

Analysis of Exposure-Time-Response Relationships Using a Spline Weight Function

Michael Hauptmann,^{1,*} Jürgen Wellmann,² Jay H. Lubin,³
Philip S. Rosenberg,³ and Lothar Kreienbrock⁴

¹GSF National Research Center for Environment and Health, Institute of Epidemiology,
Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany

²University of Münster, Institute of Epidemiology and Social Medicine,
Domagkstrasse 3, 48129 Münster, Germany

³National Cancer Institute, Division of Cancer Epidemiology and Genetics,
6120 Executive Boulevard, Bethesda, Maryland 20892, U.S.A.

⁴Hannover Veterinary School, Institute of Biometry and Epidemiology,
Bünteweg 2, 30559 Hannover, Germany

*email: hauptmann@nih.gov

** Current address: National Cancer Institute, Division of Cancer Epidemiology and Genetics,
6120 Executive Boulevard, EPS/7089, Bethesda, Maryland 20892, U.S.A.

SUMMARY. To examine the time-dependent effects of exposure histories on disease, we estimate a weight function within a generalized linear model. The shape of the weight function, which is modeled as a cubic B-spline, gives information about the impact of exposure increments at different times on disease risk. The method is evaluated in a simulation study and is applied to data on smoking histories and lung cancer from a recent case-control study in Germany.

KEY WORDS: Bootstrap; Case-control study; Exposure history; Exposure-time-response analysis; Latency; Lung cancer; Smoking; Splines; Time since exposure; Timing of exposure.

1. Introduction

In studies of acute exposures, time since exposure is clearly defined and often has great influence on the risk of disease. For extended exposures, the definition of time since exposure and its relationship to disease outcome are no longer obvious.

In a study of lung cancer where smoking histories have been collected, the question is raised how the time pattern of smoking affects disease risk. Various attempts have been made to specifically examine this relation (Doll and Peto, 1978; Lubin et al., 1984).

We propose the estimation of a time-dependent weight function. Suppose exposure histories $x_i(t)$, $t = 1, \dots, T_i$ and $i = 1, \dots, n$, are collected, where $x_i(t)$ denotes the cumulative exposure of individual i received between times $t-1$ and t prior to an index date (e.g., the date of interview in a case-control study). Without loss of generality, we assume equal time intervals.

Generally, investigators apply models in cumulative exposure $\sum_t x_i(t)$. We propose a linear combination of the incremental exposures, namely

$$\sum_t w(t)x_i(t), \quad (1)$$

where $w(t)$ is a weight function to be estimated. This linear combination can be interpreted as effective exposure. The shape of the weight function gives information about the impact of exposure at different times on disease risk. Cubic B-splines are used to model the weight function. Its parameters are estimated via a constrained maximization of the likelihood.

The method is illustrated on data from a recent German case-control study on smoking and lung cancer with about 4300 cases and a similar number of controls (Kreienbrock et al., 1992).

2. Method: The Spline Weight Function

2.1 B-Spline Weight Function Model

For a response variable y , a generalized linear model that includes time-weighted effective exposure (1) as covariate is given by $g\{E(y)\} = \eta$, where $g(\cdot)$ is the link function, and the linear predictor η_i for the i th individual is given by

$$\eta_i = \beta_0 + \beta_1 \sum_{t=1}^{T_i} w(t)x_i(t) + z_i\lambda, \quad i = 1, \dots, n. \quad (2)$$

The parameter β_1 represents the impact of effective exposure

(1) on y and λ is the column vector of parameters for additional covariates $z_i = (z_{i1}, \dots, z_{iq})$.

The weight function $w(t)$ is modeled as a cubic B-spline on $[0, T]$, where $T = \max_i\{T_i\}$. Splines are continuously differentiable piecewise polynomial functions (de Boor, 1978). B-splines are used because they have minimum support among all sets of basis functions for the space of cubic splines. This property reduces correlations among the columns of the design matrix and thereby eases computations. The B-spline representation of a cubic spline is given by

$$w(t, \theta) = \sum_{j=-3}^m \theta_j B_j(t), \tag{3}$$

where θ_j are parameters to be estimated. Conditional on inner knots $0 < t_1 < \dots < t_m < T$, the B-spline basis functions $B_j(t)$ are known functions of t . Formulas for the $B_j(t)$ are given in the Appendix.

Note that the property of the B-splines, $\sum_{j=-3}^m B_j(t) = 1$ for all $t \in [0, T]$, can be used to define a likelihood ratio test $H_0: \theta_1 = \dots = \theta_m = 1$, i.e., there is no time variation in the effects of exposure. This test has $m + 3$ d.f.

2.2 Constrained Maximum Likelihood Estimation

Since the values of the cubic spline are viewed as weights for exposure increments received at certain times in the past, we constrain the spline function such that weights are nonnegative, $\theta_j \geq 0$ for all j , and standardized, $\sum_{j=-3}^m w(t, \theta) = T$. Risk effects that depend only on cumulative exposure and not time correspond to $w(t) = 1$ for all t .

If the maximum likelihood estimate $\hat{\theta}$ lies on the boundary of the restricted parameter space, the usual asymptotic theory may not apply. In this case, confidence intervals for the weight function are obtained by nonparametric bootstrap sampling. For the bootstrap, the values of the estimated weight function at $t = 1, \dots, T$ are calculated for each set of parameter replicates and the variance of those values is used to construct asymptotically normal pointwise confidence limits for the spline weight function at $t = 1, \dots, T$. The slope of the estimated weight function and corresponding pointwise confidence intervals can also be computed.

2.3 Knot Selection

Usually one would try to select the best knots with respect to some criteria. The knots can be viewed as nonlinear parameters that have to be estimated according to a goodness-of-fit criterion. Several methods to do this are described in the literature and are in general referred to as adaptive knots (Hastie and Tibshirani, 1990; Friedman, 1991).

These methods are complicated and numerically cumbersome, so we used the following intuitive approach that is similar to percentile categorization. For m inner knots and thus $m + 1$ intervals, choose knot locations such that each interval includes $1/(m+1) \times 100\%$ of the population exposure. More precisely, choose the j th knot t_j so that $t_j = \max\{t = 1, \dots, T \mid \sum_{i=1}^n \sum_{\ell=1}^t x_i(\ell) / \sum_{i=1}^n \sum_{\ell=1}^T x_i(\ell) \leq (j-1)/(m+1)\}$.

In the following example, the number m of knots is restricted to the sequence $m = 2, \dots, 8$ to ensure flexibility of the spline weight function while avoiding overparameterization. The final model is the one that minimizes the Akaike information criterion (AIC) (McCullagh and Nelder, 1989), $AIC(m) = -2 \log L(\hat{\beta}, \hat{\theta}, \hat{\lambda}) + 2(m + q + 6)$.

3. Example: Lung Cancer and Smoking

We use the standard unconditional logistic model for the analysis of a case-control study on lung cancer and smoking. Then $E(y_i) = p_i = \Pr(y_i = 1 \mid x_i(t), t = 1, \dots, T_i, z_i)$ and the link function is $g(\mu) = \log\{\mu/(1-\mu)\}$. Parameter β_1 of model (2) represents the log odds ratio per unit effective exposure (1). Alternatively, $\beta_1 w(t)$ can be viewed as the log odds ratio per unit exposure received at time t in the past.

3.1 Data Description

We apply the method to data from a case-control study carried out from 1990 through 1996 in Germany (Kreienbrock et al., 1992). Cases include patients aged 75 years and under with histologically confirmed primary lung cancer. Controls are population based and frequency matched to cases on age (within 5 years), sex, and place of residence (23 regions).

Historical smoking data on the type and amount of tobacco products smoked from age at start of smoking to age at interview are obtained by intervals of constant smoking habit. For a cigarette smoker who also smoked cigars, cigarillos, or pipes, the tobacco exposure equivalent is added to the exposure from cigarettes. After excluding 174 individuals who smoked cigars, cigarillos, or pipes only and 43 individuals with incomplete smoking histories, the study population includes 4304 cases and 4526 controls.

Since smoking histories are based on years, exposure profiles were reconstructed in 1-year intervals from birth to interview. Exposure variables $x_i(t)$ denote the number of pack-years (1 pack-year = 365×20 cigarettes) smoked by the i th individual during the year t years prior to the interview ($t = 1, \dots, T_i$, where T_i is attained age of the i th individual). The response variable y is the case-control status.

Analyses are adjusted for the matching variables and asbestos exposure (ever/never). Models for females only are not adjusted for asbestos exposure since few women have been exposed.

3.2 Results

We specified models with four knots for females and six knots for males based on the AIC. The models fit significantly better than a simple model in cumulative exposure, i.e., $p < .001$ for males and females for testing $\theta_j = 1$ for all j . Figure 1 shows the estimated weight functions and their slopes with asymptotic pointwise 95% confidence intervals from 1000 bootstrap replications.

Estimated weight functions for both sexes show a global maximum at about 5 years before interview and are sharply decreasing thereafter until they are about one or smaller for more than 15 years in the past. For males and females, there are local maxima, 46 and 22 years, respectively. However, confidence intervals are wide and preclude meaningful interpretation. The slopes of the estimated weight functions decline sharply within the first 8 years before interview and approach zero thereafter.

The bootstrap parameter replicates for the unconstrained parameters (i.e., the intercept and the adjustment variables) are symmetric and normally shaped. However, due to constraints, the replicates are skewed to the right for some of the spline parameters.

Material may be protected by copyright law (Title 17, U.S. Code)

Weights
Slope
Fig
their
smo
wise
repl
knou
T
is 1.
for r
odds
W
sex a
years
age g
than
as th
4. S
We c
of e
patte
T =
data.
be a
of pa
defin
 $v_i \sim$
so th
mean
the p
Ch
The p
interv
a .5
If the
smok
For
mode
 β_0 at
includ
triang
t = 1,

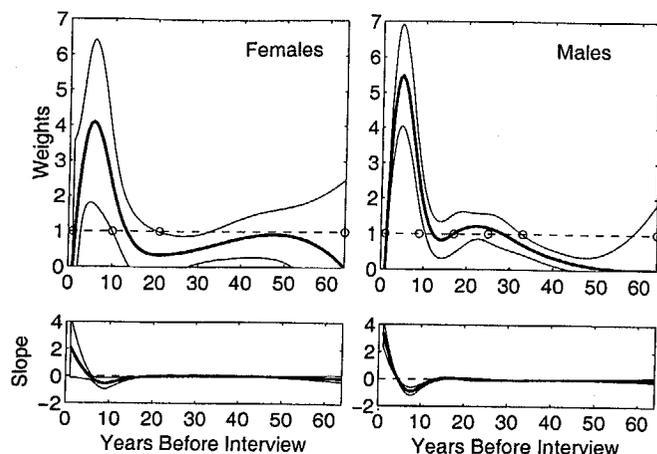


Figure 1. Estimated cubic B-spline weight functions and their slopes from logistic regression of lung cancer risk on smoking profiles for females and males. Approximate pointwise 95% confidence limits are based on 1000 bootstrap replications. Circles on the horizontal axes indicate spline knots.

The odds ratio per 1 pack-year effective exposure ($\exp(\hat{\beta}_1)$) is 1.06 (95% CI [1.01, 1.11]) for females and 1.04 (1.03, 1.05) for males. They are slightly smaller than the corresponding odds ratios from a standard model with cumulative exposure.

We also estimated the weight function within categories of sex and age. The weight functions reached a maximum 3–11 years prior to interview, before decreasing. Although younger age groups show a steeper decrease with time before interview than older age groups, these differences could be due to chance as the models were homogenous over sex and age strata.

4. Simulation Study

We carried out a simulation study to study the characteristics of estimating weights with predefined disease-exposure patterns. Hypothetical smoking profiles are generated for $T = 60$ years prior to interview based on the German smoking data. With probability .25, an individual is considered to be a lifelong nonsmoker. With probability .75, the number of pack-years smoked during the year prior to interview is defined as $x_i(1) = \max(0, .47 + .3672v_i)$, $i = 1, \dots, n$, where $v_i \sim N(0, 1)$. The truncated normal distribution is chosen so that the expectation over its nonzero support equals the mean yearly exposure in the German smoking data and that the probability of zero exposure is .1.

Changes in smoking rates over time are modeled as follows. The probability p_t of a change in smoking rate at year t before interview is defined by $.5 = (1 - p_t)^{10}$, i.e., corresponding to a .5 probability of a change in smoking rate over 10 years. If the subject has changed smoking rate, we resample a new smoking rate from the distribution above.

For $w(t)$, we generate the response based on a logistic model with covariate $\sum_{t=1}^{60} w(t)x_i(t)$ and given parameters β_0 and β_1 . We use several simple weight functions $w(t)$, including constant, linearly increasing, linearly decreasing, triangular shape, and trapezoidal shape. Weight functions for $t = 1, \dots, 60$ are standardized to sum to 60.

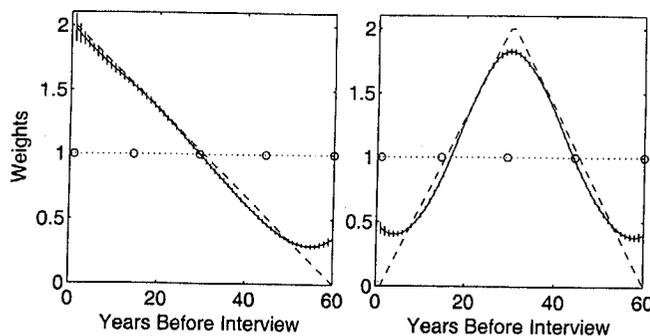


Figure 2. Spline weight function analysis of simulated data based on monotone decreasing weights (dashed line in left panel) and triangle-shaped weights (dashed line in right panel). The estimated spline weight functions (solid lines) are shown with approximate pointwise normal 95% confidence intervals. Average knot positions are indicated by circles.

Since the time-weighted exposure $\sum w(t)x(t)$ is related to cumulative exposure, the log odds ratio per pack-year estimated from the empirical data, $\log \hat{OR} = 1.06$, is chosen for β_1 .

For case-control data, β_0 is the log odds of being a case for a never-smoker, i.e., at $x(t) = 0$ for all $t = 1, \dots, 60$. Since we have half cases and half controls in our study, we set $\Pr(y = 1 | \bar{x}) = .5$, which is equivalent to $\beta_0 = -\beta_1 \bar{x}$. With a mean cumulative exposure of $\bar{x} = 22.3$ pack-years, we get $\beta_0 = -1.29$.

Data are simulated using five exposure scenarios represented by different hypothetical weight functions. For each scenario, 1000 case-control studies with 500 cases and 500 controls each are generated.

Figure 2 shows the results of the simulation study for two of the five given hypothetical weight functions (dashed lines). The estimated spline weight functions (solid lines) follow closely the given weight functions. Some deviations from the given weight functions can be observed at the tails, a well-known weakness of splines.

The approximate normal pointwise 95% confidence intervals of the estimated spline weight function values indicate a small simulation variability after 1000 replications.

Measurement error. Exposures many years in the past may be subject to greater measurement error than more recent exposures. We assess the sensitivity of the method to time-dependent error in exposure.

The previous simulation used exposure profiles $x(t)$, $t = 1, \dots, 60$. We now include exposure profiles with error $\tilde{x}(t) = x(t)e(t)$, where the time-dependent measurement error $e(t)$ is uniformly distributed on $[1 - t(1 - p)/60, 1 + t(1 - p)/60]$. The measurement error distribution is linearly increasing in time t from $U[1, 1]$ at interview to $U[p, 2 - p]$ at 60 years before interview for some chosen fraction $p \in (0, 1)$.

For the five weight functions and for a variety of sizes of error, results were similar to using the error-free profiles, i.e., error had little impact on the shape of the weight function. The simulation suggests that the patterns observed in Figure 2 are not likely a result of measurement error.

5. Discussion

Our application of a spline function to estimate weights by increments of exposure can be viewed as a subset of the class of varying-coefficient models (Hastie and Tibshirani, 1993). We apply constraints on the functional form to enhance their epidemiological interpretation.

In data from a case-control study on lung cancer, we found that the number of cigarettes smoked within 5–15 years prior to interview strongly determines an individual's risk of lung cancer and that cigarettes smoked more than 20 years before interview contribute only minimally to risk. The pattern corresponds to an observed decrease of risk with time since smoking cessation. A constant weight function is therefore not consistent with the data, and the use of cumulative exposure or average exposure rate is not appropriate.

Our choice of time before interview as the time scale was based on a preliminary analysis of age-specific weight functions. This analysis revealed a risk pattern for smoking incompatible with an age-at-exposure effect. The age-specific weight functions give no indications that exposure received in youth are especially hazardous compared to exposure received at older ages.

In a limited simulation study, we showed that increasing error with time since interview had little impact on the shape of the weight function. The method thus seems robust against nondifferential, multiplicative, time-dependent measurement error that often arises in retrospective collection of exposure histories.

An alternative exploratory approach to analyze the exposure-time-response relationship using sliding time windows is recently proposed and applied to the same data (Hauptmann et al., 2000). Both methods yield similar conclusions and are superior to the standard approach of using time since quitting smoking and duration of smoking. The ability of the standard approach to detect a decline in risk with increasing time since smoking cessation depends heavily on the prevalence of ex-smokers. While this may not be a problem with smoking, some environmental or occupational exposures are ubiquitous so that there is no complete cessation of exposure. In addition, the standard approach does not consider changes in the number of cigarettes smoked per day except for complete cessation.

RÉSUMÉ

Pour étudier les effets temporels des cursus d'exposition sur les maladies, on utilise une fonction de pondération à l'intérieur d'un modèle linéaire généralisé. La forme de cette fonction, en prenant comme modèle une cubique B-spline, fournit des informations concernant l'impact à différents moments des accroissements d'exposition sur le risque d'être malade. La méthode est évaluée à partir d'une étude de simulation et est appliquée aux données relatives à l'histoire des consommations de tabac relevées dans une récente enquête cas-témoin en Allemagne portant sur les cancers pulmonaires.

REFERENCES

- Atkinson, K. E. (1989). *An Introduction to Numerical Analysis*. New York: John Wiley.
 de Boor, C. (1978). *A Practical Guide to Splines*, Volume 27, *Applied Mathematical Science*. New York: Springer.

- Doll, R. and Peto, R. (1978). Cigarette smoking and bronchial carcinoma: Dose and time relationships among regular smokers and lifelong nonsmokers. *Journal of Epidemiology and Community Health* **32**, 303–313.
 Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics* **19**, 1–67.
 Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. New York: Chapman and Hall.
 Hastie, T. J. and Tibshirani, R. J. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B* **55**, 757–796.
 Hauptmann, M., Lubin, J. H., Rosenberg, P. S., Wellmann, J., and Kreienbrock, L. (2000). The use of sliding time windows for the exploratory analysis of temporal effects of smoking histories on lung cancer. *Statistics in Medicine* **19**, 2184–2194.
 Kreienbrock, L., Wichmann, H. E., Gerken, M., Heinrich, J., Götze, H.-J., Kreuzer, M., and Keller, G. (1992). The German radon project—Feasibility of methods and first results. *Radiation Protection Dosimetry* **45**, 643–649.
 Lubin, J. H., Blot, W. J., Berrino, F., Flamant, R., Gillis, C. R., Kunze, M., Schmaehl, D., and Visco, G. (1984). Modifying risk of developing lung cancer by changing habits of cigarette smoking. *British Medical Journal* **288**, 1953–1956.
 McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. New York: Chapman and Hall.

Received March 1999. Revised February 2000.

Accepted March 2000.

APPENDIX

A cubic spline on $[0, T]$ consists of cubic polynomials on the $m+1$ segments defined by m inner knots $0 < t_1 < \dots < t_m < T$. Adjacent polynomials are smoothly joined so that first and second derivatives agree at the knots.

Using a numerically favorable representation of cubic splines, the space of cubic splines can be spanned with $m+4$ basis functions $B_j(t)$, called B-splines. Therefore, the knot list has to be augmented by six associated arbitrary slack knots. Without loss of generality, let $t_{-3} = -3$, $t_{-2} = -2$, $t_{-1} = -1$ and $t_{m+2} = T+1$, $t_{m+3} = T+2$, $t_{m+4} = T+3$ and denote $t_0 = 0$ and $t_{m+1} = T$.

According to Atkinson (1989), the basis functions are defined by

$$B_j(t) = (t_{j+4} - t_j) \sum_{i=j}^{j+4} \frac{(t_i - t)_+^3}{\prod_{k=j, k \neq i}^{j+4} (t_k - t_i)},$$

$$j = -3, \dots, m,$$

where $(t_i - t)_+^r = (t_i - t)^r$ if $t_i > t$ and zero otherwise.

Calculations are performed in MATLAB 5.3. The design matrix of the spline weight function model is created using the function `spscol` from DeBoor's spline toolbox. For X containing the exposure profiles, Z containing the matrix of additional covariates, and `knots` containing the list of inner knots between 0 and T , the code is `design = [X * spscol(augknt([0, knots, T], 4), 4, 0: T), Z]`. The function `fmincon` does the constrained maximization of the likelihood.

1. In
It is
screen
are in
tion o
this a
ists (a
to est
proce
ologic
the m
new t
are ty
know
condit
erate
(See F
The
ogy is
each s
Ubers
sidere
time p
sumed

Material may be protected by copyright law (Title 17, U.S. Code)