Modelling of time dependencies in associations between health effects and protracted exposures

Antonio Gasparrini

Centre for Statistical Methodology

LSHTM - 25 May 2012

Outline

Outline

Definitions

In biomedical research, an exposure event is frequently associated with a risk lasting for a defined period **in the future**

Similarly, the risk at a given time is commonly assumed as the results of protracted exposures experienced **in the past**

This phenomenon is common to environmental stressors, drugs, carcinogenic substances, amongst others

Proposed **statistical methods** depend on study design, type of data, research fields

A general modelling framework

The main conceptual and analytical challenge is to model the (potentially complex) temporal pattern of risk due to time-varying exposures

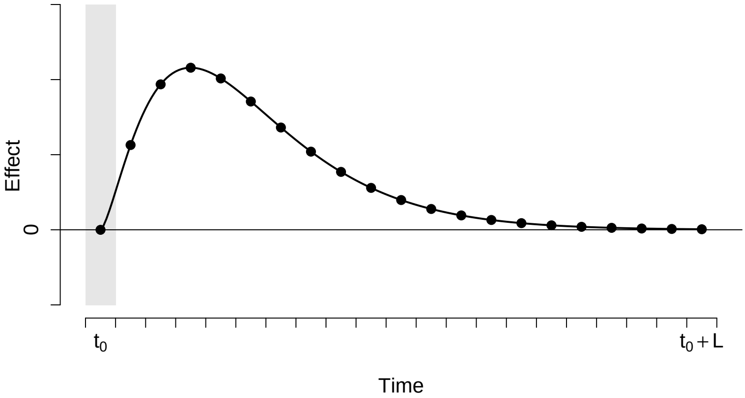
The risk is modelled in terms of contributions depending on **intensity** and **timing** of multiple exposure events:
$\Rightarrow$ **bi-dimensional** association

Aim: to develop a general conceptual and statistical framework to model this type of associations

Outline

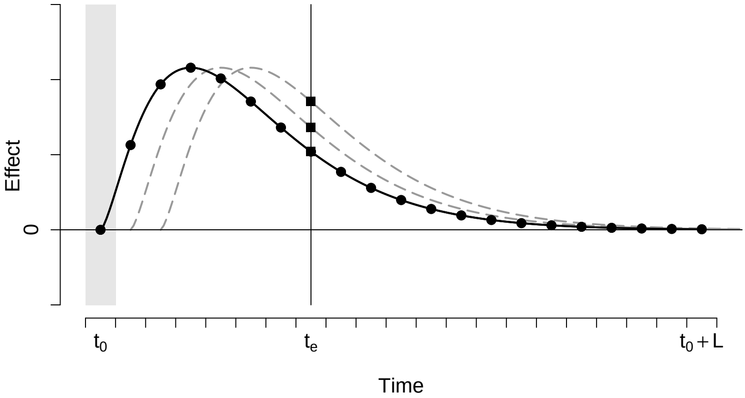
Conceptual representation

Single exposure event



Conceptual representation

Multiple exposure events



Perspectives and assumptions

Two perspectives:

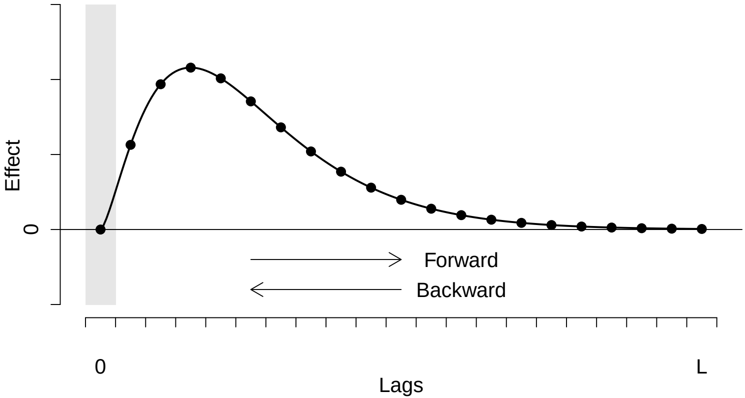
- **forward**: from fixed exposure to future outcomes
- **backward**: from fixed outcome to past exposures

Additional assumptions:

- assumption of **independence**
- assumption of **identical** effects

New lag dimension

Lag or weighting curve



Outline

Approaches for cohort data

Simple **cumulative exposure** (Checkoway *AJIM* 1992)

Exposure windows (Langholz *AJIM* 1999)

Flexible approaches based on **weighting functions** w (Thomas *SJWEH* 1983, Hauptmann *Biometrics* 2000, Richardson *Epidemiol* 2009, Sylvestre *Stat Med* 2009)

Models with weighting function

A function $s(x,t)$ for exposure-time response relationships:

$$s(x,t) = \beta_T \int_{t_0}^{t} w(t-u)\, x(u)\, du \;\approx\; \beta_T \sum_{u=t_0}^{t} w(t-u)\, x(u) \quad (1)$$

Re-parameterizing in terms of ℓ:

$$s(x,t) = \beta_T \int_{0}^{L} w(\ell)\, x(t-\ell)\, d\ell \;\approx\; \beta_T \sum_{\ell=0}^{L} w(\ell)\, x(t-\ell) \quad (2)$$

Distributed lag models (DLMs)

Developed for time series data in econometrics (Almon *Econometrica* 1965), and recently extended to epidemiology (Schwartz *Epidemiology* 2000)

Based on the concept of **lag curve**, analogous to the weighting curve described above

These models can be extended **beyond** time series data

DLMs: algebra

First define the lag vector:

$$\boldsymbol{\ell} = 0, \ldots, \ell, \ldots, L \tag{3}$$

and the vector $\mathbf{q}$ of exposure histories:

$$\mathbf{q} = [x_t, \ldots, x_{t-\ell}, \ldots, x_{t-L}]^{\mathsf{T}} \tag{4}$$

Then, given a basis matrix $\mathbf{C}$ derived from $\boldsymbol{\ell}$:

$$s(x, t; \boldsymbol{\eta}) = \mathbf{q}^{\mathsf{T}} \mathbf{C} \boldsymbol{\eta} = \mathbf{w}^{\mathsf{T}} \boldsymbol{\eta} \tag{5}$$

DLMs: interpretation

The **lag-basis** function:

$$s(x, t; \eta) = \mathbf{q}^{\mathsf{T}} \mathbf{C} \eta = \mathbf{w}^{\mathsf{T}} \eta \tag{6}$$

with $\hat{\beta} = \mathbf{C}\hat{\eta}$ as **weights** defining the lag curve

$\mathbf{Q}$ is the matrix of exposure histories, $\mathbf{W}$ the matrix of transformed variables

Different DLMs with **alternative options** for the basis to compute $\mathbf{C}$ (step, splines, etc.)

Outline

The CPUM cohort

3,347 subjects working in the Colorado Plateau mines between 1950–1960

Vital status and cause of death collected through record linkage up to 1982, with **258 lung cancer deaths**

Exposure history to **radon** (WLM) and **tobacco smoke** (pack×100) reported for 5-years age periods

Modelling strategy

Association between radon and mortality for lung cancer

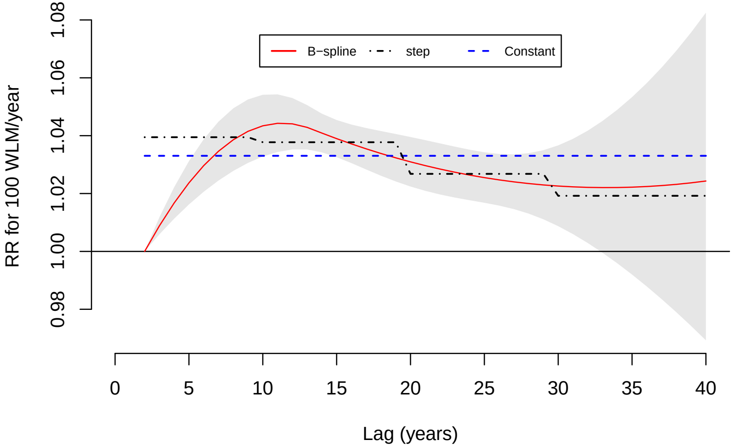
Analysis with Cox proportional hazards model for **time-varying exposures**

Age as time axis, controlling for smoking and calendar year

Exposure histories in WLM/year reconstructed for each of the **subject/periods** withing a **lag period 2–40**

Different functions used to estimate the lag curve: constant, step constant, quadratic B-spline

Lag curves for DLMs



Results

Table: DLMs with different functions for ℓ.

Function for x	*Function for ℓ*	*df*	*AIC*
Linear	Constant	1	2581.5
Linear	Step function[a]	4	2582.7
Linear	B-spline[b]	3	2583.2

[a] Cut-offs at lag 10-20-30

[b] Quadratic, 1 knot at lag 13.3, no intercept

Outline

Non-linear extension: DLNMs

Main limitation: assumption of a linear exposure-response dependency

DLMs can be extended as well to non-linear relationships, leading to **distributed lag non-linear models (DLNMs)** (Armstrong *Epidemiology* 2006, Gasparrini *Stat Med* 2010)

Link with previous work on the extension to non-linear dependencies (Abrahamowicz *Stat Med* 2007, Berhane *Stat Med* 2008)

DLNMs: algebra

Another function for modelling non-linearity applied to **q** in (4), deriving the basis matrix **R**. A DLNM is based on the **cross-basis** function:

$$s(x, t; \boldsymbol{\eta}) = \sum_{j=1}^{v_x} \sum_{k=1}^{v_\ell} \mathbf{r}_{\cdot j}^{\mathsf{T}} \mathbf{c}_{\cdot k} \eta_{jk} = \mathbf{w}^{\mathsf{T}} \boldsymbol{\eta} \tag{7}$$

Algebra is intricate, but concept is straightforward: applying 2 different functions to model a **bi-dimensional** association

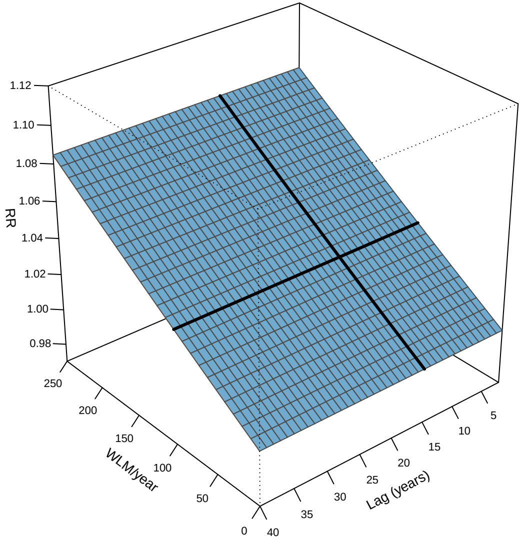
Extended results

Table: DLNMs with different functions for ℓ and x.

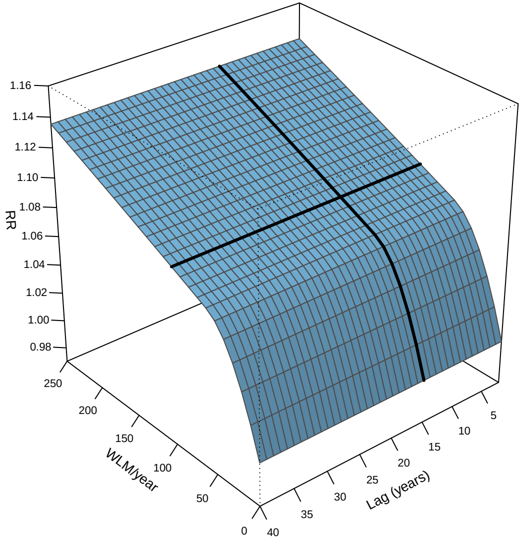
Function for x	*Function for ℓ*	*df*	*AIC*
Linear	Constant	1	2581.5
Linear	Step function[a]	4	2582.7
Linear	B-spline[b]	3	2583.2
B-spline[c]	Constant	3	2528.0
Step function[d]	Step function[a]	8	2507.0
B-spline[c]	B-spline[b]	9	2500.1

[a] Cut-offs at lag 10-20-30
[b] Quadratic, 1 knot at lag 13.3, no intercept
[c] Quadratic, 1 knot at 59.4 WLM/year, no intercept
[d] Cut-offs at 26-121 WLM

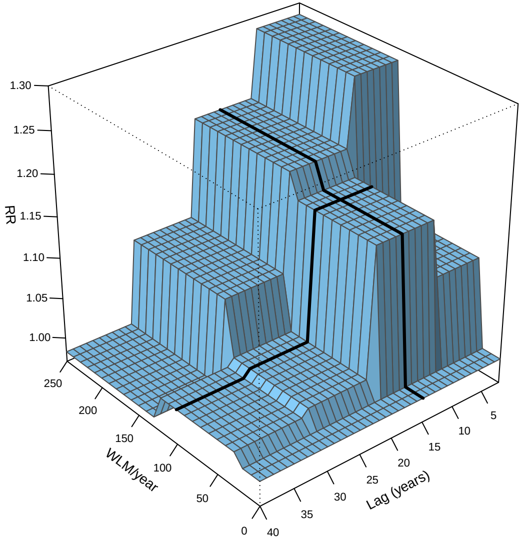
3D: linear-by-constant



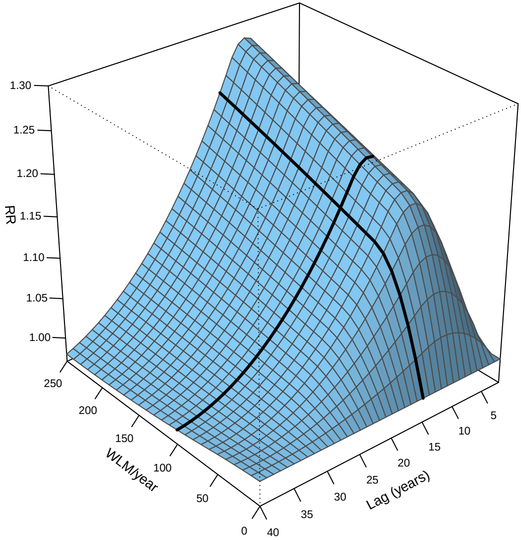
3D: spline-by-constant



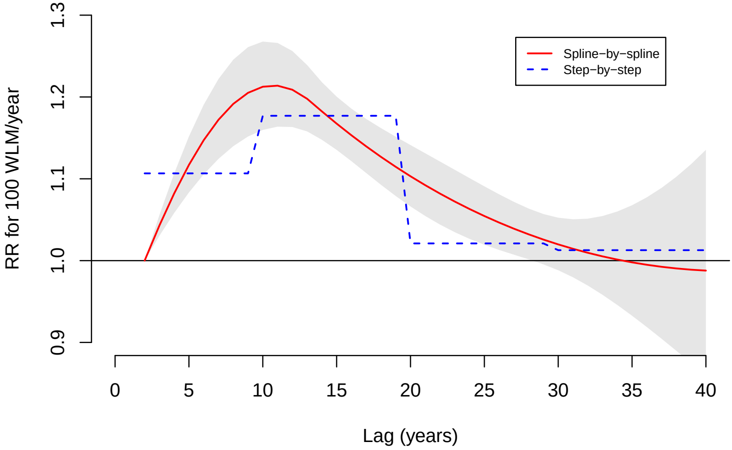
3D: step-by-step



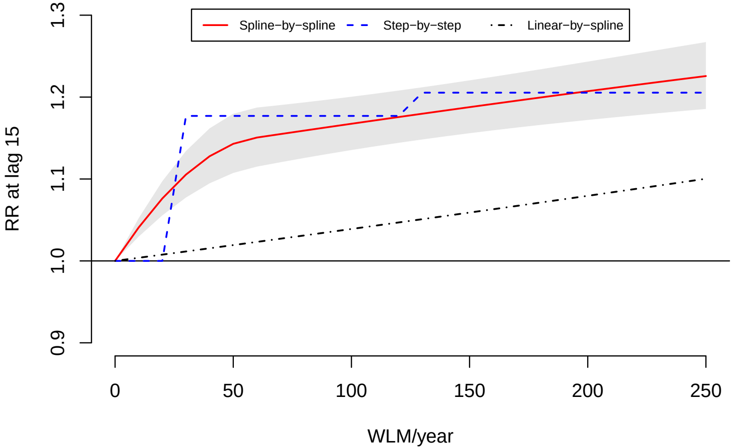
3D: spline-by-spline



Lag curves: comparison



Exposure-response: comparison

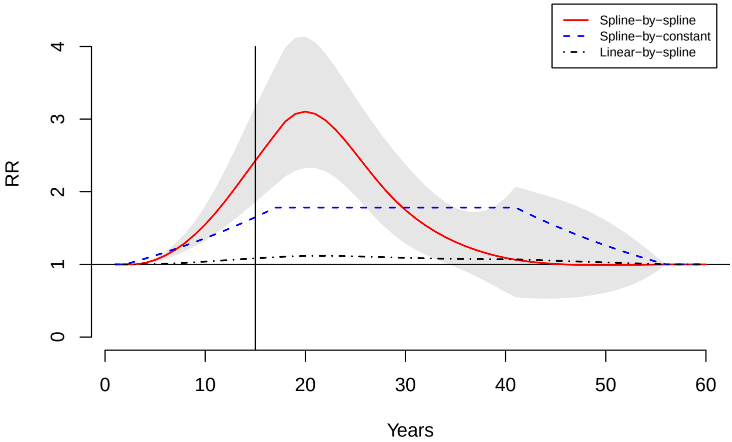


Predicted risk: 20 WLM for 10yrs

Table: Predicted risk from different models.

Model	*RR*
Linear-by-constant	1.053 (1.045–1.062)
Linear-by-spline	1.039 (1.027–1.050)
Spline-by-constant	1.361 (1.245–1.487)
Spline-by-spline	1.550 (1.331–1.805)

Evolution of risk: 20WLM for 15yrs



Outline

A unified framework

DLNMs offer a flexible way to model **complex time dependencies** in associations with protracted exposures

Potentially applicable in different setting (study design, data structure) where the risk is assumed to be associated with **protracted time-varying** exposures

The framework is based on a **simple conceptual scheme** which provides basis for interpretation and defines assumptions

Statistical and computational advantages

Statistical and **theoretical** development of DLNMs available in this extended framework

Based on entirely parametric approaches, with CI and tests off the shelf

Models can be fitted with **standard regression routines**

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

The R package dlnm

Example of code

```
library(dlnm)

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

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

pred <- crosspred(cb,model11,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")
```

Limitations

The main requirement is the availability of (possibly reconstructed) **exposure histories** for each observation (subject)

Exposure histories need to show **variability** between and within observations

More complex DLNMs need a relatively high statistical power

Model selection issues

Applications

- **Time series**: association between air pollution/temperature and mortality/morbidity – analysis in several cities/countries
- **Matched case-control**: association between gestational exposure to pesticides and autism – a study in California
- **Life course cohort**: association between body size during infancy/childhood and hip dysplasia at skeletal maturity – Bergen Hip Cohort Study

Potential applications

- Analysis of other occupational cohorts
- Association between smoking and lung cancer/cardiovascular risk
- Life course studies (Framingham Heart Study)
- ....

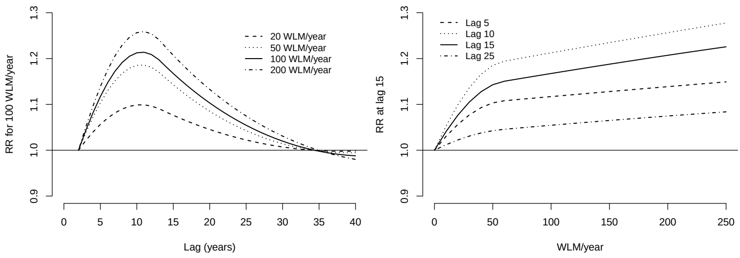
References (I)

- 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 & MacKenzie (2007). Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Statistics in Medicine*, 26(2):392408.
- 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): 624631.
- 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.
- Checkoway & Rice (1992). Time-weighted averages, peaks, and other indices of exposure in occupational epidemiology. *American Journal of Industrial Medicine*, 21(1):25-33.
- Elliott et al (2007). Long-term associations of outdoor air pollution with mortality in Great Britain. *Thorax*, 62(12):1088-94.
- 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.
- Gasparrini (2012). A general modelling framework for exposure-time-response relationships based on distributed lag non-linear models. *Statistics in Medicine*, in submission.
- Goodman et al (2004). Cause-specific mortality and the extended effects of particulate pollution and temperature exposure. *Environmental Health Perspectives*, 112(2):179185.

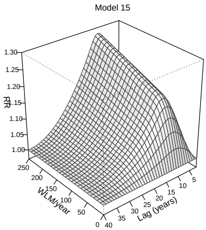
References (II)

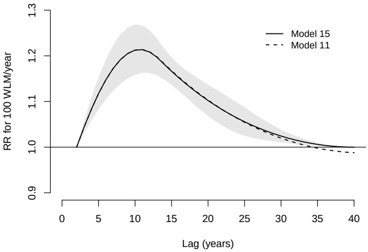
- Hauptmann et al (2000). Analysis of exposure-time-response relationships using a spline weight function. *Biometrics*, 56(4):1105-8.
- Langholz et al (1999). Latency analysis in epidemiologic studies of occupational exposures: application to the Colorado Plateau uranium miners cohort. *American Journal of Industrial Medicine*, 35(3):246-56.
- Langholz & Goldstein (1996). Risk set sampling in epidemiologic cohort studies. *Statistical Science*, 11(1):35-53
- Leffondre et al (2002). Modeling smoking history: a comparison of different approaches. *American Journal of Epidemiology*, 156(9):813.
- Richardson (2009). Latency models for analyses of protracted exposures. *Epidemiology*, 20:395399.
- Richardson et al (2009). Ionizing radiation and lymphoma mortality. *American Journal of Epidemiology*, 169(8):969-76.
- Roberts et al (2007). Maternal residence near agricultural pesticide applications and autism spectrum disorders among children in the California Central Valley. *Environmental Health Perspectives*, 115(10):1482-9.
- 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* 1983;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.

Slices


RR for 100 WLM/year
1.3
1.2
1.1
1.0
0.9
20 WLM/year
50 WLM/year
100 WLM/year
200 WLM/year
0 5 10 15 20 25 30 35 40
Lag (years)
RR at lag 15
Lag 5
Lag 10
Lag 15
Lag 25
0 50 100 150 200 250
WLM/year

Constraints





Interaction