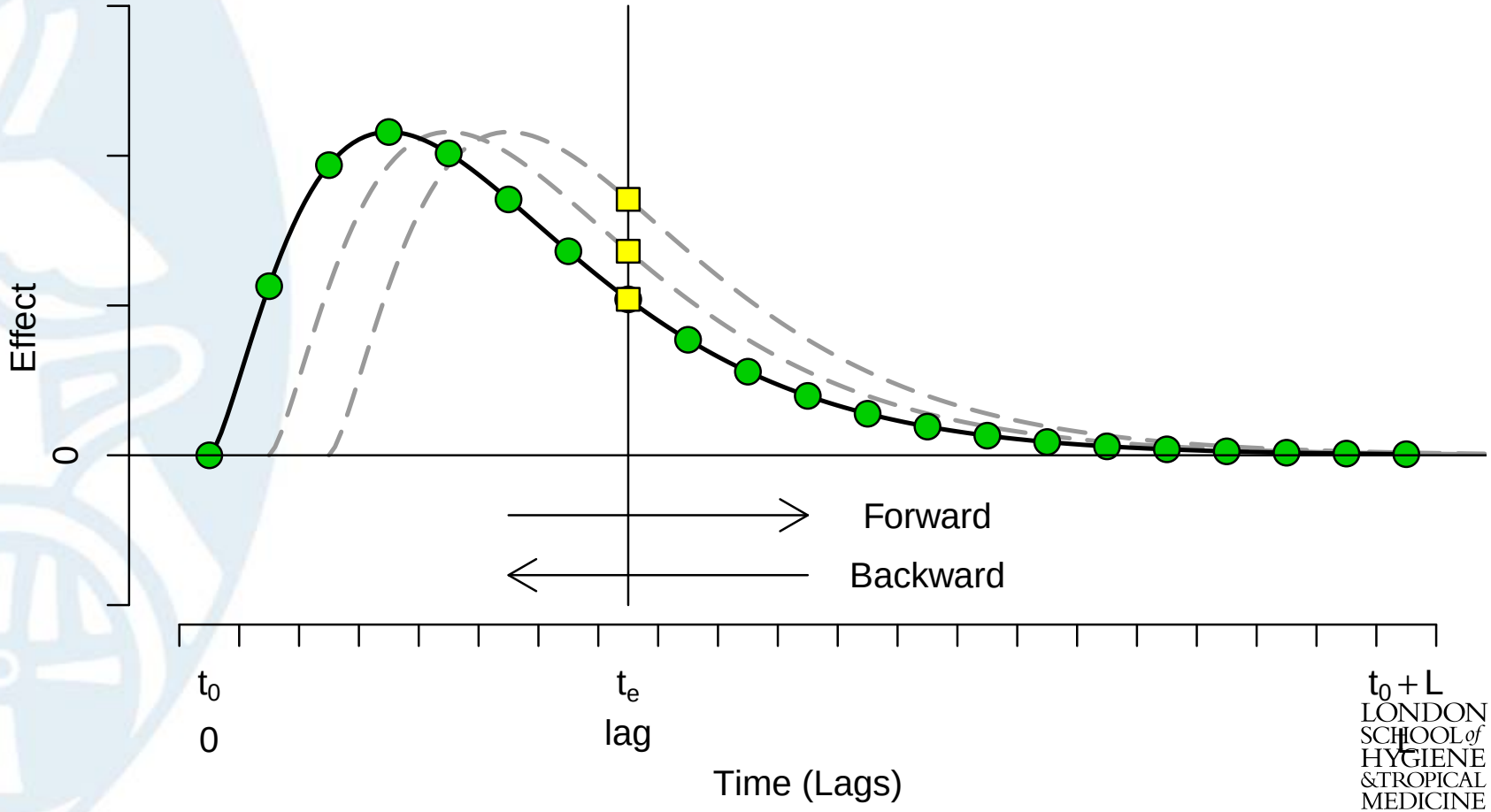
Under specific assumptions, these two perspectives can be merged together:

- assumption of **identical effects**
- (fundamental) assumption of **independency**

These conditions underpin the **conceptual framework** for defining and modelling DLNMs

Conceptual representation

New lag dimension



Exposure-lag-response associations

The risk is represented by a function $s(x_{t-\ell}, \ldots, x_{t-L})$ defined in terms of both **intensity** and **timing** of a series of past exposures, expressed through:

- an **exposure-response** function $f(x)$ for exposure x
- a **lag-response** function $w(\ell)$ for lag ℓ

Generating a bi-dimensional **exposure-lag-response** function $f \cdot w(x, \ell)$, whose integral provides:

$$s(x_{t-\ell}, \ldots, x_{t-L}) = \int_{\ell_0}^{L} f \cdot w(x_{t-\ell}, \ell)\, d\ell \approx \sum_{\ell=\ell_0}^{L} f \cdot w(x_{t-\ell}, \ell)$$

Distributed lag models (DLMs)

Given a **exposure history** at time t for lags $\ell = \ell_0, \ldots, L$:

$$\mathbf{q}_{x_t} = [x_{t-\ell_0}, \ldots, x_{t-\ell}, \ldots, x_{t-L}]^\top$$

and assuming a **linear exposure-response**, we can write:

$$s(\mathbf{q}_{x_t}; \eta) = \mathbf{q}_{x_t}^\top \mathbf{C} \eta = \mathbf{w}_{x_t}^\top \eta$$

where $\mathbf{C}$ is obtained from the lag vector $\boldsymbol{\ell} = [\ell_0, \ldots, \ell, \ldots, L]^\top$ by applying a specific **basis transformation**

Distributed lag non-linear models (DLNMs)

First the matrix $\mathbf{R}_{x_t}$ is obtained applying a **second basis transformation** to $\mathbf{q}_{x_t}$

Then we define a tensor product:

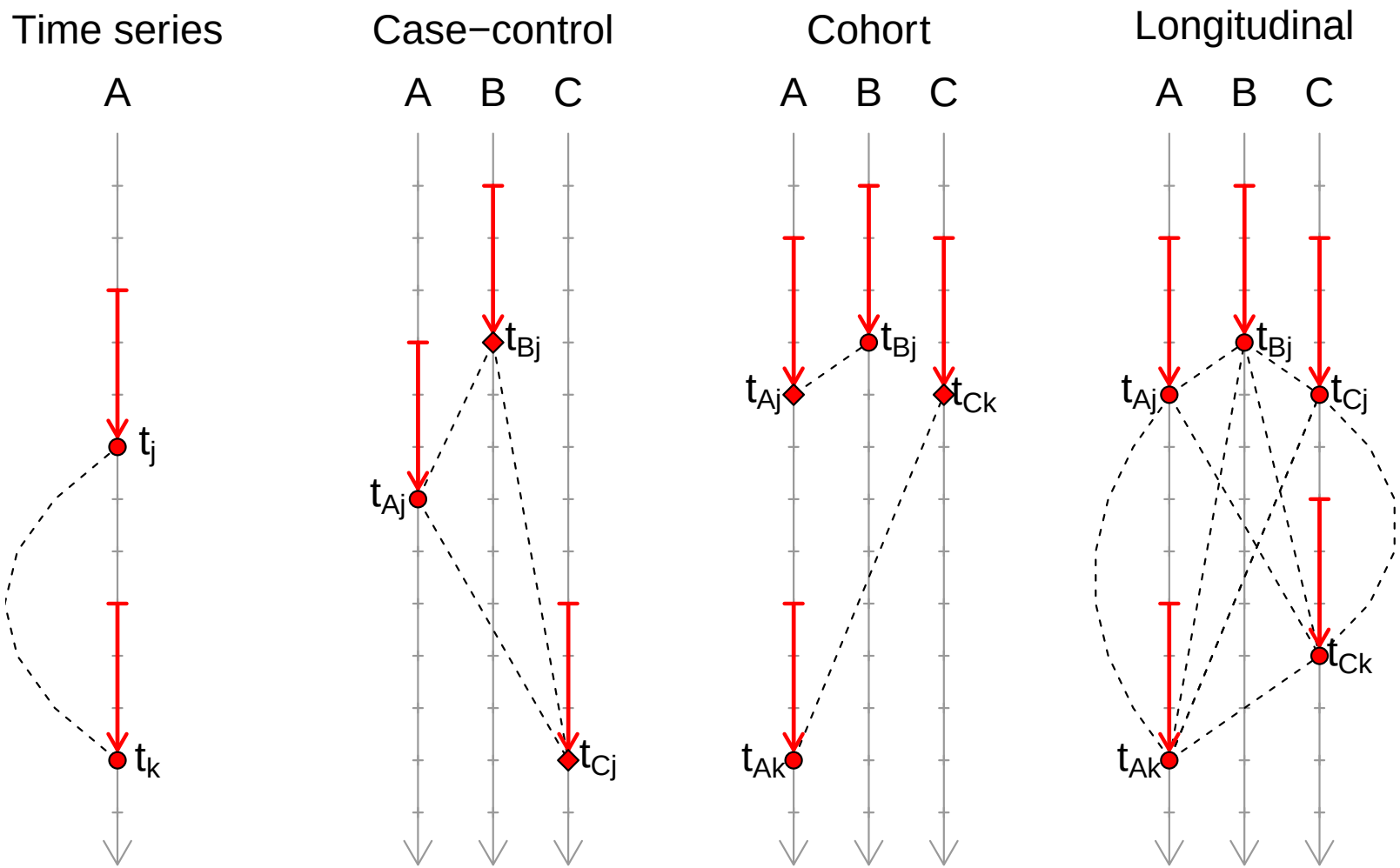
$$\mathbf{A}_{x_t} = (\mathbf{1}_{v_\ell}^\top \otimes \mathbf{R}_{x_t}) \odot (\mathbf{C} \otimes \mathbf{1}_{v_x}^\top)$$

which forms the **crossbasis**:

$$s(\mathbf{q}_{x_t}; \eta) = (\mathbf{1}_{v_x \cdot v_\ell}^\top \mathbf{A}_{x_t})\eta = \mathbf{w}_{x_t}^\top \eta$$

The problem reduces to choosing a basis for each $\mathbf{q}_{x_t}$ and ℓ, defining **exposure-response** and **lag-response functions**, respectively

Alternative study designs



First example

Temperature and all-cause mortality

Research area where DLNMs were **originally proposed**

Time series data with daily death counts and temperature measurements between 1st Jan 1993 and 31st Dec 2006 in London (845,215 deaths in total)

In this setting, exposure histories are simply derived by **'lagging'** the temperature series

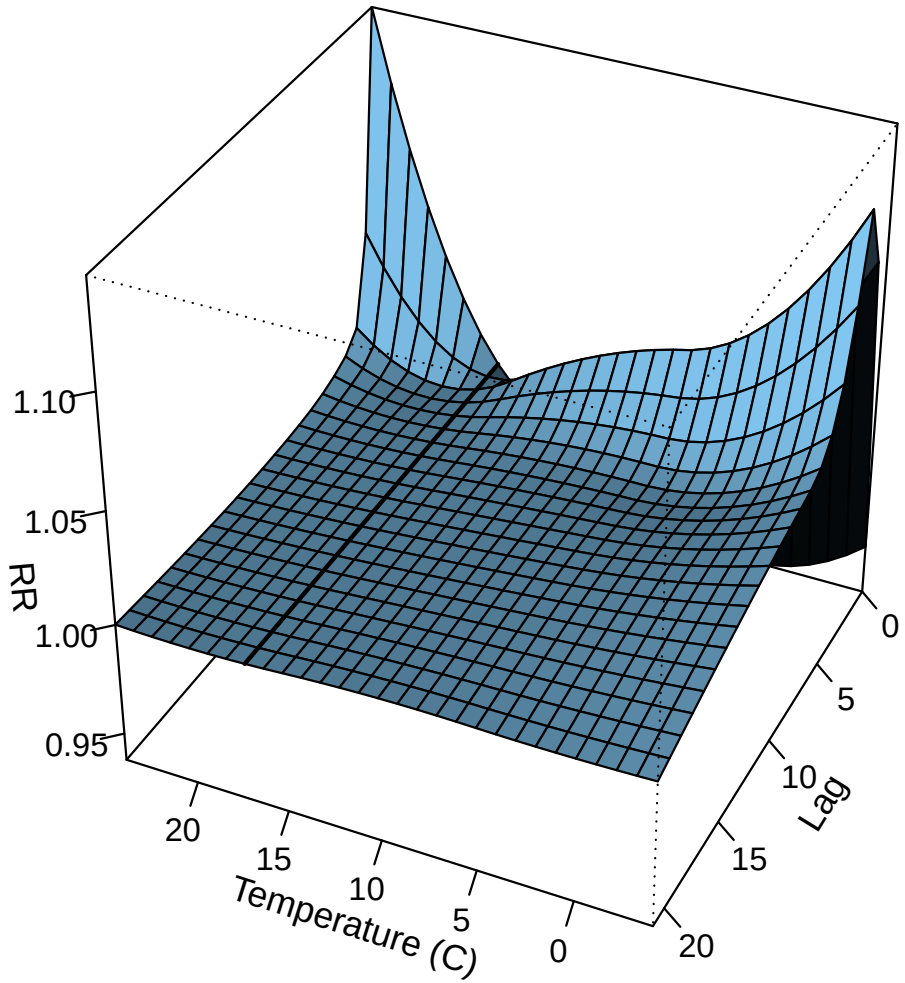
Quasi-Poisson GLM

Analysis with a generalized linear model with quasi-Poisson family, controlling for trends and day of the week

$$\log(\mu_t) = \alpha + s_x(\mathbf{q}_{x_t}; \boldsymbol{\eta}) + \sum_{p=1}^{P} s_z(z_t; \boldsymbol{\beta}_z)$$

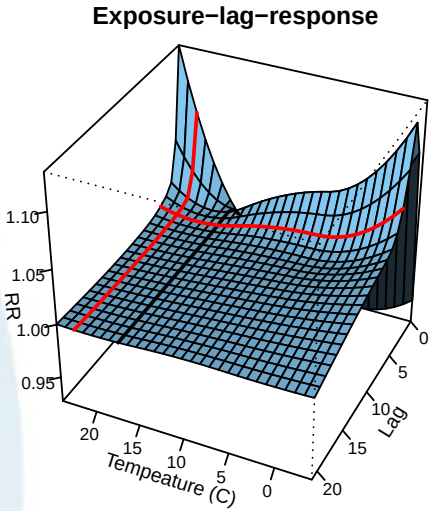
Here **spline functions** used to specify both $f(x)$ and $w(\ell)$

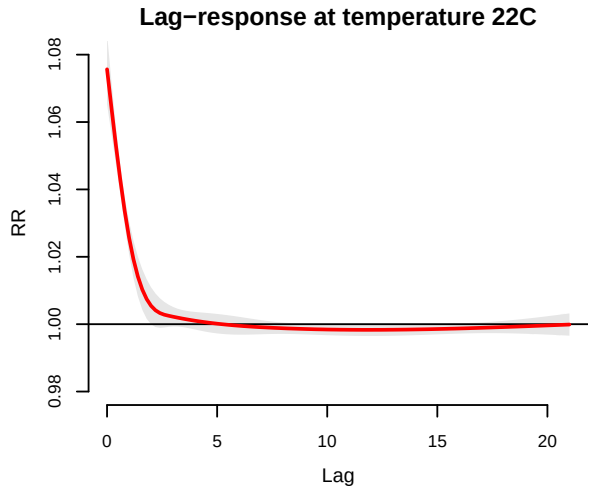
Exposure-lag-response

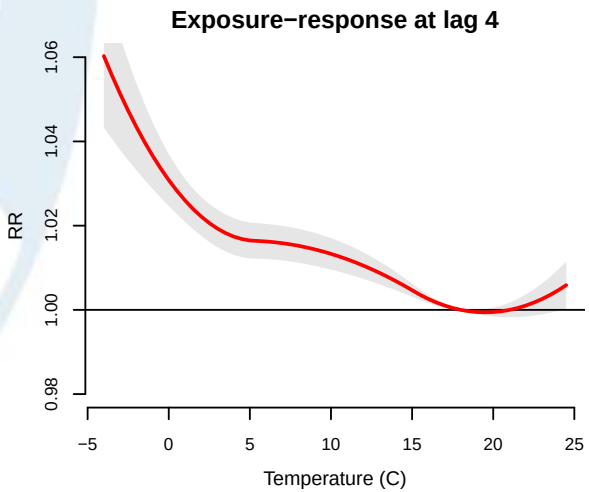


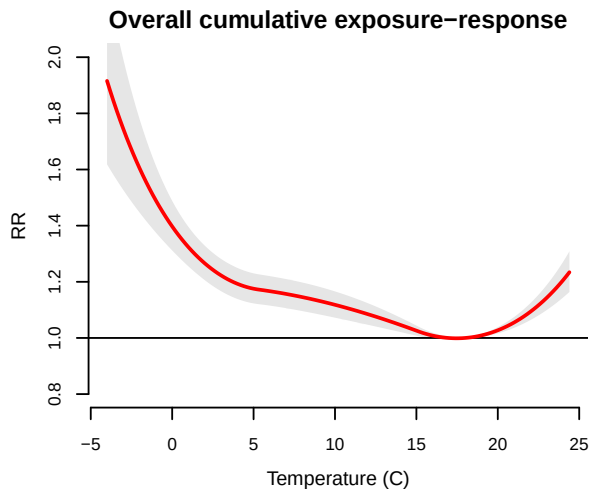
LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Summaries









LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Second example

Radon exposure and lung cancer mortality

3,347 subjects working in the Colorado Plateau mines between 1950–1960, **258 lung cancer deaths**

Yearly exposure history to **radon** (WLM) and **tobacco smoke** (pack×100) reconstructed from 5-year age periods

Proportional hazard model

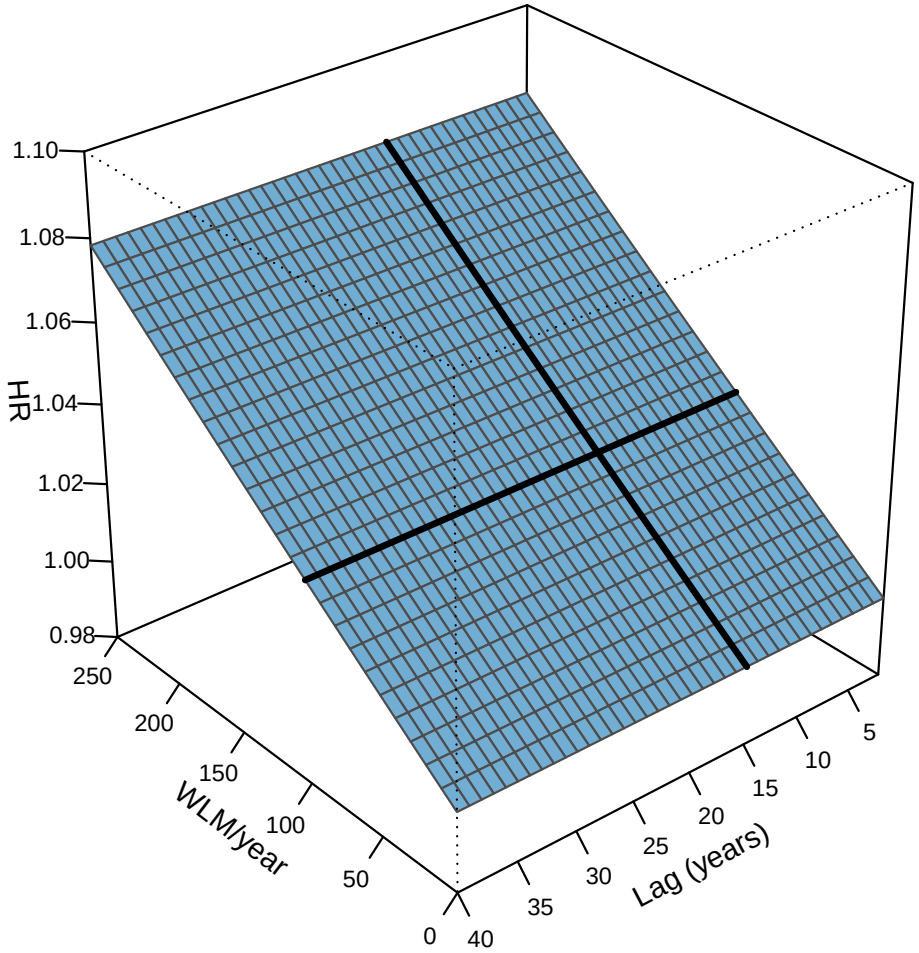
Analysis with Cox proportional hazards model using age as time axis, controlling for smoking and calendar year. For subject i:

$$\log\left[h(it)\right] = \log\left[h_0(t)\right] + s_x(\mathbf{q}_{x_{it}}; \boldsymbol{\eta}_x) + s_z(\mathbf{q}_{z_{it}}; \boldsymbol{\eta}_z) + \gamma u_{it}$$

Different functions used to specify $f(x)$ and $w(\ell)$: constant, piecewise constant, quadratic B-spline

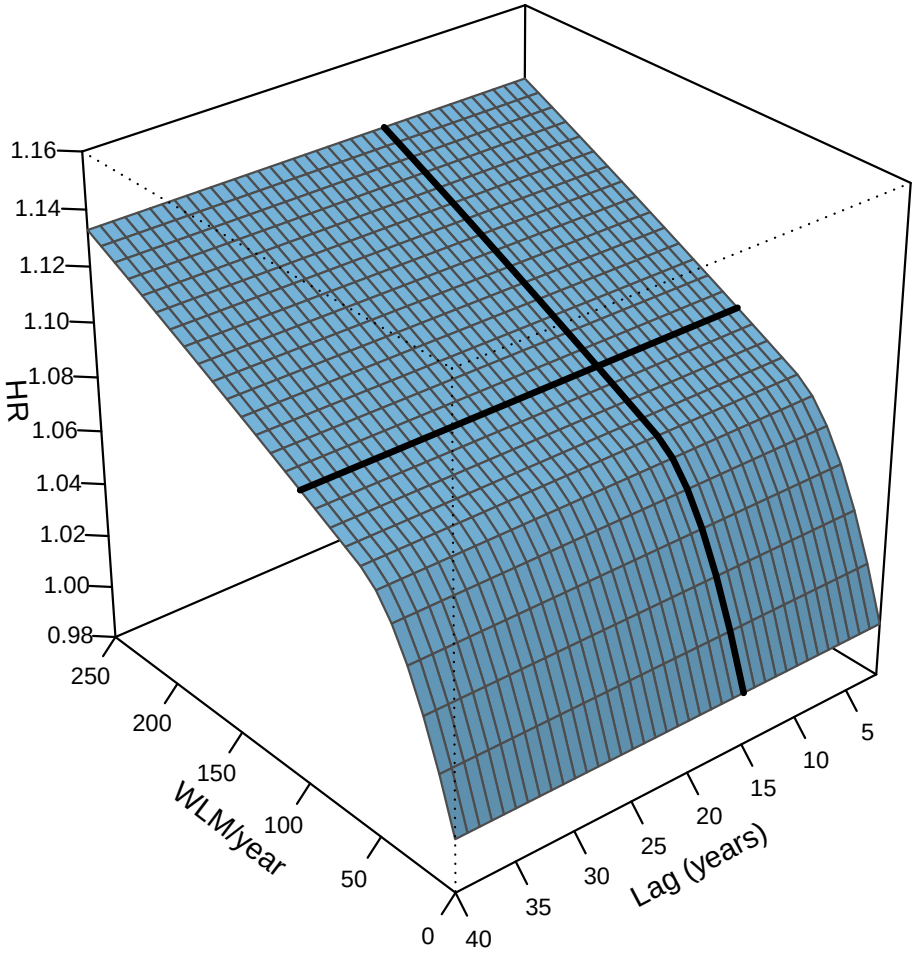
Exposure-lag-response

Linear-by-constant



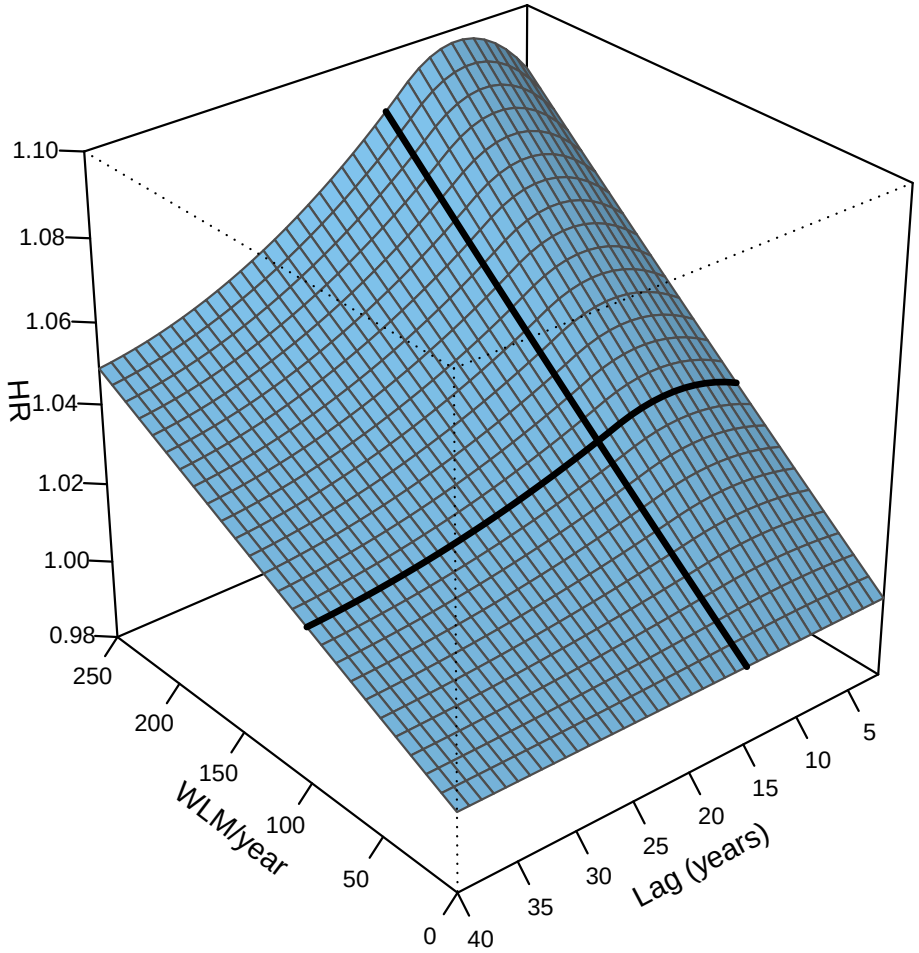
Exposure-lag-response

Spline-by-constant



Exposure-lag-response

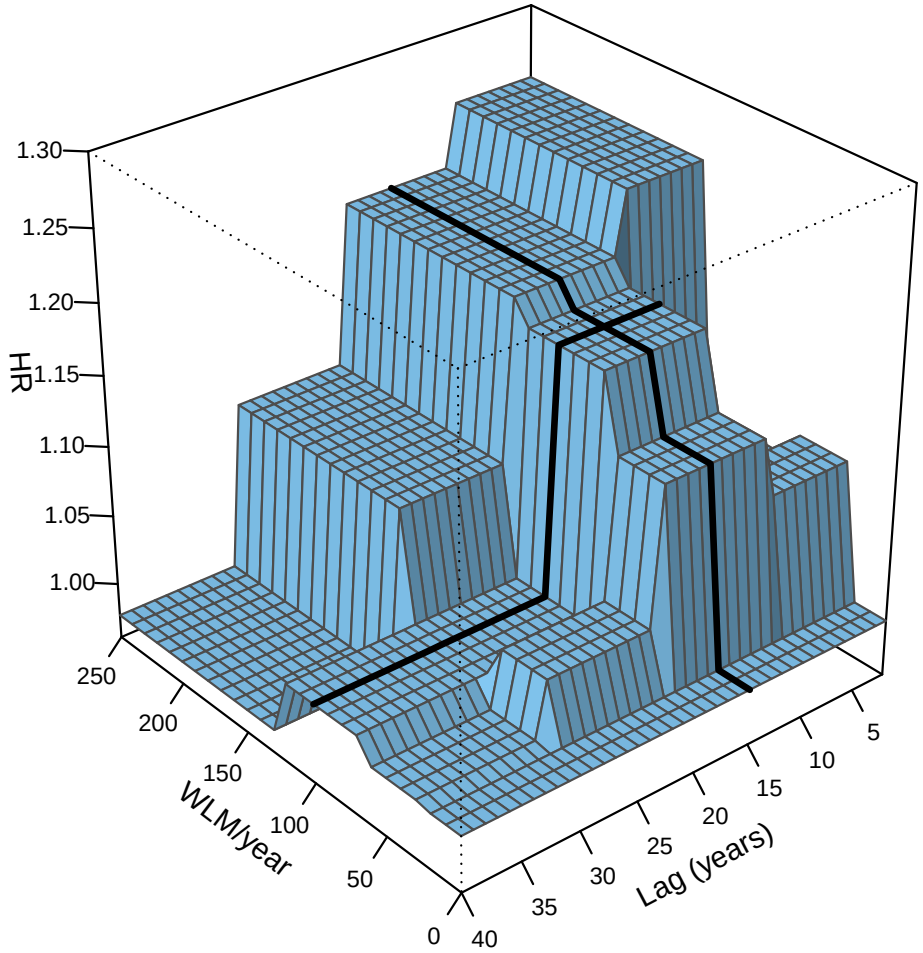
Linear-by-spline



LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Exposure-lag-response

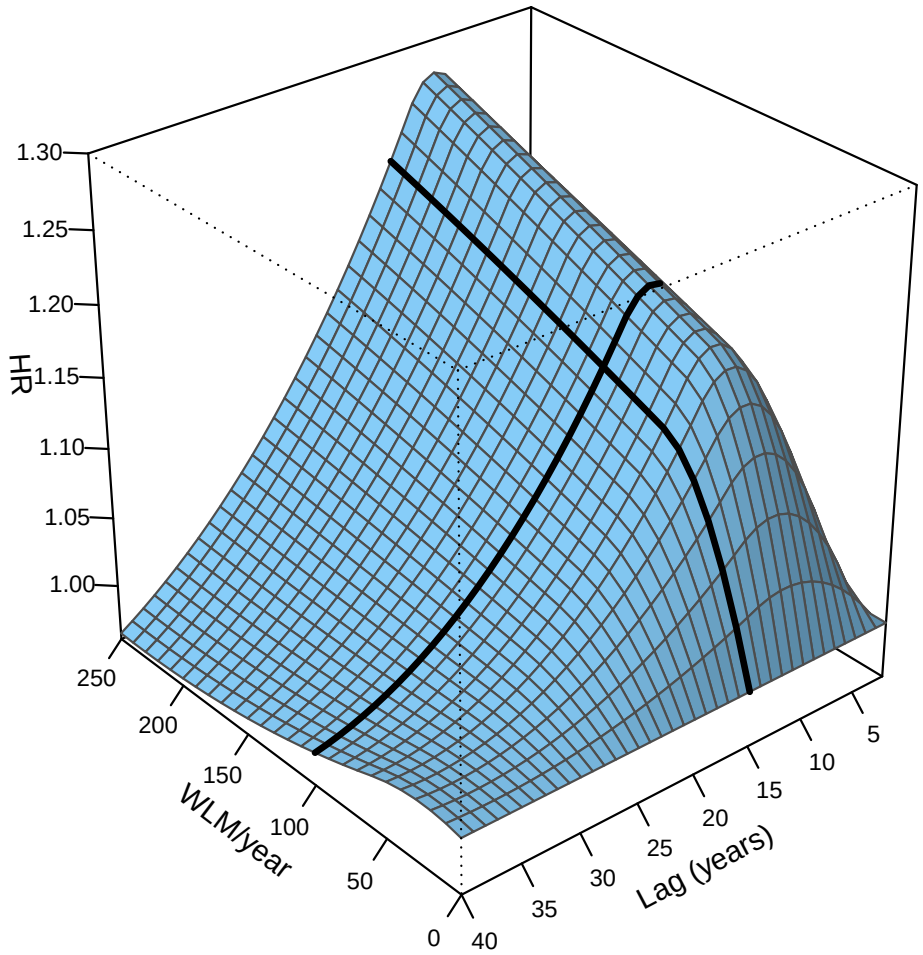
Step-by-step



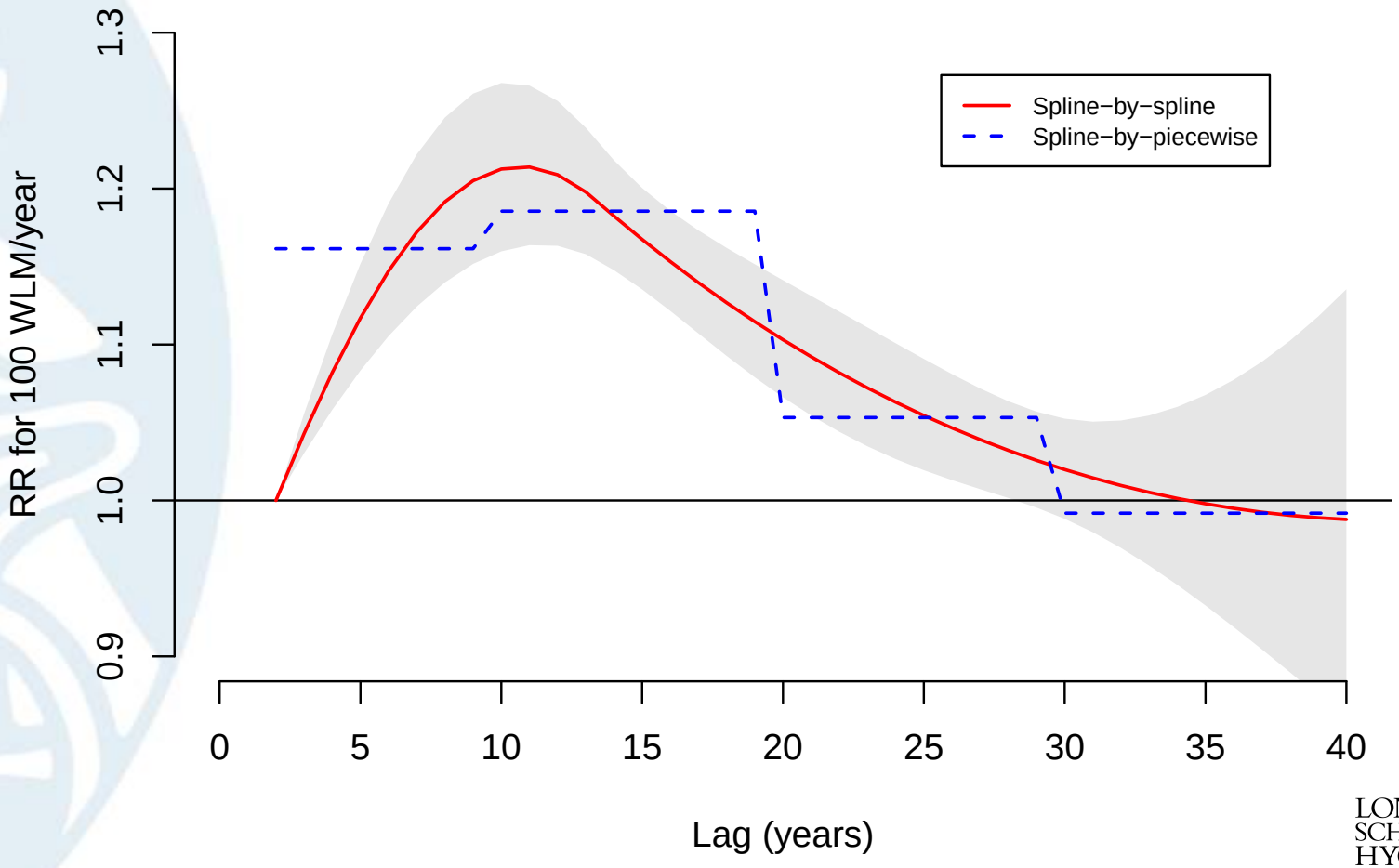
LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Exposure-lag-response

Spline-by-spline

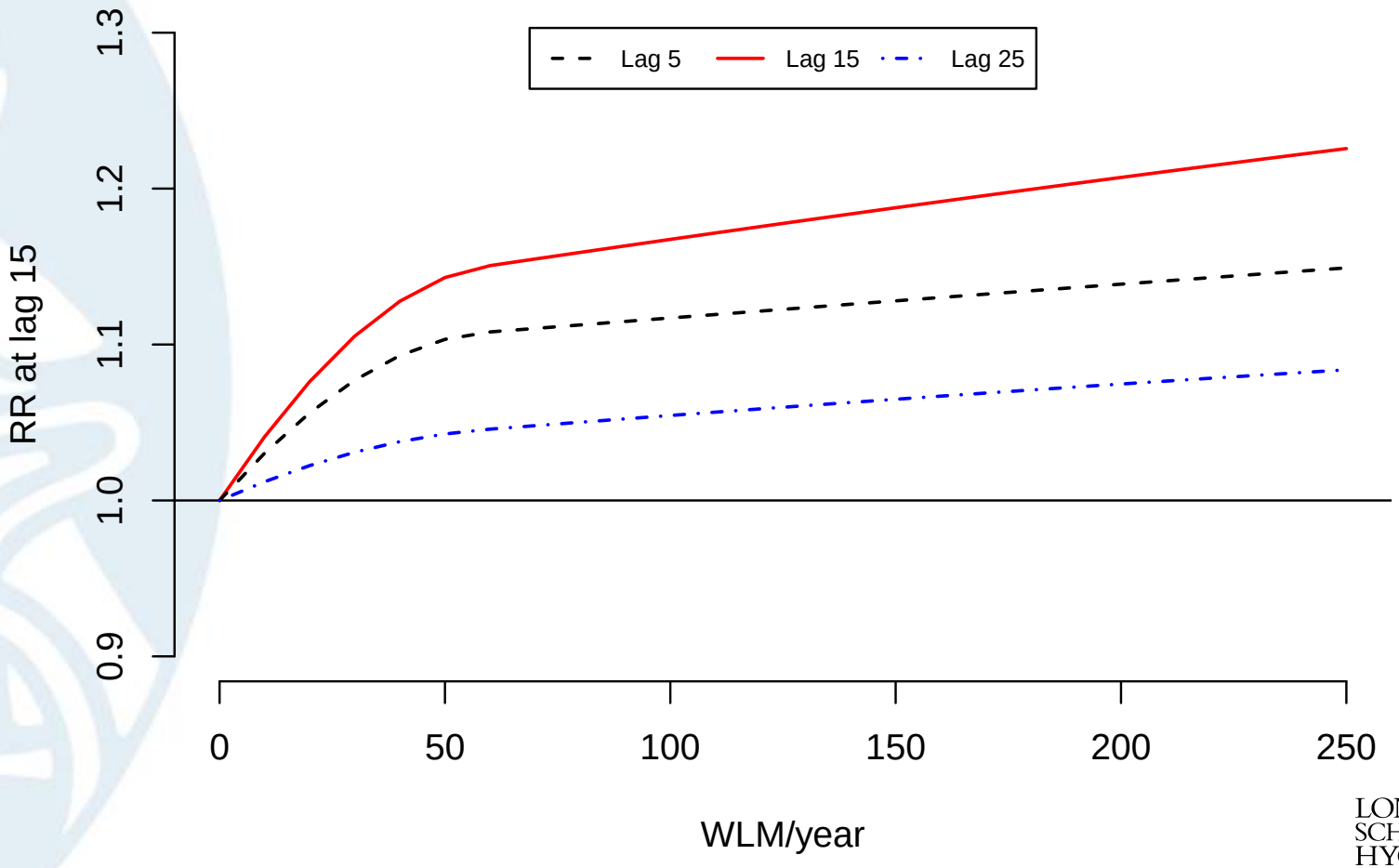


Lag-response curves from DLNMs



LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Exposure-responses at different lags



Third example

MMR vaccine and ITP risk

Data from **35** children receiving the MMR (measles, mumps, rubella) vaccine months and admitted to the hospital for idiopathic trombocytopenic purpura (ITS) within 12-24 months of age.

Replicating and extending a previous analysis using the **self-controlled case series** design (Whitaker 2006)

Conditional Poisson regression

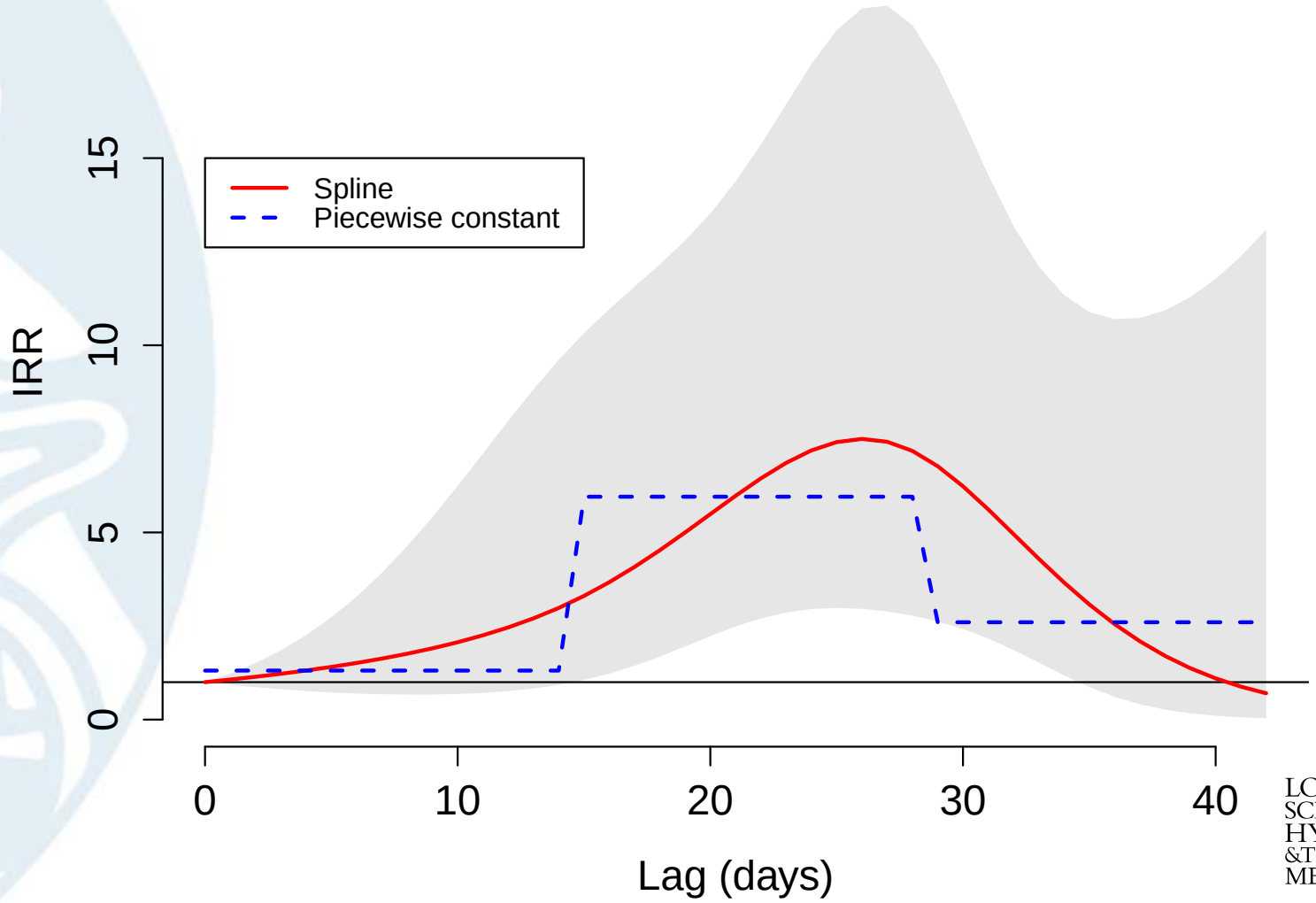
Analysis with conditional Poisson regression controlling for age. For subject i at age a:

$$\log(\lambda_{iat}) = \alpha_i + s_x(\mathbf{q}_{x_{it}}; \boldsymbol{\eta}_x) + f(a_{it}; \boldsymbol{\gamma})$$

Single exposure event modelled with a binary variable

Exposure-response assumed linear, **lag-response** modelled with spline or piecewise constant functions

Lag-response



LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Fourth example

Tobacco and lung cancer incidence

1,479 cases and **1,918** controls from three case-control studies within the Synergy network

Yearly exposure history to **tobacco smoke** (cigarette/day) reconstructed from questionnaires

Logistic regression

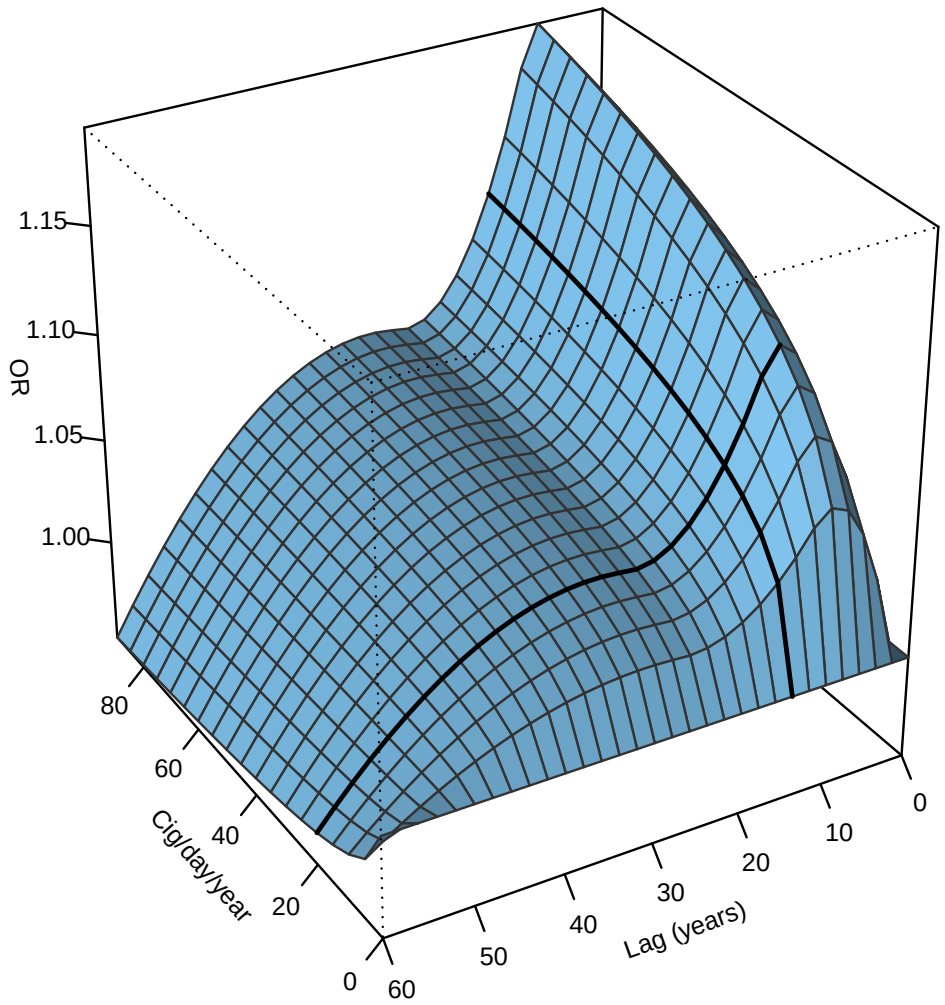
Analysis with logistic regression controlling for sex

$$\text{logit}\left(\mu_i\right) = \alpha + s_x(\mathbf{q}_{x_i}; \boldsymbol{\eta}_x) + \gamma u_i$$

Different functions used to specify $f(x)$ and $w(\ell)$: log, piecewise constant, quadratic B-spline

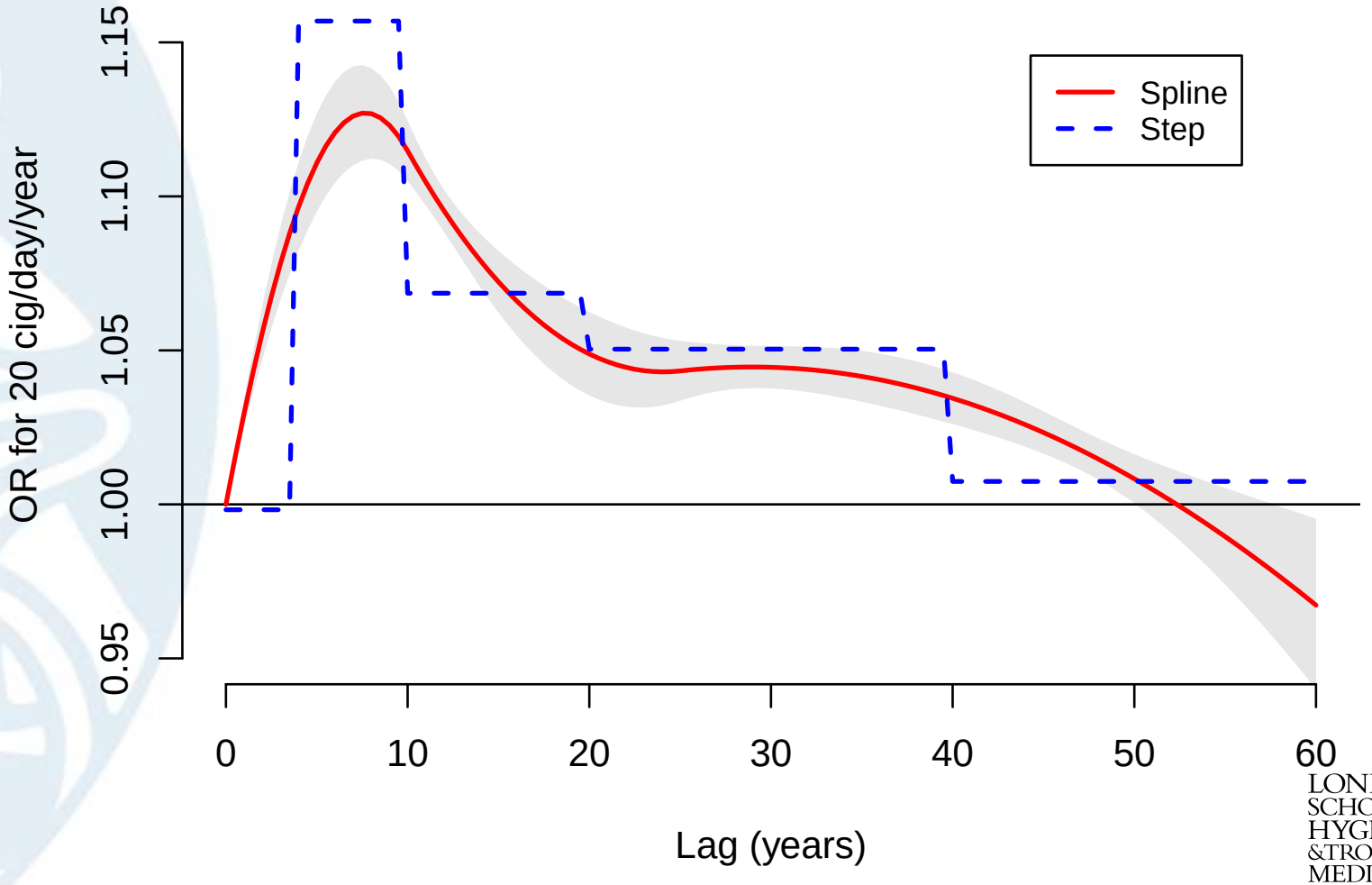
Exposure-lag-response

Log-by-spline



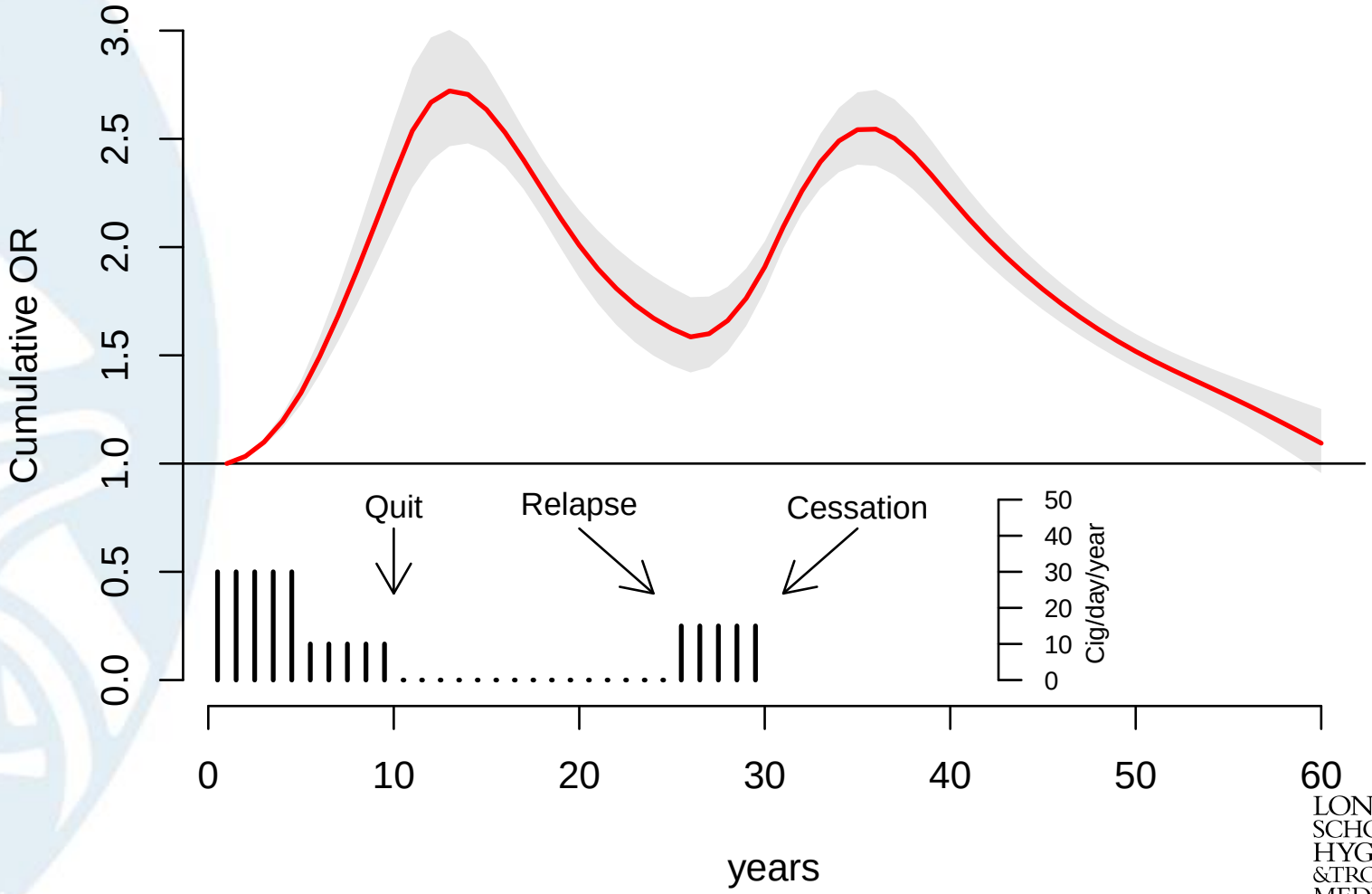
LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Lag-response curves from DLNMs



LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Dynamic prediction of risk



LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Fifth example

Trial on the effect of a drug

50 subjects followed for 4 weeks

Time-varying treatment randomly allocated in two of the four weeks, each with a different dose selected at random

Outcome measured at the end of the 28 days

Linear regression

Analysis with linear regression controlling for sex

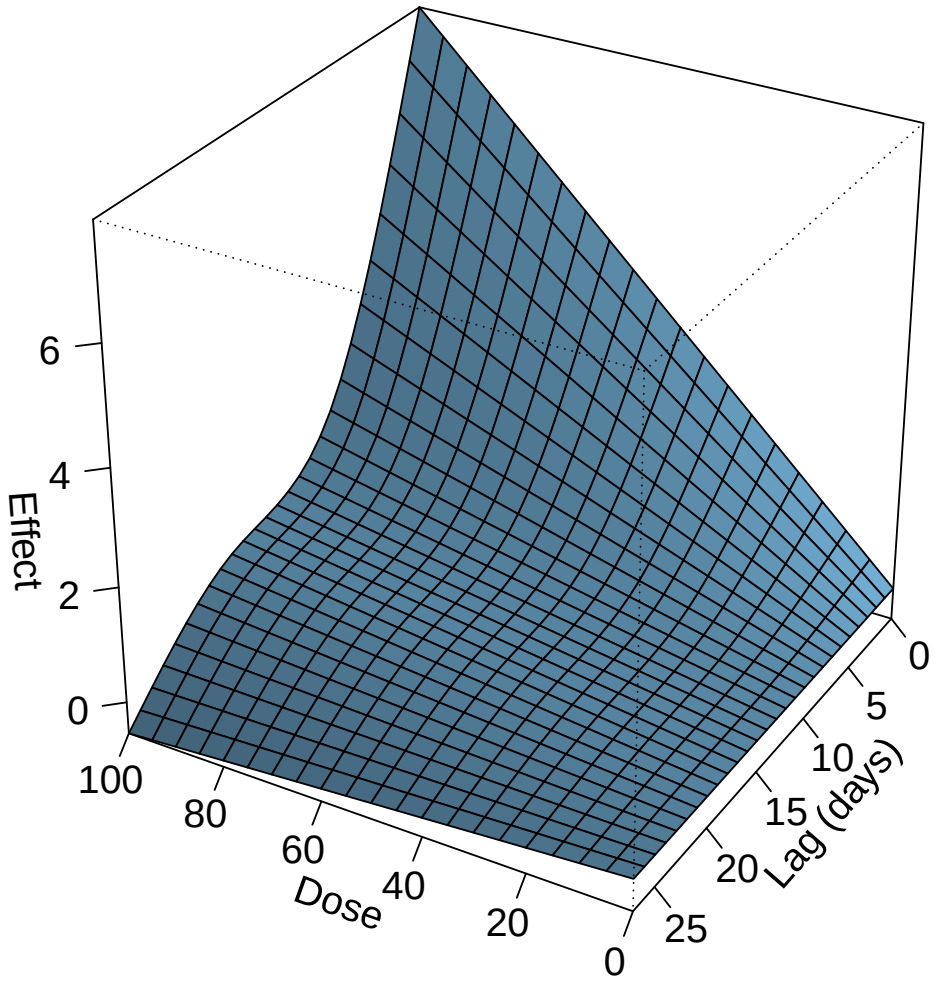
$$y_i = \alpha + s_x(\mathbf{q}_{x_i}; \boldsymbol{\eta}_x) + \gamma u_i + \epsilon_i$$

Exposure-response assumed linear

Lag-response modelled with spline or decay functions

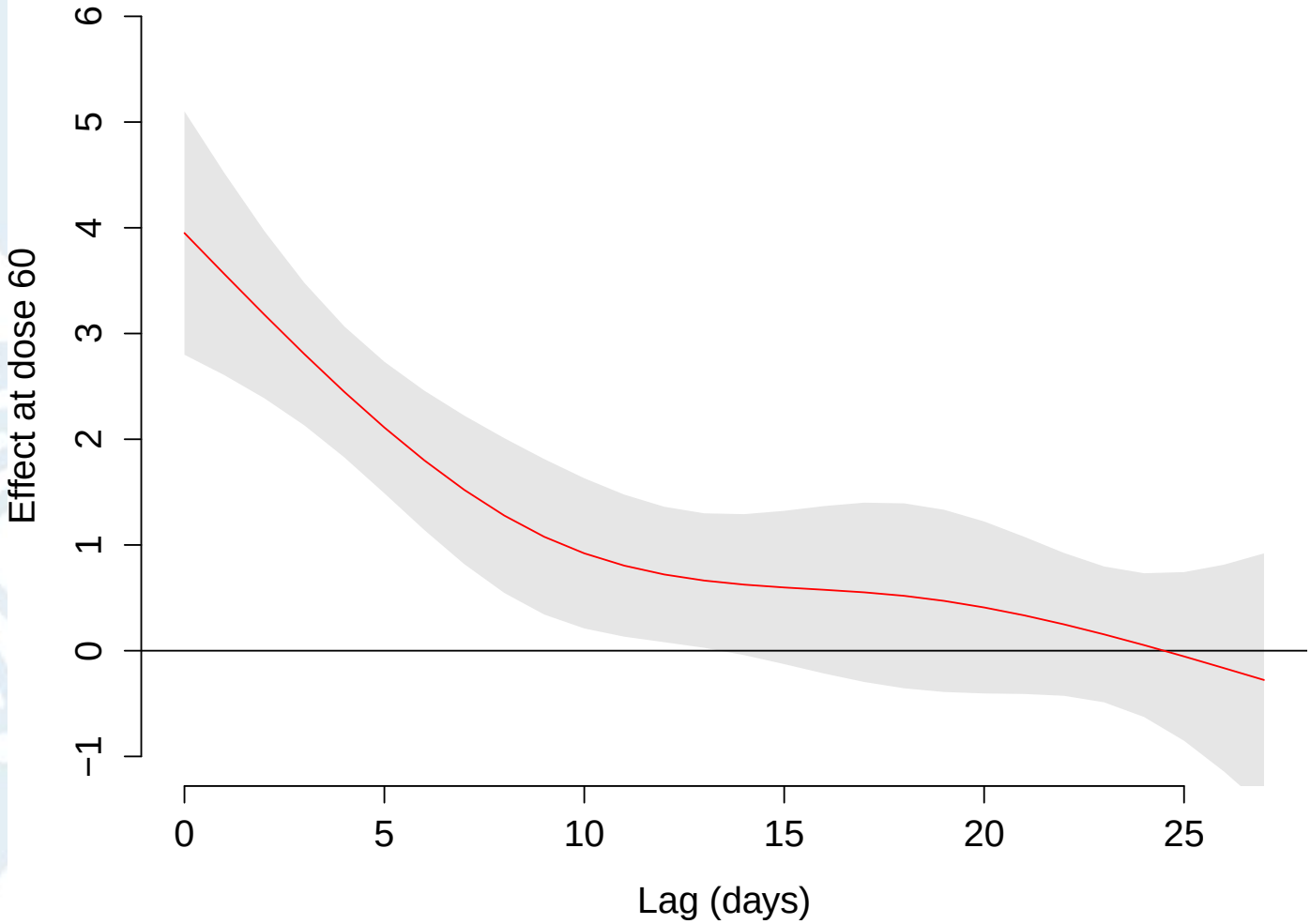
Exposure-lag-response

linear-by-spline



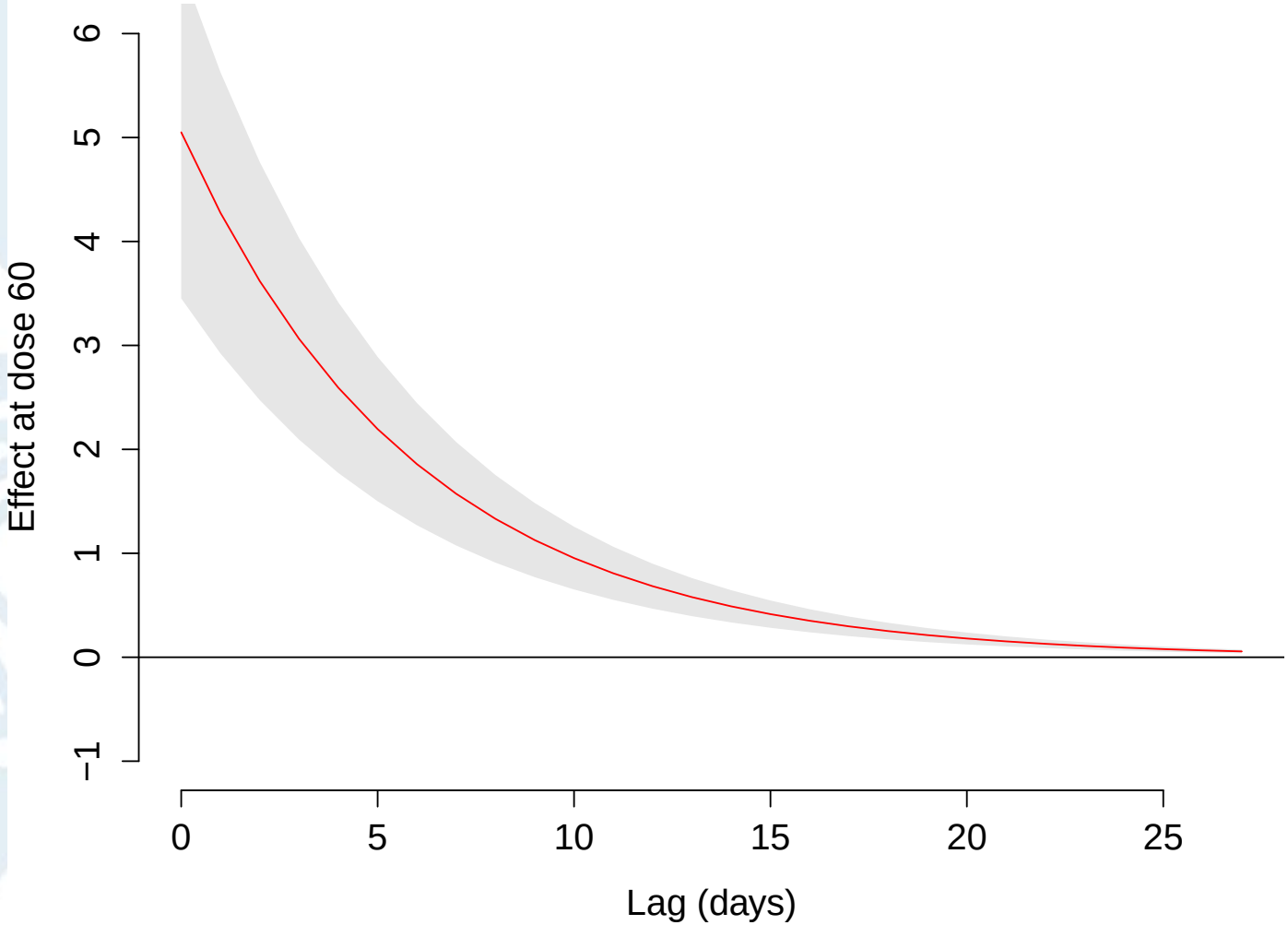
Lag-response

Spline function



Lag-response

Decay function



Software implementation

The framework is **fully implemented** in the R package `dlnm`, available from the CRAN (Gasparrini *JSS* 2011)

The package contains a **new vignette** focusing on applications beyond time series data

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

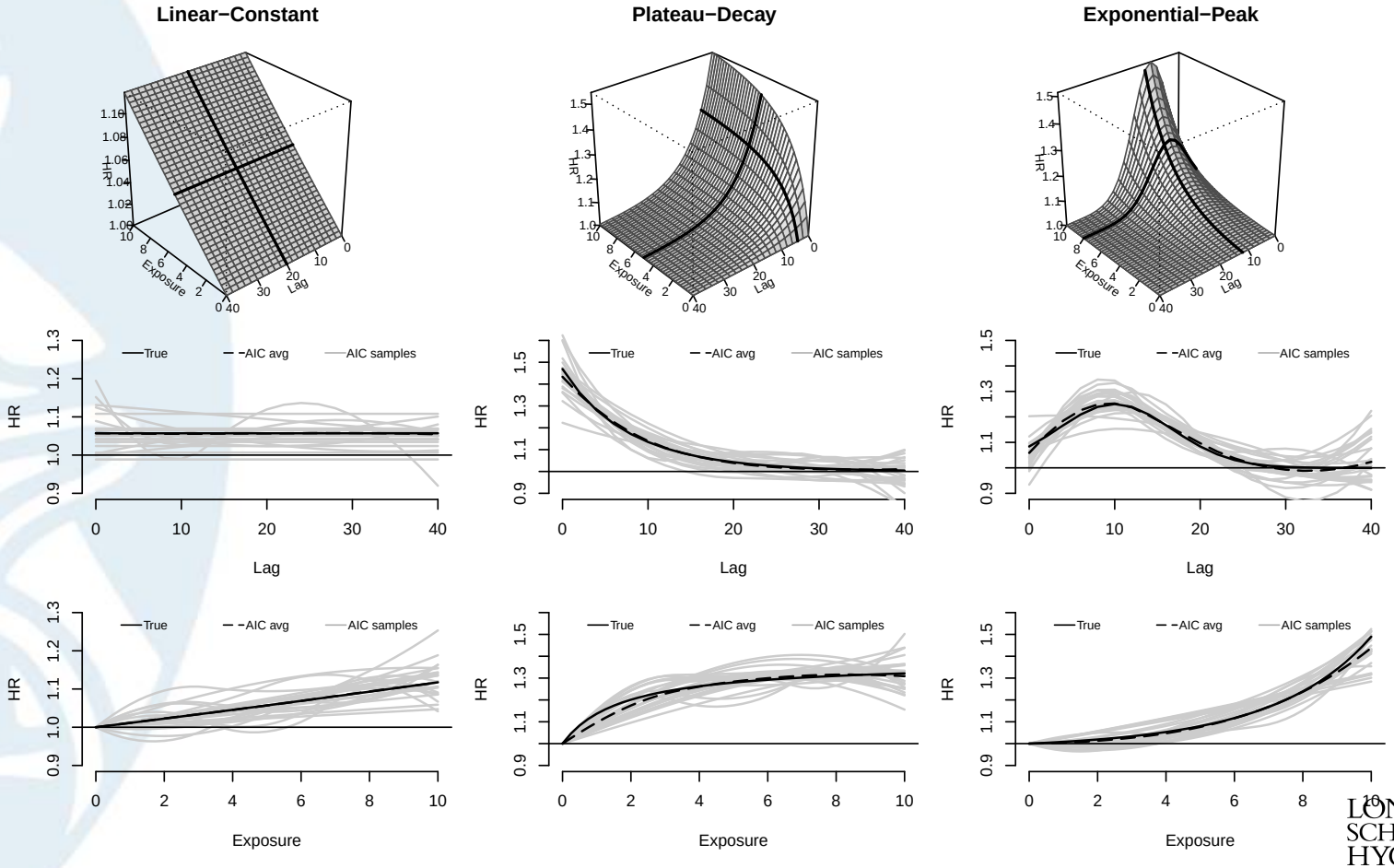
The R package dlnm

Example of code

```
library(dlnm)

cb <- crossbasis(Q,lag=c(2,40),
  argvar=list(fun="bs",degree=2,knots=59.4,cen=0),
  arglag=list(fun="bs",degree=2,knots=13.3,int=F))

model <- coxph(Surv(agest,ageexit,ind)~cb+smoke+caltime,data)

pred <- crosspred(cb,model,at=0:25*10)

plot(pred,"3d",xlab="WLM/year",ylab="Lag (years)",zlab="RR")
plot(pred,var=100,xlab="Lag (years)",ylab="RR")
plot(pred,lag=15,xlab="WLM/years",ylab="RR")
```

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Simulations



Penalized DLNMs

Currently, the bi-dimensional **exposure-lag-response** function $f \cdot w(x, \ell)$ is specified using **completely parametric** methods

However, simple DLMs also proposed in a Bayesian (Welty 2008) or penalized versions (Zanobetti 2000, Rushworth 2013, Obermeier 2015)

An obvious extension is to develop a **semi-parametric version** of DLNMs through **penalized splines**

The development may be facilitated by 'embedding' the R package `mgcv` in `dlnm`, exploiting the **existing GAM implementation**

Interactions in DLNMs

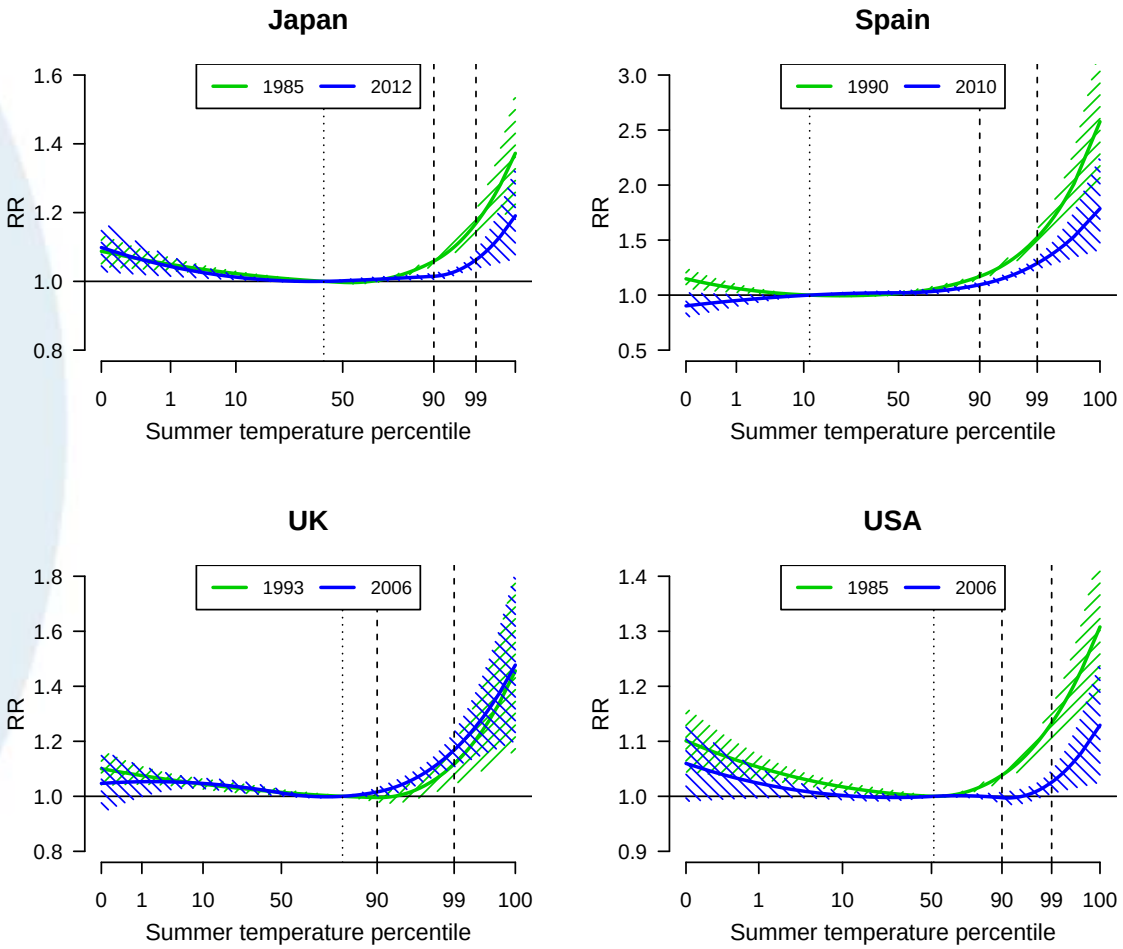
Interactions in DLNMs would allow the exposure-lag-response association **varying** depending on the value of other predictors (see also Rushworth 2013)

This corresponds to **relaxing the assumption** of identical effects

This development extends the framework to a wide range of **new applications**

However, it entails non-trivial **methodological problems**

Time-varying DLNMs



LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Some advantages

DLNMs offer a flexible way to model **exposure-lag-response** associations

Unified framework based on a general **conceptual** and **statistical** definition, applicable in various study designs

Complete software implementation, models can be fitted with **standard regression routines**

Some limitations

The DLNM framework is only applicable to **time-varying** (non-constant) exposures

It requires the **availability of exposure histories** (possibly reconstructed)

Model selection procedures still under-developed

Main references

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

Gasparrini A & Armstrong B. The R package `dlnm`. `http://cran.r-project.org/web/packages/dlnm/index.html`

E-mail: `antonio.gasparrini@lshtm.ac.uk`

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Other references (I)

- Abrahamowicz et al (2006). Modeling cumulative dose and exposure duration provided insights regarding the associations between benzodiazepines and injuries. *Journal of Clinical Epidemiology*, 59(4):393–403.
- Abrahamowicz et al (2012), Comparison of alternative models for linking drug exposure with adverse effects. *Statistics in Medicine*, 31:1014–1030.
- Almon S (1965). The distributed lag between capital appropriations and expenditures. *Econometrica*, 33(1):178–196.
- Armstrong (2006). Models for the relationship between ambient temperature and daily mortality. *Epidemiology*, 17(6): 624–631.
- Berhane et al (2008). Using tensor product splines in modeling exposure-time-response relationships: application to the Colorado Plateau Uranium Miners cohort. *Statistics in Medicine*, 27(26):5484–96.
- Heaton et al (2014). Extending distributed lag models to higher degrees. *Biostatistics*, 15(2):398–412.
- Gasparrini et al (2010). Distributed lag non-linear models. *Statistics in Medicine*, 29(21):2224–2234.
- Gasparrini (2011). Distributed lag linear and non-linear models in R: the package `dlnm`. *Journal of Statistical Software*, 43(8):1–20.
- Hauptmann et al (2000). Analysis of exposure-time-response relationships using a spline weight function. *Biometrics*, 56(4):1105–8.
- Langholz et al (1999). Latency analysis in epidemiologic studies of occupational exposures: application to the Colorado Plateau uranium miners cohort. *American Journal of Industrial Medicine*, 35(3):246–56.

Other references (II)

- Leffondre et al (2002). Modeling smoking history: a comparison of different approaches. *American Journal of Epidemiology*, 156(9):813.
- Obermeier et al (2015). Flexible distributed lag models and their application to geophysical data. *Journal of the Royal Statistical Society: Series B*, ahead of print.
- Richardson (2009). Latency models for analyses of protracted exposures. *Epidemiology*, 20:395–399.
- Rushworth et al (2013). Distributed lag models for hydrological data. Biometrics, 69:537–544.
- Schwartz (2000). The distributed lag between air pollution and daily deaths. *Epidemiology*, 11(3):320–326.
- Sylvestre & Abrahamowicz (2009). Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. *Statistics in Medicine*, 28(27):3437–53.
- Thomas (1983). Statistical methods for analyzing effects of temporal patterns of exposure on cancer risks. *Scand J Work Environ Health*, 9(4):353–366.
- Thomas (1988). Models for exposure-time-response relationships with applications to cancer epidemiology. *Annual Review of Public Health*, 9:451–82.
- Welty et al (2008). Bayesian distributed lag models: estimating effects of particulate matter air pollution on daily mortality. *Biometrics*, 65:282-291.
- Whitaker et al (2006). Tutorial in biostatistics: The self-controlled case series method. *Statistics in Medicine*, 25:1768–1797.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE