

Penalized estimation of complex, non-linear exposure-lag-response associations

ANDREAS BENDER*

Statistical Consulting Unit, StaBLab, Department of Statistics, Ludwig-Maximilians-Universität München, Ludwigstr. 33, 80539 Munich, Germany
andreas.bender@stat.uni-muenchen.de

FABIAN SCHEIPL

Department of Statistics, Ludwig-Maximilians-Universität München, Ludwigstr. 33, 80539 Munich, Germany

WOLFGANG HARTL

Department of General, Visceral, Transplantation, and Vascular Surgery, University School of Medicine, LMU Munich, Grosshadern Campus, Marchioninistraße 15, 81377 Munich, Germany

ANDREW G. DAY

Clinical Evaluation Research Unit, Kingston General Hospital, KGH Research Institute, 76 Stuart Street, Kingston, Ontario K7L 2V7, Canada

HELMUT KÜCHENHOFF

Statistical Consulting Unit, StaBLab, Department of Statistics, Ludwig-Maximilians-Universität München, Ludwigstr. 33, 80539 Munich, Germany

SUMMARY

We propose a novel approach for the flexible modeling of complex exposure-lag-response associations in time-to-event data, where multiple past exposures within a defined time window are cumulatively associated with the hazard. Our method allows for the estimation of a wide variety of effects, including potentially smooth and smoothly time-varying effects as well as cumulative effects with leads and lags, taking advantage of the inference methods that have recently been developed for generalized additive mixed models. We apply our method to data from a large observational study of intensive care patients in order to analyze the association of both the timing and the amount of artificial nutrition with the short term survival of critically ill patients. We evaluate the properties of the proposed method by performing extensive simulation studies and provide a systematic comparison with related approaches.

Keywords: Cumulative effects; Exposure-lag-response association; Penalized additive models; Reproducible research; Survival analysis; Time-dependent covariates.

*To whom correspondence should be addressed.

1. INTRODUCTION

Patients admitted to an intensive care unit (ICU) often require mechanical ventilation (MV) and artificial nutrition (enteral or parenteral). The optimal timing and amount of artificial nutrition, however, is unclear and previous observational and randomized studies have not yielded consistent results (for details, see Tables S5 and S6 in Appendix A.9 of [supplementary material](#) available at *Biostatistics* online). Modeling the association between caloric intake and survival is particularly challenging for the following reasons: First, the amount of artificial caloric intake during ICU stay varies daily on a per patient basis. Second, effects of artificial nutrition may vary over time, and hazard rates at a particular point in time may depend on multiple past caloric intakes (i.e., the exposure history). Third, the caloric intake may have a delayed impact on the outcome, and may “wear off” after some time. Consequently, contradictory results of recent observational studies may be explained by their inability to account for the interplay between the timing and amount of artificial nutrition. Older observational studies have only considered the average caloric intake over a defined period (usually 1–2 weeks) and ignored the exact timing of caloric intake. In general, our proposal aims at modeling the cumulative effect of past exposure as a function of both the timing and amount of exposure which might additionally vary over the follow-up time. In the context of the application, this time-variation means that the contribution to the hazard of a given amount of nutrition on Day 4 to the hazard at Day 7 may be different than that of an identical amount of nutrition on Day 6 to the hazard at Day 9, even though the time between exposure and risk is the same in both cases.

Let us define some necessary terminology. We denote effects of time-constant covariates that can vary over time as *time-varying effects*. Variables whose values change over time (here nutrition on the ICU), are referred to as *time-dependent covariates* (TDC). We denote the follow-up time by t and the times at which the TDC was observed by t_e . The value of the TDC at time t_e is then denoted by $z(t_e)$ and a subject's exposure history by $\mathbf{z} = (z(t_{e_1}), z(t_{e_2}), \dots)$. The *lag* time t_{lag} is defined as the length of the delay until the TDC recorded at time t_e starts to affect the hazard. The *lead* time t_{lead} defines the duration of the effect of the TDC recorded at t_e . That is, $z(t_e)$ can affect the hazard for all $t \in [t_e + t_{\text{lag}}, t_e + t_{\text{lag}} + t_{\text{lead}}]$ and, vice versa, the hazard at time t is affected by exposures $\{z(t_e) : t_e \in [t - t_{\text{lag}} - t_{\text{lead}}, t - t_{\text{lag}}]\}$. In this work, we propose a novel approach for the modeling of such *exposure-lag-response associations* (ELRAs; [Gasparrini, 2014](#)) that extends previous research by (i) taking into account all three dimensions relevant to the effect of a time-dependent exposure: the timing of exposures t_e as well as their amounts recorded in a subject's exposure history \mathbf{z} , whose effect can vary over follow-up time t and (ii) by allowing for more flexible definitions of the window of effectiveness defined by t_{lag} and t_{lead} . Thus, we present a general framework for the estimation of effects like

$$g(\mathbf{z}, t) = \int_{t - t_{\text{lag}}(t_e) - t_{\text{lead}}(t_e)}^{t - t_{\text{lag}}(t_e)} h(t, t_e, z(t_e)) dt_e, \quad (1.1)$$

where $g(\mathbf{z}, t)$ denotes the cumulative effect of exposure history \mathbf{z} on the log-hazard at time t . The term $h(t, t_e, z(t_e))$ denotes individual contributions of exposures $z(t_e)$ recorded at specific exposure times t_e . We refer to these contributions as *partial effects*. The cumulative effect $g(\mathbf{z}, t)$ is defined as the integral of these partial effects within a pre-specified time window defined by lag and lead times and can be approximated by a weighted sum over the respective partial effects.

Previous approaches are special cases of this general definition: *weighted cumulative exposure* (WCE) models ([Sylvestre and Abrahamowicz, 2009](#)) assume that $h(t, t_e, z(t_e)) \stackrel{!}{=} \tilde{h}(t - t_e)z(t_e) \forall t, t_e$, i.e., partial effects that are linear in $z(t_e)$ and depend only on the latency $t - t_e$. [Sylvestre and Abrahamowicz \(2009\)](#) estimate $\tilde{h}(t - t_e)$ using B-Spline basis expansions and smoothness is controlled with BIC-based knot selection. [Berhane and others \(2008\)](#) use unpenalized B-Spline tensor product bases to model the relative risk of dying as a bivariate non-linear function of radon exposure $z(t_e)$ and latency $t - t_e$, i.e., they set

$h(t, t_e, z(t_e)) \stackrel{!}{=} \tilde{h}(t - t_e, z(t_e)) \forall t, t_e$. Estimation is performed using logistic regression models, which only applies to discrete-time models (or discretized time) and they do not provide uncertainty estimates for the bivariate surface estimate. [Gasparrini \(2014\)](#) defines the framework of distributed lag non-linear models (DLNM), where partial effects are specified as in [Berhane and others \(2008\)](#) and a penalized likelihood estimation approach for DLNMs is described in [Gasparrini and others \(2017\)](#). This DLNM can be obtained as a special case of (1.1) in the same way as for [Berhane and others \(2008\)](#). In all previous approaches, the weight of exposure $z(t_e)$ only depends on the latency $t - t_e$, not the specific combination of t and t_e . Thus, for example, $h(t = 30, t_e = 3) \stackrel{!}{=} h(t = 40, t_e = 13) \stackrel{!}{=} \tilde{h}(t - t_e = 27)$, which can be a perfectly reasonable assumption in many cases, but is often unknown a priori. Relaxing this assumption leads to more flexible versions of the WCE and DLNM. A more general WCE model thus could be defined using partial effects $h(t, t_e) \cdot z(t_e)$, which can not be defined as a special case of the DLNM. In Section 3, we apply this extension of the WCE to the nutrition data with a three categorical $z(t_e)$. A more flexible version of the DLNM, defined by partial effects $h(t, t - t_e, z(t_e))$, is used in simulation study Part A, Section 4.2. When we define ELRAs in Section 2.2.4, we allow additional flexibility, as t_{lag} and t_{lead} can also be functions of exposure time t_e , thus the number of past exposures that contribute to the cumulative effect can also depend on (exposure) time.

The remainder of this article is structured as follows. Our model, which is an extension of the piece-wise exponential model (PEM), is described in detail in Section 2. The proposed method for estimating ELRAs is presented in Section 2.2.4. Section 3 shows an application of our approach to observational data of almost 10 000 critically ill patients. In Section 4, we present results of an extensive simulation study to assess properties of the proposed modeling approach and to investigate its behavior if modeling assumptions are violated. A summary and discussion are presented in Section 5. Details regarding reproducibility of our analyses can be found in Section 6.

2. MODELS AND METHODS

In the following, we outline the proposed framework for fitting ELRAs, starting from simpler models and gradually increasing their complexity. This motivates the general flexibility of the proposed model class (cf. [Argyropoulos and Unruh, 2015](#)) and helps to understand ELRAs as natural extensions of time-varying effects for TDCs.

2.1. Piece-wise exponential model

Given a partition of the follow-up period $(0, t_{\text{max}}]$ into J intervals with $J + 1$ cut-points $0 = \kappa_0 < \dots < \kappa_J = t_{\text{max}}$, where t_{max} is the maximal follow-up time, and assuming the baseline hazard rate $\lambda_0(t)$ in each interval j to be constant, such that $\lambda_0(t) = \lambda_{0j}, \forall t \in (\kappa_{j-1}, \kappa_j], t > 0$, a proportional hazards model for subjects $i = 1, \dots, n$ can be written as

$$\log(\lambda_i(t|\mathbf{x}_i)) = \log(\lambda_{0j}) + \mathbf{x}'_i \boldsymbol{\beta} \quad \forall t \in (\kappa_{j-1}, \kappa_j], \quad (2.1)$$

with $\mathbf{x}'_i = (x_{i,1}, \dots, x_{i,P})$ a row-vector of P time-constant covariates.

This is the PEM of [Whitehead \(1980\)](#) and [Friedman \(1982\)](#), whose likelihood is equivalent to that of a Poisson generalized linear model (GLM)

$$\log(\mathbb{E}(y_{ij}|\mathbf{x}_i)) = \log(\lambda_{0j}) + \mathbf{x}'_i \boldsymbol{\beta} + \log(t_{ij}) \quad (2.2)$$

with (i) one observation for each interval $j = 1, \dots, J$ under risk for subject i , (ii) responses as event indicators for subject i for interval j , i.e., $y_{ij} = I(t_i \in (\kappa_{j-1}, \kappa_j] \wedge t_i = T_i)$, and (iii) $t_{ij} = \min(t_i - \kappa_{j-1}, \kappa_j -$

κ_{j-1}), representing the time subject i spent under risk in interval j . We define $t_i := \min(T_i, C_i)$ as the observed right-censored time under risk for subject i with event time T_i and censoring time C_i , which is assumed to be non-informative for this model class.

In practice, when fitting the respective Poisson regression model, $\log(\lambda_{0j})$ is incorporated in the linear predictor $\mathbf{x}'_i \boldsymbol{\beta}$ and $\log(t_{ij})$ enters as an offset. In the following sections, our model specifications for the log-hazard rate do not contain the offset term, but the offset must be included at the estimation stage. This model structure lends itself easily to include TDCs, as a covariate can change its value at each κ_j , i.e., *additional* interval cut-points can be chosen at the time-points at which a change in the TDC is recorded (see Section 2.2.1 for more details on the choice of cut-points). Then (2.1) can be extended to $\log(\lambda_i(t|\mathbf{x}_{ij})) = \log(\lambda_{0j}) + \mathbf{x}'_{ij} \boldsymbol{\beta}$. Additionally, time-varying effects can be incorporated by creating a TDC for time itself, e.g., by using the interval midpoints $\tilde{t}_j := (\kappa_j - \kappa_{j-1})/2$, and including interaction terms of selected covariates with \tilde{t}_j or transformations thereof, in the linear predictor.

2.2. Piece-wise exponential additive model

Model (2.2) is a GLM, in which covariate effects are assumed to be linear in x and constant or linear over t as well as non-cumulative, and observations are assumed to be independent. In order to remove these restrictions, we introduce the class of *piece-wise exponential additive mixed models* (PAMMs) (cf. [Argyropoulos and Unruh, 2015](#)), in analogy to the extension of GLMs to generalized additive mixed models (GAMMs). By doing so, we can simultaneously (i) achieve reliable estimates of the baseline hazard parameters $\lambda_{0j}, j = 1, \dots, J$ even for very fine partitions $[\kappa_0, \dots, \kappa_J]$ of the follow-up with large J (Section 2.2.1), (ii) include random frailty effects to model the heterogeneity of and dependence between observations with a known grouping structure, e.g., patients from different hospitals in a multicenter study (Section 2.2.2), (iii) include non-linearly time-varying and non-linear effects of both time-constant or TDCs (Section 2.2.3), (iv) include cumulative, time-lagged, and time-varying effects (i.e., ELRAs) of TDCs (Section 2.2.4).

The basic idea—detailed in the subsections below—is to use penalized estimation of spline basis representations of the respective effects or rates in order to obtain models that are highly flexible and yet allow for reliable inference (cf. our Section 2.3, [Argyropoulos and Unruh, 2015](#), p. 13f). In particular, both the time variation and the non-linear functional shape of covariate effects do not have to be specified *a priori*, instead they are estimated from the data.

2.2.1. Baseline hazard. In [Whitehead \(1980\)](#)'s definition of PEMs (2.2), the baseline hazard $\lambda_0(t)$ is a step function and interval-specific hazard rates λ_{0j} are estimated by including interval-specific dummy variables in the model matrix. The choice of interval cut-points can be problematic ([Demarqui and others, 2008](#)) and choosing a very high number of cut-points increases the number of parameters that need to be estimated. The latter reduces the stability and precision of the individual estimates $\hat{\lambda}_{0j}$ and often leads to implausibly large changes in the estimated baseline hazard from one interval to the next. To avoid these issues, we estimate the baseline log-hazard step function with a spline evaluated at the interval midpoints \tilde{t}_j . First, this means that a very fine partition of the follow-up can be used since the number of parameters for the baseline hazard is given by the dimension of the spline basis, not by the number of intervals J , thus reducing the importance of the location of cut-points. Second, given a sufficiently large spline basis and a fine partition of the intervals, the baseline hazard can be estimated very flexibly, as it changes at each of the closely adjacent interval cut-points. Such closely adjacent cut-points also make the implausibility of restricting the estimated hazard rate to a *step function* with steps at κ_j negligible for practical purposes. Third, large hazard rate changes between adjacent intervals that are not strongly supported by the data are avoided since the estimate of $\log(\lambda_{0j}) = f(\tilde{t}_j)$ is suitably penalized (cf. Section 2.3).

Table 1. Overview of possible effect specifications. $h_p(\cdot)$ denotes a smooth function of its arguments

Effect specification	Description
$f_p(x_p, t) = \beta_p x_{i,p}$	Linear, time-constant effect
$f_p(x_p, t) = h_p(x_{i,p})$	Smooth, time-constant effect
$f_p(x_p, t) = \beta_p x_{i,p} + \beta_{p:t}(x_{i,p} \cdot t)$	Linear, linearly time-varying effect
$f_p(x_p, t) = h_p(x_{i,p}) \cdot t$	Smooth, linearly time-varying effect
$f_p(x_p, t) = x_{i,p} \cdot h_p(t)$	Linear, smoothly time-varying effect
$f_p(x_p, t) = h_p(x_{i,p}, t)$	Smooth, smoothly time-varying effect

2.2.2. *Frailties and random cluster effects.* For data with a known grouping structure, we can extend (2.2) by Gaussian random effects (i.e., frailties) affecting the group-specific hazard rates. Defining $\ell = 1, \dots, L$ as the index for different groups and ℓ_i as the grouping level to which subject i belongs, we write

$$\log(\lambda_i(t|\mathbf{x}_i, \ell_i)) = \log(\lambda_0(t)) + \mathbf{x}'_i \boldsymbol{\beta} + b_{\ell_i}$$

with Gaussian random effects b_{ℓ_i} that capture heterogeneity and dependence induced by the grouping structure. For example, our application contains random effects for the ICUs where the patients were treated. Their inverse variance is estimated from the data as one of the model's smoothing parameters (cf. Section 2.3).

2.2.3. *Flexible covariate effects.* The effects $\mathbf{x}'_i \boldsymbol{\beta}$ of time-constant covariates \mathbf{x}_i in equation (2.1) denote simple linear and time-constant associations with the log-hazard rate. Much more generally, PAMMs can contain possibly non-linear, possibly time-varying effects of time-constant covariates on the log hazard, denoted by $f_p(x_{i,p}, t)$. We extend equation (2.1) to

$$\log(\lambda_i(t|\mathbf{x}_i, \ell_i)) = \log(\lambda_0(t)) + \sum_{p=1}^P f_p(x_{i,p}, t) + b_{\ell_i}.$$

Note that, like the baseline, all time-varying effects in a PAMM are step functions over the partition given by $\kappa_j, j = 0, \dots, J$, so that $f_p(x_{i,p}, t) \equiv f_p(x_{i,p}, \tilde{t}_j) \forall t \in (\kappa_{j-1}, \kappa_j]$. Table 1 shows the variety of covariate effects subsumed into this notation, which range from the simple linear, time constant effects in (2.2) to varying coefficients $x_{i,p} h_p(t)$ or $h_p(x_{i,p}) t$ (Hastie and Tibshirani, 1993) to non-linear, smoothly time-varying covariate effects $h_p(x_{i,p}, t)$ modeled as bivariate function surfaces, parameterized as tensor product smooths (Wood and others, 2013). The smooth functions $f_p(\cdot)$ can be represented as splines of the form $f_p(\cdot) = \sum_{m=1}^M \gamma_{m,p} B_{m,p}(\cdot)$, where $B_{m,p}$ are covariate specific (tensor product) basis functions and $\gamma_{m,p}$ are the associated spline coefficients estimated from the data controlling the effect's shape. Specification $f_p(x_{i,p}, t)$ is the most flexible and should be employed whenever prior information or domain specific knowledge regarding the relationship is absent and a sufficiently large number of cases is available for reliable estimates. Note that time-varying effects imply non-proportional hazards and that (non-)linearity is irrelevant for effects of categorical covariates represented by dummy variables. All this flexibility, of course, raises the question of model selection (cf. discussion in Section 5).

2.2.4. *Exposure-lag-response associations.* For the sake of notational simplicity, we present a model with only one ELRA of a single TDC of primary interest, the *exposure* $z_i(t_e)$. An extension to multiple

ELRAs, however, is straightforward, as will be illustrated in the application example (cf. Section 3). Denote a subject's *exposure history* by $\mathbf{z}_i = (z_i(t_{e,1}), \dots, z_i(t_{e,Q_i}))$, with $t_{e,1}, \dots, t_{e,Q_i}$ assumed to be sufficiently dense over t_e so that all relevant changes of z_i are recorded in \mathbf{z}_i .

Define the set of intervals in which hazard rates are potentially affected by exposure at time t_e as $\mathcal{J}(t, t_e) := \{(\kappa_{j-1}, \kappa_j] : \kappa_{j-1} > t_e + t_{\text{lag}}(t_e) \wedge \kappa_j \leq t_e + t_{\text{lag}}(t_e) + t_{\text{lead}}(t_e)\}$ where $t_{\text{lag}}(t_e)$ and $t_{\text{lead}}(t_e)$ are the lag and lead times as defined in the introduction (cf. Section 1, Equation (1.1)). Vice versa, exposure times potentially affecting hazards in interval j are given by $\mathcal{T}_e(j) := \{t_e : (\kappa_{j-1}, \kappa_j] \in \mathcal{J}(t, t_e)\}$. In the following, we refer to the set $\mathcal{T}_e(j)$ as *lag-lead window* (see top of Figure 1 for examples and intuition). Currently, lag and lead times must be defined *a priori*. In the general definition, $t_{\text{lag}}(t_e)$ and $t_{\text{lead}}(t_e)$ can depend on the exposure time, i.e., the width of the lag-lead window can change over time. In many applications, however, constant lag and lead times will be a reasonable assumption, which can be incorporated in our approach directly, by defining $t_{\text{lag}}(t_e) = t_{\text{lag}} \forall t_e$ and $t_{\text{lead}}(t_e) = t_{\text{lead}} \forall t_e$.

An ELRA $g(\mathbf{z}, t)$ represents the cumulative, time-varying effect of \mathbf{z} on the log-hazard, so we define its contribution to the model's additive predictor as

$$g(\mathbf{z}_i, t) = \int_{\mathcal{T}_e(j)} h(\tilde{t}_j, t_e, z_i(t_e)) dt_e \approx \sum_{q: t_{e,q} \in \mathcal{T}_e(j)} \Delta_q h(\tilde{t}_j, t_{e,q}, z_i(t_{e,q})) \quad \forall t \in (\kappa_{j-1}, \kappa_j], \quad (2.3)$$

where $h(t, t_e, z(t_e))$ is the partial effect of exposure value $z(t_e)$ observed at t_e on the hazard in interval j to which t belongs. The total (cumulative) ELRA effect at time t is then given by the integral of all partial effects of exposures within the window of effectiveness given by $\mathcal{T}_e(j)$, approximated by a weighted sum. Like all time-varying effects in a PAMM, $g(\mathbf{z}_i, t)$ is a step function over the partition of t given by the κ_j , evaluated at the interval mid points $\tilde{t}_j, j = 1, \dots, J$. Quadrature weights $\Delta_q = t_{e,q} - t_{e,q-1}$ for numerical integration are given by the time between two consecutive exposure measurements.

If we restrict the ELRA to be linear in the exposure, i.e., $h(z_i(t_e), t_e, t) = \tilde{h}(t_e, t) \cdot z_i(t_e)$ we can simplify (2.3) to $g(\mathbf{z}_i, t) \approx \sum_{q=1}^Q \tilde{\Delta}_{i,q} \tilde{h}(t_{e,q}, t)$, with $\tilde{\Delta}_{i,q} = \begin{cases} z_i(t_{e,q}) \Delta_q & \text{if } t_{e,q} \in \mathcal{T}_e(j) \\ 0 & \text{else} \end{cases}$. In order to set

up a spline basis for the bivariate function $\tilde{h}(t_e, t)$, we use a tensor product B-spline basis with marginal bases $B_m(t_e), m = 1, \dots, M$ and $B_k(t), k = 1, \dots, K$ defined over the exposure and hazard time domains, respectively. Such tensor product bases allow the construction of multivariate basis functions over disparate dimensions, where M and K delimit the maximal complexity of the ELRA over t_e and t , respectively. Thus, we write $\tilde{h}(t_e, t) = \sum_{m=1}^M \sum_{k=1}^K \gamma_{m,k} B_m(t_e) B_k(t)$.

Combining these two equations, the linear ELRA's basis function expansion is then given by

$$g(\mathbf{z}_i, t) \approx \sum_{m=1}^M \sum_{k=1}^K \gamma_{m,k} \tilde{B}_{i,m}(t_e, t) B_k(t), \quad (2.4)$$

where $\tilde{B}_{i,m}(t_e, t) = \sum_{q=1}^Q \tilde{\Delta}_{i,q} B_m(t_e)$. The penalized estimation of $h(\cdot)$ (cf. Section 2.3) implies an assumption of smoothness on $\tilde{h}(t_e, t)$, which ensures that effects of exposures at consecutive exposure times t_e, t'_e are similar and that effects of a given exposure history \mathbf{z}_i on the hazards in neighboring intervals j, j' are similar as well. Paired with an anisotropic penalty, the specification via a tensor product basis allows for different amounts of roughness over t_e and t . This is crucial if (i) the effect's complexity is different over the two time variables, e.g., if the timing of the exposure is largely irrelevant, but its effect is strongly time-dependent so that $\tilde{h}(t_e, t)$ is approximately constant over t_e but highly variable over t or if (ii) exposure and hazards are observed on different time domains. For example, in this article's motivating application, nutrition information is available on a daily basis only for the second to twelfth calendar days of ICU stay,

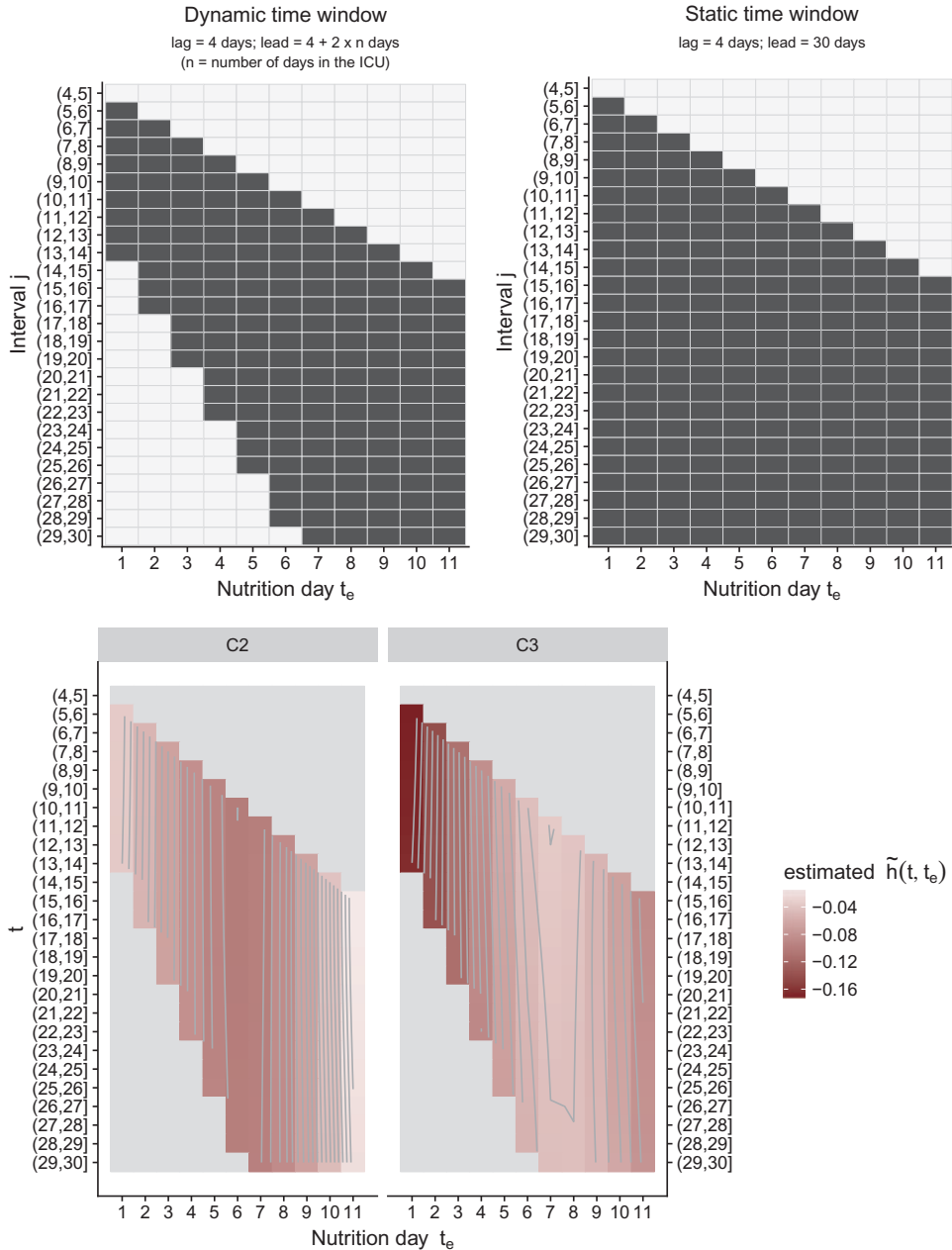


Fig. 1. *Top*: Two possible specifications of the lag-lead window $\mathcal{T}_e(j), j = 5, \dots, 30$. Left panel shows the *dynamic* $\mathcal{T}_e(j)$ ($t_{\text{lag}} = 4$, $t_{\text{lead}} = t_{\text{lag}} + 2 \times t_e$), right panel depicts the *static* $\mathcal{T}_e(j)$ ($t_{\text{lag}} = 4$, $t_{\text{lead}} = 30$) with a longer and constant lead period. This implies that once in effect, partial effects contribute to the cumulative effect until the end of the follow up. *Bottom*: Estimated partial effects surface $\tilde{h}(t, t_e)$ (cf. Equation (3.1)) for nutrition categories C2 (left panel) and C3 (right panel), respectively. Both estimations are based on the dynamic time window definition (Top, left panel). No estimates are given for values (t, t_e) outside of the window of effectiveness $\mathcal{T}_e(j)$. Estimates can be interpreted as “additive change of the log-hazard at time t if patient is in category C2 [C3] instead of category C1 at exposure time t_e ”.

while hazards are modeled beginning 96 h after ICU admission for up to 30 days. More generally, prior exposure could be measured over the course of years or decades and survival after hospitalization observed over months or weeks. Non-linear ELRAs based on smooth functions $\tilde{h}(t, t_e, z(t_e))$ are constructed analogously to bivariate smooth functions in 2.2.3 via straightforward extension to three-dimensional tensor products (Wood, 2006, sec. 4.1.8). The complete specification of the model is then given by:

$$\log(\lambda_i(t|\mathbf{x}_i, \mathbf{z}_i, \ell_i)) = \log(\lambda_0(t)) + \sum_{p=1}^P f_p(x_{i,p}, t) + g(\mathbf{z}_i, t) + b_{\ell_i} \quad (2.5)$$

2.3. Estimation and inference

Reliable likelihood-based methods for the parameter estimation of the proposed model have been developed in Wood (2011) in the context of penalized models of the form $D(\boldsymbol{\gamma}) + \sum_s v_s \boldsymbol{\gamma}' \mathbf{K}_s \boldsymbol{\gamma}$, where $D(\boldsymbol{\gamma})$ is the model deviance, $\boldsymbol{\gamma}$ contains all spline basis coefficients and random effects representing model (2.5), and v_s and \mathbf{K}_s , $s = 1, \dots, S$ are the smoothing parameters and penalty matrices for the individual smooth terms, respectively. In our case, $D(\boldsymbol{\gamma})$ is the deviance of the Poisson GAMM implied by (2.5) in analogy to (2.2). Given $\mathbf{v} = (v_1, \dots, v_S)$, parameter estimates can be obtained by penalized iteratively re-weighted least squares (P-IRLS). To guarantee convergence, Wood (2011) employs P-IRLS based on nested iterations, i.e., after each P-IRLS step, estimation of \mathbf{v} is updated given the current $\boldsymbol{\gamma}$ estimates. Estimation of \mathbf{v} proceeds by numerical optimization of the (restricted) maximum likelihood. Subsequent articles (Marra and Wood, 2011, 2012; Wood, 2012) develop shrinkage based procedures for simultaneous smoothness and variable selection and methods for confidence intervals (CIs) and significance tests for penalized model terms. In the following sections, we extend these methods to the context of PAMMs and, particularly, ELRAs. Note that these methods provide valid inference only for estimates derived from a single, pre-specified model. As always, any subsequent model and variable selection based on fits from multiple model specifications would require appropriate adjustments by methods of post selection inference. Without such adjustments, P -values and CIs will no longer be valid as estimation uncertainty is likely to be underestimated.

2.3.1. Confidence intervals. CIs with good coverage properties for smooth terms are developed in Marra and Wood (2012) and are directly applicable to ELRAs in (2.5). Let $\boldsymbol{\gamma}_z = (\gamma_{1,1}, \dots, \gamma_{M,K})$ be the vector of estimated basis coefficients in (2.4), and $\mathbf{V}_{\hat{\boldsymbol{\gamma}}_z}$ the empirical Bayesian covariance matrix (Wood, 2006, Ch. 4.8) of their estimates $\hat{\boldsymbol{\gamma}}_z$. Let further \mathbf{Z} be the $J \times MK$ design matrix for a specific exposure history \mathbf{z} , where J is the number of intervals into which the follow-up period has been partitioned, and MK is the number of columns associated with the tensor product basis of the ELRA term. The interval-wise CIs of the cumulative effect are given by

$$\mathbf{Z} \hat{\boldsymbol{\gamma}}_z \pm \zeta_{1-\alpha/2} \sqrt{\text{diag}(\mathbf{Z} \mathbf{V}_{\hat{\boldsymbol{\gamma}}_z} \mathbf{Z}^T)} = \hat{\mathbf{g}}_z \pm \zeta_{1-\alpha/2} \widehat{\mathbf{SE}}_z, \quad (2.6)$$

where ζ_q is the q -quantile of the standard normal distribution. In (2.6), $\hat{\mathbf{g}}_z$ is the length J vector estimate of $\mathbf{g}_z = (g(\mathbf{z}, \tilde{t}_1), \dots, g(\mathbf{z}, \tilde{t}_J))$, representing the cumulative effect of exposure history \mathbf{z} on the log hazard and $\widehat{\mathbf{SE}}_z$ its standard errors in intervals $j = 1, \dots, J$. In order to quantify hazard rate differences over time given different exposure histories $\mathbf{z}_{(1)}$ and $\mathbf{z}_{(2)}$, we define $\mathbf{Z} := \mathbf{Z}_{(1)} - \mathbf{Z}_{(2)}$ in (2.6). We can now obtain estimated differences $\hat{g}(\mathbf{z}_{(1)}, \tilde{t}_j) - \hat{g}(\mathbf{z}_{(2)}, \tilde{t}_j)$ in cumulative effects for each interval and respective pointwise CIs for these differences given different exposure histories. We demonstrate this approach in Section 3.3 (cf. Figure 2) and investigate properties of such CIs by means of a simulation study in B.1 of supplementary material available at *Biostatistics* online.

2.3.2. Hypothesis testing. The method introduced in Section 2.3.1 provides a way to quantify the uncertainty associated with a cumulative effect's estimate at individual time-points of the follow-up and thereby, to assess if the effect differs from zero at certain times t . In some applications, however, it is also of interest to assess the overall effect of an ELRA term. To perform such a global test, we can use significance tests for individual smooth terms of the form $H_0 : \mathbf{f}_q = \mathbf{0}$ (Wood, 2012), where \mathbf{f}_q could be any of the model terms in (2.5) and particularly the ELRA $g(\mathbf{z}, t)$ in 2.4. Using the notation from Section 2.3.1 we define the overall test by

$$H_0 : \mathbf{g}_z = \mathbf{0}. \quad (2.7)$$

The general idea of the test is straightforward and uses the representation of the smooth term as a linear transformation of basis coefficients $\boldsymbol{\gamma}_z$ such that $\mathbf{Z}\boldsymbol{\gamma}_z = \mathbf{g}_z$. An appropriate test-statistic has the familiar quadratic Wald-type form $T_r = \hat{\mathbf{g}}_z^T \mathbf{V}_{\mathbf{g}_z}^{r-} \hat{\mathbf{g}}_z$. Here $\mathbf{V}_{\mathbf{g}_z}^{r-}$ is the rank r pseudo inverse of $\mathbf{V}_{\mathbf{g}_z} = \mathbf{Z}\mathbf{V}_{\boldsymbol{\gamma}_z}\mathbf{Z}^T$. The difficult part then becomes choosing the appropriate r in the context of penalized estimation, as naive choices (e.g., rank of $\mathbf{V}_{\mathbf{g}_z}$) lead to reduced power (see Wood, 2012 for details). Given r , which in this context can be a non-integer number, T_r follows a mixture of χ^2 distributions, from which P -values can be obtained routinely (Wood, 2012, p. 4). In B.1 of [supplementary material](#) available at *Biostatistics* online, we show that this *Overall Test* works well when testing individual ELRA terms.

3. ASSOCIATION BETWEEN CALORIC INTAKE AND MORTALITY IN ICU PATIENTS

3.1. Data and objective

We apply our method in a retrospective analysis of a large international multicenter study with $n = 9661$ critically ill patients (after pre-processing and application of exclusion criteria) with a maximal follow up of 60 days or until release from hospital (cf. Heyland and others, 2011). On the day of admission (Day 0), goal calories for each patient were determined by a nutritionist or physician and the actual caloric intake provided by the hospital staff was recorded for a maximum of 11 *calendar* days after the date of ICU admission, which we denote by $t_e \in \{1, \dots, 11\}$. The follow-up time t was partitioned into 24 h periods starting with ICU admission. We are interested in the association between caloric adequacy and acute mortality, that is, mortality within $t_{\max} = 30$ days (720 h) after ICU admission. In total, 1974 (20.4%) patients died within this period. For our main analysis, we assume that patients released from the hospital survived at least until $t = 30$ (see Discussion in Appendix A.8 of [supplementary material](#) available at *Biostatistics* online). We only included patients that survived at least 96 h (4 days), consequently, we began evaluation in interval (4, 5]. For the purpose of the analysis presented here, all patients still alive after $t = 30$ were censored.

3.2. Modeling approach

We adjusted for the most relevant potential confounders, including subject specific covariates age, body mass index (BMI), sex, diagnosis at admission and admission category, and the Apache II Score (an overall measure of the patients' health status at admission) as well as patient unrelated covariates like year of admission and a random effect (Gaussian frailty) for the ICUs. Since we model the mortality risk beginning in interval (4, 5] (due to application of exclusion criteria), we also included variables that describe the patients' ICU stay up to that point, namely number of days under MV and number of days with additional OI, number of days with parenteral nutrition (PN), and number of days receiving Propofol (PF) on the first three full calendar days of the ICU stay, respectively. To be able to compare different caloric intakes independently of a patient's weight and caloric requirements, we defined a patient's caloric adequacy as $CA = \frac{\text{actual daily caloric intake (in kcal)}}{\text{goal calories (in kcal)}}$. In order to account for unmeasured additional OI, we used a discretized

version of CA , with three categories $C1 : CA < 30\%$, $C2 : 30\% \leq CA < 70\%$, and $C3 : CA \geq 70\%$. If patients received additional OI, they were moved up one category (more details in Appendix A.1 and Table S3 of [supplementary material](#) available at *Biostatistics* online).

The effect of nutrition is represented in the model by two ELRAs $g_{C2}(\mathbf{z}_i^{C2}, t)$ and $g_{C3}(\mathbf{z}_i^{C3}, t)$ for dummy variables $z_i^{C2}(t_e), z_i^{C3}(t_e)$ indicating whether nutrition at time t_e was in category $C2$ or $C3$, respectively. Both terms have the structure defined in (2.4). Category $C1$ is the “reference” category, thus direct interpretation of g_{C2} and g_{C3} is only possible with respect to a (hypothetical) patient that received $C1$ on all 11 protocol days. Equation (3.1) shows the model specification:

$$\log(\lambda_i(t|\mathbf{x}_i, \mathbf{z}_i, \ell_i)) = \log(\lambda_0(t)) + \sum_{p=1}^P f_p(x_{i,p}, t) + g_{C2}(\mathbf{z}_i^{C2}, t) + g_{C3}(\mathbf{z}_i^{C3}, t) + b_{\ell_i}, \quad (3.1)$$

where $\lambda_0(t)$ represents the baseline hazard rate, $\sum_{p=1}^P f_p(x_{i,p}, t)$ incorporates all linear and non-linear effects of time-constant covariates, g_{C2} and g_{C3} are non-linear time-varying cumulative effects of the nutritional intake (see Section 2.2.4 for details), and b_{ℓ_i} is an independent identically distributed Gaussian random intercept term attributed to different ICUs in the data set. Details on model specification and on the estimated confounder effects are in A.3 and A.5 of [supplementary material](#) available at *Biostatistics* online. All non-linear functions of time-constant covariates $f_p(\mathbf{x}_{\cdot,p}, t)$ were estimated using P-Splines (Eilers and Marx, 1996) with second order difference penalties and $M = 10$ basis functions on equidistant knots. For the $\tilde{h}(t_e, t)$ terms associated with the ELRA (cf. Equation (2.4)), $M = K = 5$ basis functions were used for each dimension and first order difference penalties were used for exposure time t_e .

The lag-lead window $\mathcal{T}_e(j)$ (see Equation (3.1)) was defined based on substantive considerations with $t_{\text{lag}}(t_e) \equiv 4$ and $t_{\text{lead}}(t_e) = t_{\text{lag}} + 2 \cdot t_e$ and is depicted in Figure 1 (top, left). Viewed column-wise, each panel in Figure 1 shows the intervals at which a specific protocol day ($t_e \in \{1, \dots, 11\}$) can *potentially* affect the estimated hazard. Viewed row-wise, the figures show protocol days t_e , which contribute to the cumulative nutrition effect in interval j . We will refer to this specification as *dynamic* lag-lead $\mathcal{T}^{\text{dynamic}}$. The clinical considerations underlying this definition are summarized in A.4 of [supplementary material](#) available at *Biostatistics* online. To investigate the impact of a possible misspecification of \mathcal{T} , we conducted a simulation study (cf. B.1 of [supplementary material](#) available at *Biostatistics* online), which closely resembles the data structure and effects of this application example. Section 5 provides discussion on \mathcal{T} and some justification for the choice of t_{lag} .

3.3. Results

We are mostly interested in the relationship between caloric intake and survival. Therefore, in the following, we will only present the results of this association. Estimated confounder effects are presented in A.5 of [supplementary material](#) available at *Biostatistics* online. The estimated partial effect surfaces $\hat{h}_{C2}(t_e, t)$ and $\hat{h}_{C3}(t_e, t)$ for cumulative effects g_{C2} and g_{C3} , respectively, are presented at the bottom of Figure 1 (see also Figure S5 of [supplementary material](#) available at *Biostatistics* online for CIs). Both effects are estimated to be almost constant over t (vertical axis) and vary meaningfully only over t_e (horizontal axis). Estimates can be interpreted as the estimated difference in log-hazard rates at time t between patients who were in category $C2$ ($C3$) at time t_e compared to a patient who was in category $C1$ at time t_e if the two had identical exposures at all other exposure time-points $t'_e \neq t_e$ and identical values for random effects and all other covariates $x_p, p = 1, \dots, P$. For example, a patient with $C2$ instead of $C1$ at $t_e = 7$ is estimated to have a decreased hazard by a factor of about $e^{-0.1} \approx 0.90$ over $t \in (11, 30]$, compared with a patient with identical covariate values and otherwise identical exposure history. Direct interpretation of these partial effects, however, is hindered by the fact that hazard at time t is affected by different numbers of exposures depending on the structure of $\mathcal{T}_e(j)$, since the total effect of exposure history \mathbf{z} on the hazard is given by

Table 2. Overview of evaluated comparisons with nutrition categories C1 (lower), C2 (mid) and C3 (upper) as defined in Table S3 [supplementary material](#) available at Biostatistics online

Comparison	$\mathbf{z}_{(1)}$	$\mathbf{z}_{(2)}$
Comparison A	Days 1–11: C1	Days 1–4: C1, Days 5–11: C2
Comparison B	Days 1–11: C1	Days 1–11: C2
Comparison C	Days 1–4: C1, Days 5–11: C2	Days 1–11: C2
Comparison D	Days 1–11: C1	Days 1–11: C3
Comparison E	Days 1–11: C2	Days 1–4: C2, Days 5–11: C3
Comparison F	Days 1–11: C2	Days 1–11: C3

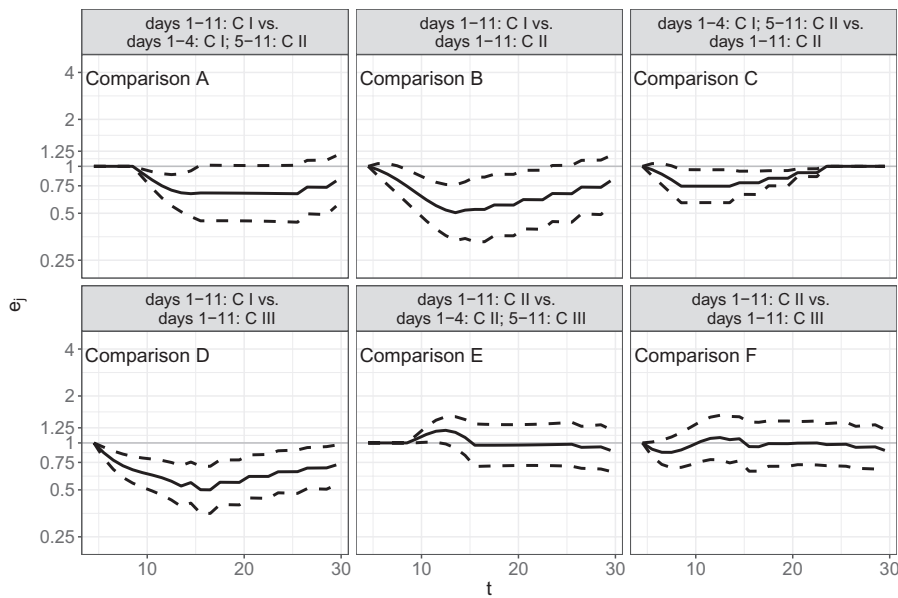


Fig. 2. Estimated hazard ratios $\hat{e}_j (\pm 2\widehat{SE}_{\hat{e}_j})$ for the comparison of different nutrition protocols (all other covariates being equal), as defined in equation (3.2). Each facet depicts one comparison. The comparisons are indicated in the facet headers and summarized in Table 2. $\hat{e}_j < 1$ indicate an increase in the hazard rate for the first protocol compared with the second. Dashed lines depict approximate 95% CI as described in Section 2.3.1.

$g_{C2}(\mathbf{z}^{C2}, t) + g_{C3}(\mathbf{z}^{C3}, t)$, and since this effect represents the difference in the log-hazard compared with a patient with constant category C1 nutrition. For these reasons, we analyze and interpret estimated hazard ratios between hypothetical patients with different clinically relevant exposure histories that can easily be computed from the partial effects. Therefore, we show the cumulative effects of nutrition, evaluated at interval midpoints $\tilde{t} \in \{4.5, 5.5, \dots, 29.5\}$, as hazard ratios

$$e_j = \frac{\lambda(\tilde{t}_j | \mathbf{z}_2)}{\lambda(\tilde{t}_j | \mathbf{z}_1)} = \exp(\mathbf{g}_{\mathbf{z}_2} - \mathbf{g}_{\mathbf{z}_1}) \quad (3.2)$$

for patients with different nutrition protocols \mathbf{z}_1 and \mathbf{z}_2 and identical values for all other covariates. Six clinically relevant comparisons are summarized in Table 2. The results are depicted in Figure 2 and suggest

that (i) hypocaloric (category C1) nutrition is associated with increased hazard rates throughout the follow-up period (Comparisons B, D and to a lesser extent Comparison A); (ii) based on this model, moving from constantly medium (C2) to constantly full (C3) nutrition is not associated with a decrease of the hazard rate (Comparisons E, F); (iii) the (small) hazard rate increases associated with hypocaloric nutrition in the first few days of the ICU stay may persist for up to 25 days after ICU admission (Comparison C).

In A.7 [supplementary material](#) available at *Biostatistics* online, we compare our main model with two alternative models, which only use time-aggregated cumulative nutrition information as TDCs instead of fitting cumulative effects of the nutrition exposure histories. The findings suggest that, for these data, results qualitatively similar to the ones presented here could have been obtained using simpler approaches. We argue, however, that the suggested approach is preferable: Since the true association structure is typically not known in advance and since the penalized estimation of our very flexible proposal seems to be able to avoid overfitting in settings where the association has a simple linear or time-constant structure. Allowing for complex associations entails little cost and helps to avoid both possible model misspecification and issues of post-selection inference that come with *post hoc* comparisons of simpler models using different pre-specified variations of cumulating the (effects of) TDCs (cf. A.7 of [supplementary material](#) available at *Biostatistics* online for details).

4. SIMULATION STUDY

We performed extensive simulation studies to investigate the performance of the proposed modeling approach. As general performance of this model class has already been evaluated in other publications (cf. [Wood, 2011](#) for GAMs and [Argyropoulos and Unruh, 2015](#) for their application to survival analysis), our simulation focuses on the evaluation of ELRAs and is divided in two parts. In *Part A* (Section 4.2), we demonstrate the ability of our approach to estimate the effects of a complex nonlinear DLNM as well as a time-varying extension thereof, i.e., a three-dimensional function of the form $g(\mathbf{z}, t) = \int h(t, t - t_e, z(t_e)) dt_e$.

Simulation *Part B* is modeled after the application example, especially with respect to data structure and the structure, magnitude, and shape of simulated effects and focuses on estimation of cumulative effects $g(\mathbf{z}, t)$ instead of partial effects $h(t, t_e, z(t_e))$. Details on data generation and results of Simulation Part B are provided in B.1 of [supplementary material](#) available at *Biostatistics* online. They show that simulated effects are generally estimated well and CIs as well as hypothesis tests discussed in Sections 2.3.1 and 2.3.2 have good properties, often even when the model is misspecified (cf. B.1.2 and Figures S19, S20 of [supplementary material](#) available at *Biostatistics* online).

4.1. Data generation

In both simulation parts, random survival times were drawn from the piece-wise exponential distribution. To do so, one needs to specify a vector of piece-wise constant hazards $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_J)$ in intervals defined by time-points $\kappa = (\kappa_0, \dots, \kappa_J)$. In general, λ can be defined through a function of time t , covariates \mathbf{x} and exposure histories \mathbf{z} such that $\lambda_j, j = 1, \dots, J$ is given by $\lambda(t|\mathbf{x}, \mathbf{z}) = f(t, \mathbf{x}, \mathbf{z})$, evaluated at times $t = \kappa_j, j = 1, \dots, J$. The algorithm by which survival times are drawn from the piece-wise exponential distribution ($t \sim \text{PEXP}(\lambda, \kappa)$) is described in detail in B of [supplementary material](#) available at *Biostatistics* online, specifically Table S7. Note that drawing from PEXP generates continuous event times, thus for each subject we obtain tuples (t_i, x_i, \mathbf{z}_i) .

4.2. Simulation Part A

We adapt a simulation study performed in [Gasparrini and others \(2017\)](#) for Poisson responses to the context of time-to-event data. Our objective is to demonstrate the ability of our method to estimate DLNMs and

time-varying DLNMs. We distinguish two scenarios, both of which simulate data from a model with a single cumulative effect of a TDC.

Scenario (1): Simulate a “classical” DLNM for survival data via

$$\log(\lambda(t, \mathbf{z})) = \log(\lambda_0(t)) + \int h(t - t_e, z(t_e)) dt_e \quad (4.1)$$

Scenario (2): Extend scenario (1) such that the ELRA additionally varies over time

$$\begin{aligned} \log(\lambda(t, \mathbf{z})) &= \log(\lambda_0(t)) + g(\mathbf{z}, t) = \log(\lambda_0(t)) + \int \tilde{h}(t, t - t_e, z(t_e)) dt_e \\ &= \log(\lambda_0(t)) + \int f(t) \cdot h(t - t_e, z(t_e)) dt_e \end{aligned} \quad (4.2)$$

For the functional shape of $h(t - t_e, z(t_e))$ in (4.1) and (4.2) we used the “complex” ELRA, which is depicted at the right panel (“TRUTH”) in Scenario (1) in Figure 3 (see also Figure 1 in [Gasparrini and others, 2017](#)). With Φ_{μ, σ^2} the density function of the normal distribution with mean μ and variance σ^2 , the mathematical representation of h is given by the functions $f_1(z(t_e)) = \Phi_{1.5, 2}(z(t_e)) + 1.5 \cdot \Phi_{7.5, 1}(z(t_e))$, $f_2(t - t_e) = 15 \cdot \Phi_{8, 10}(t - t_e)$, $f_3(t - t_e) = 5 \cdot (\Phi_{4, 6}(t - t_e) + \Phi_{25, 4}(t - t_e))$ and finally $h(t - t_e, z(t_e)) = \begin{cases} f_1(z(t_e)) \cdot f_3(t - t_e) & \text{if } z(t_e) \geq 5 \\ f_1(z(t_e)) \cdot f_2(t - t_e) & \text{if } z(t_e) < 5 \end{cases}$ (cf. Appendix in [Gasparrini and others, 2017](#)).

In (4.2), we set $f(t) = -\cos(\pi t / t_{\max})$, such that the partial effects are negative at the beginning and then become positive after the first half of the follow up (cf. right panel (“TRUTH”) of Scenario (2) in Figure 3). In both Scenarios, in addition to fitting a correctly specified model to the simulated data, we also fit a underspecified model for comparison. In Scenario (1), we fit a WCE with partial effects $h(t - t_e)z(t_e)$, thus neglecting the non-linearity in $z(t_e)$. In Scenario (2), we fit a standard (non time-varying) DLNM with partial effects $h(t - t_e, z(t_e))$, thus neglecting the additional time-variation. All models were estimated by PAMMs.

Evaluation was performed by graphical comparison of the estimated $\hat{h}_{t, t_e, z(t_e)} \equiv \hat{h}(t - t_e, z(t_e))$ [Scenario 2: $\hat{h}(t, t - t_e, z(t_e))$] to the respective true (simulated) surface $h_{t, t_e, z(t_e)}$ and with summary statistics $RMSE = \frac{1}{R} \sum_{r=1}^R \sqrt{\frac{1}{400} \sum_{t-t_e=0}^{40} \sum_{z(t_e)=0}^{10} (h_{t, t_e, z(t_e)} - \hat{h}_{t, t_e, z(t_e)})^2}$ and $\text{coverage}_\alpha = \frac{1}{R} \sum_{r=1}^R \left[\frac{1}{400} \sum_{t-t_e=0}^{40} \sum_{z(t_e)=0}^{10} I \left(h_{t, t_e, z(t_e)} \in [\hat{h}_{t, t_e, z(t_e)} \pm \zeta_{1-\alpha/2} \hat{\sigma}_{\hat{h}_{t, t_e, z(t_e)}}] \right) \right]$, where ζ_q is the q -quantile of the standard normal distribution and $\hat{\sigma}_{\hat{h}_{t, t_e, z(t_e)}}$ is the estimated standard deviation of the partial effect estimate. In Scenario (2) RMSE and coverage are additionally averaged over each time-point $t = 1, \dots, 40$.

The results for Scenario (1) are shown at the top of Figure 3. The simulated effects are estimated very well when the model is specified correctly—“PAM (DLNM)” has a mean RMSE of ≈ 0.037 and close to nominal coverage of $\approx 97\%$. For comparison, the misspecified model “PAM (WCE)”, which wrongly assumes linearity in $z(t_e)$, yields an 1.6 times increased RMSE (≈ 0.06) and a coverage of $\approx 36\%$. The results for Scenario (2) are presented at the bottom of Figure 3. The tri-variate function $h(t, t - t_e, z(t_e))$ was not estimated quite as precisely by the correctly specified model (“PAM (TV DLNM)”) as in the time-constant Scenario (1). In general, however, the ELRA, as well as its variation over time were fit very well, which illustrates the ability of our approach to estimate truly three-dimensional cumulative effects (4.2) (mean RMSE of about 0.024 and mean coverage of about 0.96). The simpler model (“PAM (DLNM)”), that ignores possible time-variation, consequently has the same value at all time-points and is close to zero, i.e., the estimate averages out the risk reducing effects at the beginning and the risk

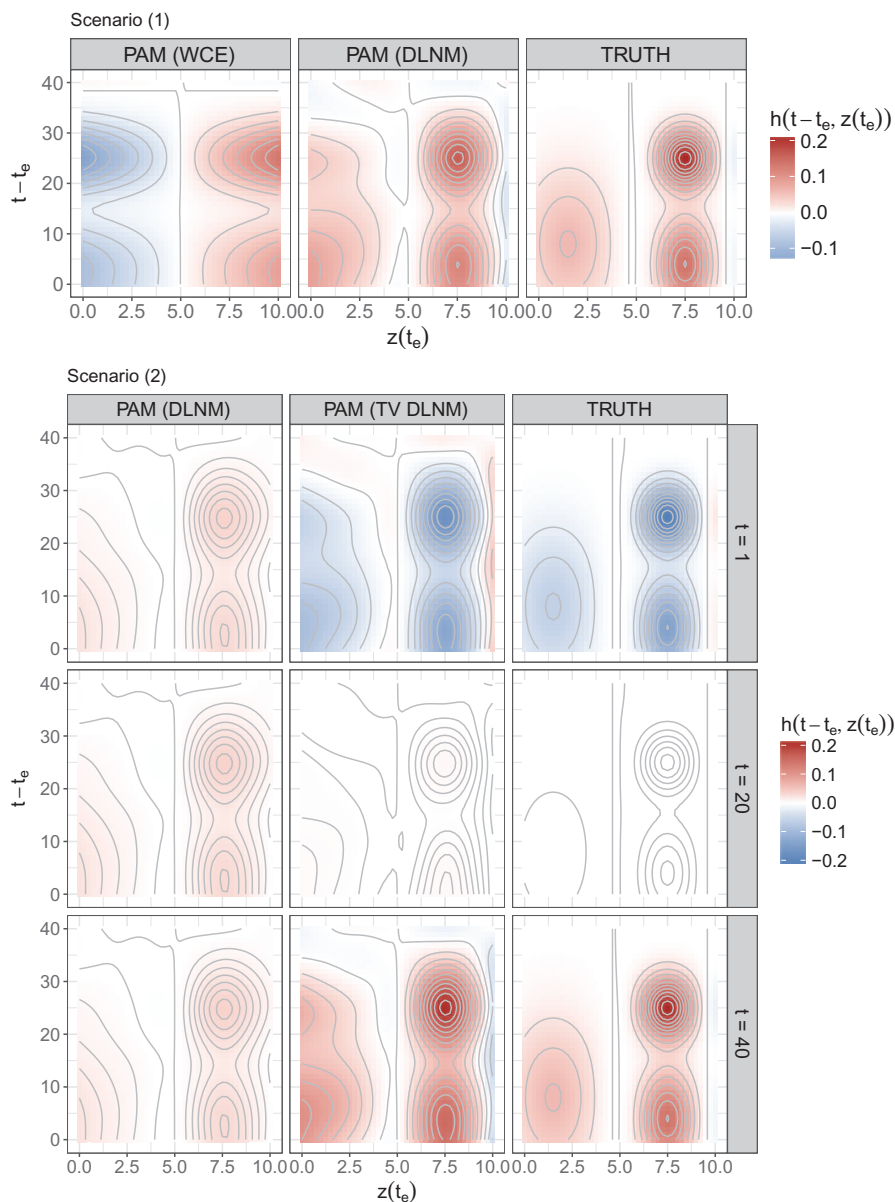


Fig. 3. Simulation Part A *Scenario (1)*: estimated WCE ($\hat{h}(t - t_e, z(t_e))$; left), estimated DLNM ($\hat{h}(t - t_e, z(t_e))$; middle), and true simulated bi-variate effect surfaces ($h(t - t_e, z(t_e))$; right) for different values of latency $t - t_e$ and exposure $z(t_e)$. *Scenario (2)*: estimated (left and mid panel) and true simulated (right panel) bi-variate effect surfaces $\hat{h}(t - t_e, z(t_e))$, and $h(t - t_e, z(t_e))$, respectively, for different values of latency $t - t_e$, exposures $z(t_e)$ and for follow-up times $t \in \{1, 20, 40\}$. In both Scenarios, displayed estimates are obtained by averaging over all estimates from $R = 500$ simulation runs.

increasing effects at the end of the follow-up, respectively ($\text{RMSE} \approx 0.038$, coverage $\approx 71\%$). Note that RMSEs for Scenario (2) are generally smaller, due to small effects and estimates toward the middle of the follow-up.

5. DISCUSSION

By embedding the concept of PEMs into the framework of additive models, we define a very versatile model class for life-time data analysis that inherits the robust and flexible tools for modeling, estimation and validation available for GAMMs. In contrast to traditional PEMs, the baseline and time-varying effects are represented as flexible, potentially non-linear penalized splines. Our general approach to flexibly modeling cumulative effects of TDCs (ELRAs: exposure-lag-response associations) takes into account the entire exposure history, i.e., both the timing and amount of exposure. The proposed presentation of ELRAs in form of hazard ratio trajectories for pairs of realistic exposure histories provides an accessible alternative to established visualization techniques like contour or wire-frame plots.

The general formulation, we propose defines a broad model class that includes previous approaches for cumulative effects like the WCE model (Sylvestre and Abrahamowicz, 2009) and the DLNM (Gasparrini and others, 2017) as special cases. The application example (Section 3) is a direct demonstration of this generalization, and—although the improvement in predictive accuracy was small since the estimated partial effects did not vary much over t in this case — also provides an example for the estimation of a more flexible WCE model, as the weight function $h(t, t_e)$ was allowed to depend on the specific combination of t and t_e instead of restricting the dependency to the latency $t - t_e$ a priori. Simulation study Part B (cf. Section 4.2) additionally illustrates the practical feasibility of a “time-varying” extension of the DLNM.

In general, our simulation studies (Section 4) confirm that our method is able to estimate complex ELRAs and is relatively robust to misspecification. When no true exposure effect is present, both the coverage of the proposed CIs for all comparisons and the Type I error rate of the hypothesis tests maintained near nominal levels, regardless of the specification of the penalty and the correctness of the specified lag and lead times (cf. supplement Figures S19 (upper panel, Setting IV) and S20), but CIs can sometimes have sub-nominal coverage for misspecified non-null models, especially in the case of P_1 penalties. It is also apparent that a misspecification of lag and lead times can induce bias, potentially causing an underestimation of effects at the end of the follow up (if the lead time was specified as too short), or an overestimation (if the lead time was specified as too long). Going forward, data driven selection of the relevant time window is our most important research goal. One approach for such a procedure could include additional penalties similar to the double penalty approach by Obermeier and others (2015) for distributed lag models.

From a practical perspective, the interpretation of (cumulative) effects of TDCs is problematic whenever the exogeneity of such variables is unclear. For example, although nutrition is administered by the hospital staff, the amount of nutrition provided is likely to depend on patients’ health status, at least to some degree: patients undergoing procedures due to life-threatening complications presumably receive less calories or feeding could be stopped due to the decision to withdraw life support. Handling such variables always requires a trade-off with respect to the recency of the covariate (Crowder, 2012, Ch. 3.6.) that may result in better adjustment for confounding for more recent values of the covariate, but may also be fully indicative of the outcome and thus induce indication bias (Signorello and others, 2002). In our application, we address this issue by including a minimum lag time of 4 days for the nutritional effects. In future research, our method could be combined with frameworks that take into account that TDCs can simultaneously be confounders and mediators of past exposures. Xiao and others (2014) recently presented one such extension of the WCE to marginal structural models. Additionally, as discussed in Supplement A.8, extending the proposed estimation of cumulative effects to the competing risks setting, similar to Danieli and Abrahamowicz (2017), would also be of great value for practical applicability.

Finally, the flexibility of this model class implies challenges with respect to variable and model selection. Wynant and Abrahamowicz (2014) address the issue of model selection for non-linear and/or time-varying effects in the context of partial likelihood models, and they demonstrate the usefulness of backward elimination for this purpose. Boosting based approaches, which have been shown to perform well for similarly complex GAMs in the context of functional data analysis and for complex time-to-event models might also be of interest here.

6. SUPPLEMENTARY MATERIAL

Supplementary material is available at <http://biostatistics.oxfordjournals.org>. Data and code to reproduce the analyses are available at <https://github.com/adibender/elra-biostats>.

ACKNOWLEDGMENTS

The authors would like to thank Daren Heyland for providing the data set and useful discussion. We also want to thank the referees for their detailed and helpful comments, which greatly improved the manuscript. *Conflict of Interest:* None declared.

FUNDING

Fabian Scheipl was supported by the German Research Foundation through the Emmy Noether Programme (GR 3793/1-1 to Sonja Greven).

REFERENCES

- ARGYROPOULOS, C. AND UNRUH, M. L. (2015). Analysis of time to event outcomes in randomized controlled trials by generalized additive models. *PLoS One* **10**, e0123784.
- BERHANE, K., HAUPTMANN, M. AND LANGHOLZ, B. (2008). Using tensor product splines in modeling exposure-time-response relationships: application to the colorado plateau uranium miners cohort. *Statistics in Medicine* **27**, 5484–5496.
- CROWDER, M. J. (2012). *Multivariate Survival Analysis and Competing Risks*. Chapman & Hall/CRC Texts in Statistical Science. Hoboken: CRC Press.
- DANIELI, C. AND ABRAHAMOWICZ, M. (2017). Competing risks modeling of cumulative effects of time-varying drug exposures. *Statistical Methods in Medical Research*, doi:10.1177/0962280217720947.
- DEMARQUI, F. N., LOSCHI, R. H. AND COLOSIMO, E. A. (2008). Estimating the grid of time-points for the piecewise exponential model. *Lifetime Data Analysis* **14**, 333–356.
- EILERS, P. H. C. AND MARX, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**, 89–121.
- FRIEDMAN, M. (1982). Piecewise exponential models for survival data with covariates. *The Annals of Statistics* **10**, 101–113.
- GASPARRINI, A. (2014). Modeling exposure-lag-response associations with distributed lag non-linear models. *Statistics in Medicine* **33**, 881–899.
- GASPARRINI, A., SCHEIPL, F., ARMSTRONG, B. AND KENWARD, M. G. (2017). A penalized framework for distributed lag non-linear models. *Biometrics* **73**, 938–948.
- GERDS, T. A., KATTAN, M. W., SCHUMACHER, M. AND YU, C. (2013). Estimating a time-dependent κ^2 concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine* **32**, 2173–2184.

- HASTIE, T. AND TIBSHIRANI, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)* **55**, 757–796.
- HEYLAND, D. K., CAHILL, N. AND DAY, A. G. (2011). Optimal amount of calories for critically ill patients: depends on how you slice the cake! *Critical Care Medicine* **39**, 2619–2626.
- MARRA, G. AND WOOD, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis* **55**, 2372–2387.
- MARRA, G. AND WOOD, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics* **39**, 53–74.
- OBERMEIER, V., SCHEIPL, F., HEUMANN, C., WASSERMANN, J. AND KÜCHENHOFF, H. (2015). Flexible distributed lags for modelling earthquake data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **64**, 395–412.
- SIGNORELLO, L. B., MCLAUGHLIN, J. K., LIPWORTH, L., FRIIS, S., SØRENSEN, H. T. AND BLOT, W. J. (2002). Confounding by indication in epidemiologic studies of commonly used analgesics. *American Journal of Therapeutics* **9**, 199–205.
- SYLVESTRE, M.-P. AND ABRAHAMOWICZ, M. (2009). Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. *Statistics in Medicine* **28**, 3437–3453.
- WHITEHEAD, J. (1980). Fitting Cox's regression model to survival data using GLIM. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **29**, 268–275.
- WOOD, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: CRC Press.
- WOOD, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**, 3–36.
- WOOD, S. N. (2012). On p-values for smooth components of an extended generalized additive model. *Biometrika* **100**, 221–228.
- WOOD, S. N., SCHEIPL, F. AND FARAWAY, J. J. (2013). Straightforward intermediate rank tensor product smoothing in mixed models. *Statistics and Computing* **23**, 341–360.
- WYNANT, W. AND ABRAHAMOWICZ, M. (2014). Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis. *Statistics in Medicine* **33**, 3318–3337.
- XIAO, Y., ABRAHAMOWICZ, M., MOODIE, E. E. M., WEBER, R. AND YOUNG, J. (2014). Flexible marginal structural models for estimating the cumulative effect of a time-dependent treatment on the hazard: Reassessing the cardiovascular risks of didanosine treatment in the Swiss HIV cohort study. *Journal of the American Statistical Association* **109**, 455–464.

[Received February 6, 2017; revised January 15, 2018; accepted for publication January 16, 2018]