HHS Public Access

Author manuscript

*Stat Modelling*. Author manuscript; available in PMC 2015 December 01.

Published in final edited form as:

*Stat Modelling*. 2015 June 1; 15(3): 256–278. doi:10.1177/1471082X14565526.

# Cox Regression Models with Functional Covariates for Survival Data

**Jonathan E. Gellar**[1], **Elizabeth Colantuoni**[1], **Dale M. Needham**[2], and **Ciprian M. Crainiceanu**[1]

[1]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

[2]Pulmonary & Critical Care Medicine, and Physical Medicine & Rehabilitation, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

## Abstract

We extend the Cox proportional hazards model to cases when the exposure is a densely sampled functional process, measured at baseline. The fundamental idea is to combine penalized signal regression with methods developed for mixed effects proportional hazards models. The model is fit by maximizing the penalized partial likelihood, with smoothing parameters estimated by a likelihood-based criterion such as AIC or EPIC. The model may be extended to allow for multiple functional predictors, time varying coefficients, and missing or unequally-spaced data. Methods were inspired by and applied to a study of the association between time to death after hospital discharge and daily measures of disease severity collected in the intensive care unit, among survivors of acute respiratory distress syndrome.

### Keywords

functional data analysis; survival analysis; Cox proportional hazards model; nonparametric statistics; intensive care unit

## 1 Introduction

We introduce the proportional hazard functional regression model for data where the outcome is the possibly censored time to event and the exposure is a densely sampled functional process measured at baseline. The methodology was inspired by and applied to a study of the association between survival time after hospital discharge of survivors of acute respiratory distress syndrome (ARDS), and daily measures of organ failure during hospitalization.

The main innovation of our approach is to provide a fast and easy to use Cox model for functional regression based on modern developments in nonparametric smoothing, survival analysis methods and software, and functional data analytic concepts. More precisely, the method we propose has three important characteristics: 1) it employs penalized splines

**Address for correspondence:** Jonathan E. Gellar, Department of Biostatistics, Bloomberg School of Public Health, 615 N. Wolfe. St., #E3031, Baltimore, MD 21205, USA. jgellar1@jhu.edu. **Phone:** (+1) 410 502 3357. **Fax:** (+1) 410 955 0958..

(Eilers and Marx, 1996; O'sullivan et al., 1986; Ruppert et al., 2003) to model the functional coefficient(s); 2) it treats the proportional hazards model that incorporates the functional parameter as a hazard rate with a mixed effects format on the log scale and uses modern survival data fitting techniques that incorporate nonparametric smoothing and frailty regression (Gray, 1992; Verweij and Van Houwelingen, 1993; Therneau and Grambsch, 1998); and 3) it estimates the amount of smoothing of the functional parameter using a likelihood-based information criterion such as AIC (Verweij and Houwelingen, 1994; Therneau et al., 2003), and thus avoids the use of principal component truncation, which may lead to highly unstable coefficient shapes.

Functional data regression is under intense methodological development (Marx and Eilers, 1999; Cardot et al., 1999, 2003; James, 2002; Müller and Stadtmüller, 2005; Cardot and Sarda, 2005; Ferraty and Vieu, 2006; Reiss and Ogden, 2007; Ramsay et al., 2009; James et al., 2009; Goldsmith et al., 2011; Harezlak and Randolph, 2011; Ferraty, 2011; McLean et al., 2012), though there are only a few modeling attempts in the case when outcome is time-to-event data. Probably the first paper to consider this topic was James (2002), who used a functional generalized linear model to model right-censored life expectancy, where censored outcomes are handled using the procedure of Schmee and Hahn (1979). The procedure has two steps: 1) use a truncated normal distribution to estimate the unobserved failure times; and 2) incorporate these estimates in standard least squares equations. The entire procedure is then iterated using the EM algorithm. This method for accounting for censored survival times is similar in spirit to the better known Buckley-James estimator (Buckley and James, 1979; Miller and Halpern, 1982). Müller and Zhang (2005) developed an alternative model for the mean remaining lifetime given a longitudinal covariate up to a given time, though their model does not allow for censored outcomes. Chiou and Müller (2009) incorporate the ideas of functional data analysis by modeling hazard rates as random functions, but they do not allow for functional predictors.

In parallel with these efforts in functional data analysis, important developments have been achieved by researchers in survival data analysis. Specifically, non-parametric smoothing approaches have been introduced to account for the possibly nonlinear effects of scalar covariates in a proportional hazards model. Therneau and Grambsch (1998) showed how to fit a model with a smooth covariate effect using penalized splines, and how this model was closely related to the more well-studied gaussian frailty model. Kneib and Fahrmeir (2007) developed an expanded mixed model that allows for non-parametric effects of scalar covariates, spatial effects, time-varying covariates, and frailties. Strasak et al. (2009) allowed for smooth effects of time-varying covariates. These techniques have now matured, are accompanied by high quality software (Belitz et al., 2013; Therneau, 2012, 2014), and continue to be developed. Our approach will take advantage of decades of development in survival data analysis, functional data analysis, and nonparametric smoothing. In particular, we will identify the most robust, easiest to use combination of approaches to achieve our goal: modeling survival data with nonparametric functional regression parameters.

Thus, we take advantage of these methods in order to model the e ect of a functional covariate on a (possibly censored) survival time, by including the functional covariate as a term in a Cox proportional hazards model (Cox, 1972). This approach has several

advantages. First, the proportional hazards model is one of the most popular regression models ever, has been well-studied, and has a form that is familiar to a general audience. Second, the Cox model is widely considered to be the standard in regression for survival data due to its interpretability, applicability, and inferential flexibility. Third, it allows simple extensions of the fitting procedure to incorporate functional covariates. These properties allowed us to develop a model that is easy to implement and computationally efficient. Indeed, our software will be made publicly available as part of the pcox package in R (R Development Core Team, 2011), and fitting the model requires only one line of code.

We now briefly review some of the existing functional regression techniques that are relevant to this article. Scalar-on-function regression models the relationship between a scalar outcome and a functional covariate. Suppose that $\{Y_i\}$ are a set of scalar outcomes, $\{X_i(t)\}$ are functional covariates defined on the interval [0, 1], and $\{Z_i\}$ non-functional covariates, where $i \in \{1, 2, \ldots, N\}$. Then the generalized functional linear model that relates $Y_i$ to $\boldsymbol{Z}_i$ and $X_i(t)$ is $g(\mu_i) = \alpha + \boldsymbol{Z}_i\gamma + \int_0^1 X_i(t)\beta(t)\,dt$, where $Y_i$ follows an exponential family distribution with mean $\mu_i$, and $g(\cdot)$ is an appropriate link function (McCullagh and Nelder, 1989; James, 2002; Müller and Stadtmüller, 2005). The key feature of this model that differentiates it from a standard (non-functional) generalized linear model is the integral term, $\int_0^1 X_i(t)\beta(t)\,dt$, which captures the contribution of the functional covariate $X_i(t)$ towards $g(\mu_i)$. The integration essentially serves as a weighting mechanism for $X_i(t)$, where the weights are given by the coefficient function $\beta(t)$. Thus, we think of $\beta(t)$ as the optimal weight function to express the contribution of $X_i(t)$ towards $g(\mu_i)$.

In the next section, we propose an extension of the Cox proportional hazards model that incorporates a similar integral term for functional covariates, and describe how the parameters in such a model may be estimated. Section 3 describes a number of extensions to the model that allow for added flexibility. Section 4 assesses the performance of our model in a simulation study. In Section 5, we apply our model to our application of interest, and we conclude with a discussion of our findings in Section 6. Additional details regarding our software implementation using R's pcox package may be found in the supplemental material.

## 2 Cox Model with Functional Covariates

### 2.1 Proportional Hazards Model

Let $T_i$ be the survival time for subject $i$, and $C_i$ the corresponding censoring time. Assume that we observe only $Y_i = \min(T_i, C_i)$, and let $\delta_i = I(T_i \leq C_i)$. We also assume that for each subject, we have a collection of covariates $\boldsymbol{Z}_i = \{Z_{i1}, Z_{i1}, \ldots, Z_{ip}\}$. The Cox proportional hazards model (Cox, 1972) for this data is given by $\log h_i(t; \gamma) = \log h_0(t) + \boldsymbol{Z}_i\gamma$, where $h_i(t; \gamma)$ is the hazard at time t given covariates $\boldsymbol{Z}_i$ and $h_0(t)$ is a non-parametric baseline hazard function. The parameter vector $\exp(\gamma)$ is commonly referred to as the vector of hazard ratios, because it represents the multiplicative change in the hazard function for a one-unit increase in $\boldsymbol{Z}_i$.

Suppose now that in addition to $\boldsymbol{Z}_i$, we have also collected a functional covariate, $X_i(s) \in \mathscr{L}^2[0,1]$, for each subject. Without loss of generality, we assume that $X_i(s)$ is centered by subtracting an estimator of the population mean function from the observed data. We propose the following functional proportional hazards model

$$log h_i[t;\gamma,\beta(\cdot)] = log h_0(t) + \boldsymbol{Z}_i\gamma + \int_0^1 X_i(s)\beta(s)\,ds \quad (2.1)$$

The functional parameter, β(s), is slightly more difficult to interpret than its nonfunctional counterpart, γ. One interpretation is that the term $\exp\left\{\int_0^1 \beta(s)\,ds\right\}$ corresponds to the multiplicative increase in one's hazard of death if the entire covariate function, $X_i(s)$, was shifted upwards by 1 unit (with $\boldsymbol{Z}_i$ held constant). In this context, one can refer to β(s) as a functional log hazard ratio. More generally, β(s) serves as a weight function for $X_i(s)$ to obtain its overall contribution towards one's hazard of mortality. The model assumes proportional hazards; i.e., this contribution is the same over the entire domain *t* of the hazard function.

We note in particular the the domain of the functional predictor, *s*, is not the same as the time domain *t* over which the event is followed. We assume that $X_i(s)$ is fully available at baseline, before the event can occur. If the functional predictor is measured concurrently with the event, then this predictor is a time-varying covariate. Alternative methods exist for this scenario, ranging from traditional approaches (Cox, 1972) to more modern developments in joint modeling of longitudinal and survival data (Tsiatis and Davidian, 2004; Ibrahim et al., 2010; Rizopoulos, 2012).

As proposed above, (2.1) is under-determined unless we make assumptions on the form of the functional coefficient β(s). We will take the common approach of approximating β(s) using a spline basis (Marx and Eilers, 1999; Cardot and Sarda, 2005; Ramsay and Silverman, 2005; Goldsmith et al., 2011). Let $\varphi(s) = \{\varphi_1(s), \varphi_2(s), \ldots, \varphi_{k_b}(s)\}$ be a spline basis over the s-domain, so that $\beta(s) = \sum_{k=1}^{k_b} b_k \phi_k(s)$. Then (2.1) becomes

$$\begin{aligned} log\, h_i(t;\gamma,\boldsymbol{b}) &= log\, h_0(t) + \boldsymbol{Z}_i\gamma \int_0^1 X_i(s)\phi(s)\,\boldsymbol{b}\,ds \\ &= log\, h_0(t) + \boldsymbol{Z}_i\gamma + c_i' b \end{aligned} \quad (2.2)$$

where $\boldsymbol{b} = \{b_1, b_2, \ldots, b_{K_b}\}$ and $\boldsymbol{c}_i$ is a vector of length $K_b$ with $k^{th}$ element $\int_0^1 X_i(s)\phi_k(s)\,ds$. Note that this integral is based only on the covariate function and the (known) basis functions, and may be calculated using numerical integration.

As a choice of basis, we prefer penalized B-splines (Eilers and Marx, 1996), also known as P-splines. B-splines adapt flexibly to the data, have no boundary effects, and are fast to compute. In addition, by using a large number of knots and applying a roughness penalty, we prevent overfitting while eliminating the necessity to choose the number and precise location of the knots (O'sullivan et al., 1986; O'Sullivan, 1988). We have fit the model using other penalized bases, with minimal change to the estimated coefficient function.

### 2.2 Estimation via Penalized Partial Likelihood

For notational convenience, let $\theta = [\gamma \; \boldsymbol{b}]$ and $\eta_i(\boldsymbol{\theta}) = \boldsymbol{Z}_i\gamma + c_i'\boldsymbol{b}$. Then the partial likelihood for this model is $L^{(p)}(\theta) = \Pi_{i:\delta_i=1}[\exp\{\eta_i(\theta)\}/\Sigma_{j:Y_j \geq Y_i}\exp\{\eta_j(\theta)\}]$. In order to ensure the smoothness of the coefficient function $\beta(t)$, we will impose a penalty on the spline coefficients, $\boldsymbol{b}$. Based on the above partial likelihood, we define the penalized partial log-likelihood (PPL) as

$$\ell_\lambda^{(p)}(\theta) = \sum_{i:\delta_i=1}\left[\eta_i(\theta) - log\left(\sum_{j:Y_j \geq Y_i} exp\{\eta_j(\theta)\}\right)\right] - \lambda P(\boldsymbol{b})$$

where $P(\boldsymbol{b})$ is an appropriate penalty term for the spline coefficients $\boldsymbol{b}$, and $\lambda$ is parameter that controls the smoothness of the resulting coefficient function.

The use of a penalized partial likelihood function in Cox models is not new. Gray (1992) introduced the function to allow for smooth effects of scalar covariates in a Cox model, and this method was extended and implemented in R by Therneau and Grambsch (1998). Verweij and Houwelingen (1994) and Therneau et al. (2003) exploited the well-known connection between mixed effects models and penalized splines to estimate frailty models via the penalized partial likelihood. Here, we follow a similar procedure to incorporate functional predictors into the Cox model.

We will assume that we can express the penalty term as $P(\boldsymbol{b}) = \frac{1}{2}\boldsymbol{\lambda}\theta'\boldsymbol{D}\boldsymbol{\theta}$, where $\boldsymbol{D}$ is a symmetric, non-negative definite penalty matrix. Also, let $\boldsymbol{W}_i = \begin{bmatrix}\boldsymbol{Z}\boldsymbol{i} & c_i'\end{bmatrix}$ be the row of the design matrix corresponding to subject $i$, such that $\eta_i(\boldsymbol{\theta}) = \boldsymbol{W}_i'\boldsymbol{\theta}$. Then for a given $\lambda$, the first and second derivatives (i.e., the gradient and hessian matrix) of $\ell_\lambda^{(p)}(\boldsymbol{\theta})$ may be easily computed. We can estimate the regression coefficients $\theta$ by maximizing the partial log-likelihood (for a given $\lambda$) using a Newton-Raphson procedure. In the results presented in Sections 4 and 5, we use a second-order difference penalty. This penalty is a discrete approximation to the integrated second derivative penalty, is computationally efficient, and is commonly used in P-splines (Eilers and Marx, 1996).

### 2.3 Optimization of the smoothing parameter

An essential step in the algorithm is the optimization of the smoothing parameter, $\lambda$. Unfortunately, typical optimization criteria, such as Allen's PRESS (cross-validated residual sum of squares) and Mallow's $C_p$, are not appropriate for Cox models. Verweij and Van Houwelingen (1993) proposed the cross-validated log likelihood (CVL) for the purpose of optimizing the smoothing parameter of a penalized partial likelihood. Let $\hat{\boldsymbol{\theta}}_{(-i)}^{\lambda}$ be the value of $\theta$ that maximizes $\ell_{\lambda,(-i)}^{(p)}(\boldsymbol{\theta})$, the penalized partial log-likelihood when observation $i$ is left out. Then the CVL for a given value of $\lambda$ is given by $\mathrm{CVL}\{\lambda\} = \sum_{i=1}^{n}\ell_{\lambda,i}^{(p)}\left\{\hat{\boldsymbol{\theta}}_{(-i)}^{\lambda}\right\}$, where

$\ell_{\lambda,i}^{(p)}(\cdot) = \ell_{\lambda}^{(p)}(\cdot) - \ell_{\lambda,(-i)}^{(p)}(\cdot)$ is the contribution of subject *i* to the penalized partial log-likelihood.

This expression is quite computationally intensive, as it requires calculating $\ell_{\lambda,i}^{(p)}\left(\hat{\boldsymbol{\theta}}_{(-1)}^{\lambda}\right)$ for each *i*. In practice, a computationally efficient alternative is the penalized version of the AIC, $\mathrm{AIC}(\lambda) = \ell_{\lambda,i}^{(p)}\left(\hat{\boldsymbol{\theta}}^{\lambda}\right) - e(\lambda)$, where $e(\lambda) = tr\left[H_{\lambda}\left(\hat{\boldsymbol{\theta}}^{\lambda}\right)^{-1}\boldsymbol{H}\left(\hat{\boldsymbol{\theta}}^{\lambda}\right)\right]$ is the effective dimension and $\boldsymbol{H}(\cdot)$ is the unpenalized portion of the Hessian matrix. Although the AIC does not approximate the CVL directly, it is known that the changes in AIC or CVL from the null model are approximately equal, making the AIC a useful surrogate (Verweij and Houwelingen, 1994). pcox makes the AIC available, as well as two related criteria: the corrected AIC ($\mathrm{AIC}_c$) of Hurvich et al. (1998), and the EPIC criterion of Shinohara et al. (2011). The $\mathrm{AIC}_c$ criterion has been recommended in cases with small *n* or large number of parameters to avoid overfitting (Burnham and Anderson, 2002), while the EPIC criterion corresponds to a likelihood ratio test at the $\alpha = 0.05$ level for testing the significance of one additional parameter in two nested models. Practically, the three optimization methods offer three different levels of smoothness for the functional coefficient, with AIC resulting in the least smooth and EPIC the smoothest estimate.

## 2.4 Inference

There have been at least two proposals for the variance-covariance estimate of the parameter estimates in penalized partial likelihood models. Gray (1992) suggested using $\boldsymbol{V}_1 = \boldsymbol{H}_{\lambda}\left(\hat{\boldsymbol{\theta}}^{\lambda}\right)^{-1}\mathscr{I}\left(\hat{\boldsymbol{\theta}}^{\lambda}\right)\boldsymbol{H}_{\lambda}\left(\hat{\boldsymbol{\theta}}^{\lambda}\right)^{-1}$. On the other hand, Verweij and Houwelingen (1994) use $\boldsymbol{V}_2 = \boldsymbol{H}_{\lambda}\left(\hat{\boldsymbol{\theta}}^{\lambda}\right)^{-1}$, and refer to the square root of the diagonal elements of this matrix as "pseudo-standard errors". Therneau et al. (2003) make both of these estimates available in their implementation, and we choose to take the same approach. For either choice of $\boldsymbol{V}$, a pointwise 95% confidence interval for $\beta(s_0) = \varphi(s_0)\boldsymbol{b}$ may be constructed as $\hat{\beta}(s_0) \pm 1.96\sqrt{\phi(s_0)\boldsymbol{V}_{22}\phi(s_0)}$, where $\boldsymbol{V}_{22}$ is the lower-right $K_b \times K_b$ matrix of $\boldsymbol{V}$.

Nonetheless, both proposals above for the variance-covariance matrix $\boldsymbol{V}$ are only valid when the smoothing parameter $\lambda$ is fixed. When $\lambda$ is optimized using one of the methods discussed in Section 2.3, these proposals may underestimate the true standard error. Thus, in addition to these two methods we propose to use a bootstrap of subjects (Crainiceanu et al., 2012a) to calculate the pointwise and joint confidence intervals for the model parameters. The performance of these four types of confidence intervals will be compared via simulation in Section 4.

# 3 Extensions

## 3.1 Additional penalized model covariates

An advantage of using penalized splines is that they are modular (Ruppert et al., 2003), making it very easy to extend Model (2.1) to include additional penalized terms. This makes the inclusion of additional functional predictors, smooth effects of scalar covariates, or frailty terms straightforward. Time-varying coeffcients for scalar covariates, $X_i\beta(t)$, may also be included by applying a penalized spline basis to the coeffcient function, allowing for non-proportional hazards (Zucker and Karr, 1990; Grambsch and Therneau, 1994). Each of these terms requires a corresponding penalty term in the penalized partial likelihood, with a separate smoothing parameter for each one. The pcox software package allows for each of these specialized terms in the model formula.

## 3.2 Full Likelihood Approach

An alternative estimation procedure may be employed by assuming a spline basis for the baseline hazard, and maximizing the penalized full data log likelihood (PFL). Such an approach was originally proposed by Cai et al. (2002) using a linear spline basis for the baseline hazard without any penalized covariates, and has been further extended to include penalized predictors including smooth covariate effects, spatial effects, and frailties by Kneib and Fahrmeir (2007); Strasak et al. (2009).

Since our model treats functional predictors as penalized regression terms, the PFL approach may be used to fit (2.2). Let $\psi(t) = \{\psi_1(t), \psi_2(t), \ldots, \psi_{K0}(t)\}$ be a spline basis over the time domain $t$, such that $log\, h_0(t) = g_0(t) = \sum_{k=1}^{K_0} b_{0k}\psi_k(t)$ is a spline approximation of the log baseline hazard, with spline coefficient $\boldsymbol{b}_0 = \{b_{01}, b_{02}, \ldots, b_{0k0}\}$.Then the PFL is

$$\ell_\lambda^{(f)}(\theta) = \sum_{i=1}^{n}\left\{\delta_i\eta_i(T_i,\theta) - \int_0^{T_i} e^{\eta_i(u,\theta)}du\right\} - \frac{\lambda_0}{2}\boldsymbol{b}_0'\boldsymbol{D}_0\boldsymbol{b}_0 - \frac{\lambda_1}{2}\boldsymbol{b}'\boldsymbol{D}\boldsymbol{b}$$

where $\eta_i(t,\theta) = \boldsymbol{Z}_i\gamma + \boldsymbol{b}_0\psi(t) + c_i'\boldsymbol{b}$ and $\lambda_0$ and $\boldsymbol{D}_0$ are the smoothing parameter and penalty matrix for the baseline hazard, respectively. The full likelihood approach has been shown to have advantages over the partial likelihood approach in cases where data is interval-censored (Cai and Betensky, 2003). However, in our application we do not have interval-censored data and we expect minimal benefit to using the PFL approach over the PPL approach. Due to the increased computational demand of using the PFL approach, we have chosen to use the PPL approach throughout this paper. The PFL approach will be offered in our software package.

## 3.3 Missing and unequally-spaced data

A common complication in functional regression occurs when the observed functional predictors are observed at widely spaced, unequal time intervals. This could occur for example when the functional predictor is measured at follow-up times that are not the same for each subject, or when there is substantial missingness in these observations. Goldsmith et al. (2011) showed how a functional principal components (FPCA) basis could be used to

pre-smooth the observed data in a functional regression context, with minimal loss of information. We follow this approach in our modeling strategy. We perform FPCA by smoothing the empirical covariance matrix, as described in Staniswalis and Lee (1998) and Yao et al. (2003), and implemented using the fpca.sc() function in the R package refund (Crainiceanu et al., 2012b).

## 4 Simulation Study

### 4.1 Simulation design

In order to assess the performance of our model under a variety of conditions, we conducted an extensive simulation study. Of interest was our model's ability to accurately identify the coefficient function $\beta(s)$. For simplicity, we consider only the scenario where there are no non-functional covariates $\boldsymbol{Z}$, and only a single functional predictor $X_i(s)$, of fixed domain. Let $\{s_j = j/100 : j = 0, 1, \ldots, J = 100\}$ be our grid of time points over the interval [0, 1]. For each subject $i \in 0, \ldots, N$, we generate the survival time $T_i$ and functional predictor $X_i(s)$ based on the model $h_i(t) = h_0(t) \exp(\eta_i)$, where $\eta_i = \frac{1}{J}\sum_{j=1}^{J} X_i(s_j)\beta(s_j)$ and $X_i(s_j) = u_{i1} + u_{i2}s_j + \sum_{k=1}^{10}\{v_{ik1} \sin(2\pi k s_j) + v_{ik2}\cos(2\pi k 10 s_j)\}$. Here, $h_i(t)$ is the hazard of $T$ for subject $i$, $h_0(t)$ is the baseline hazard, $u_{i1} \sim N(0, 25)$, $u_{i2} \sim N(0, 4)$, and $v_{ik1}$, $v_{ik2} \sim N(0, 1/k^2)$. This model for simulating our functional predictors is based on the procedure employed by Goldsmith et al. (2011), which was in turn adapted from Müller and Stadtmüller (2005). We generated random survival times according to this proportional hazards model by following Bender et al. (2005). The baseline hazard was chosen to follow a Weibull distribution with shape parameter 0.75 and mean 600, where time is assumed measured in days. All subjects are censored at $C_i = 730$ days. These values were chosen to approximate the data that was used in our application. Based on the baseline hazard and censoring mechanism, we expect approximately 27% of subjects to be censored.

We apply three data-generating coefficient functions: $\beta_1(s) = 2\sin\left(\frac{\pi s}{5}\right)$, $\beta_2(s) = 2(s/10)^2$, and $\beta_3(s) = -2\varphi(s|2, 0.3) + 6\varphi(s|5, 0.4) + 2\varphi(s|7.5, 0.5)$, where $\varphi(\cdot|\mu, \sigma)$ is the density of a normal distribution with mean $\mu$ and standard variance $\sigma^2$. The three coefficient functions appear in the top row of Figure 1.

We consider three different sample sizes, $N \in \{100, 200, 500\}$ subjects. We are also interested in the amount of information that is lost when there is a large amount of missing data in the covariate measurements. In order to address this issue, we generate an "incomplete" version of the full covariate dataset as follows. First, for each subject we randomly select the number of measured values $J_i$ as a random integer between 10 and 51, with equal probability. We then randomly select $J_i$ of the $J = 100$ observations that are to be included in the incomplete dataset, again with equal probability. The result constitutes the observed, incomplete dataset. The incomplete coefficient functions are then smoothed using a functional principal components basis that retains enough principal components to explain 99% of the variability in the covariate functions.

For each combination of sample size, true coefficient function, and level of missingess (full vs. incomplete dataset), we generate $R = 1000$ simulated datasets and apply our model for Cox regression with a single functional predictor. The three versions each use a different criterion for selecting the smoothing parameter: AIC, $AIC_C$, or EPIC. In all cases, we use penalized B-splines to model the coefficient function (Marx and Eilers, 1999), using the difference penalty of Eilers and Marx (1996).

### 4.2 Evaluation criteria

The primary measure of model performance is its ability to estimate the true coefficient function. This is assessed by the average mean squared error (AMSE), defined as AMSE $\left(\hat{\beta}(s)\right) = \frac{1}{J+1}\sum_{j=0}^{J}\left\{\hat{\beta}(s_j) - \beta(s_j)\right\}^2$, where $\hat{\beta}(s_j)$ is the estimated coefficient function at $s = s_j$ and $\beta(s_j)$ is the value of the true coefficient function at this location. A secondary measure of model performance is its predictive ability, which we measure with the cross-validated concordance probability, or C-Index (Harrell et al., 1996; van Houwelingen and Putter, 2012). The C-Index is the proportion of all pairs of observations for which the order of survival times are concordant with the model-based linear predictor; we use a 10-fold cross validated version of this statistic.

We also evaluated the coverage probability of four different pointwise 95% confidence intervals. The first two confidence intervals are formed using the two model-based estimates of the variance, as described in Section 2.4. The second two are bootstrap estimates based on 100 bootstrapped samples. One of these is a Wald-type confidence interval based on the bootstrap estimate of the variance, and the other is based on 2.5% and 97.5% quantiles of the bootstrap distribution of the estimates.

### 4.3 Simulation results

Box plots of the AMSE and cross-validated C-Index appear in Figure 1, along with the coefficient function estimates that had median AMSE for the scenario when $N = 200$. Overall, there is very little difference in the performance between the three optimization criteria. The EPIC criterion tends to have slightly better performance under $\beta_2(s)$, as this coefficient function is very smooth and EPIC favors smoother estimates, but these gains are minimal.

$\beta_3(s)$, with its sharp peaks and valleys, is by far the most difficult coefficient function to estimate, especially in the incomplete data case. Interestingly, despite its estimation resulting in the highest AMSE measurements, the models under this coefficient perform fairly well in terms of predictive ability, with median cross-validated C-indices near 90%. This observation reflects that this coefficient function's peaks, while difficult to estimate precisely, are relatively high in magnitude, which causes the model to “target” certain predictor functions as being more strongly or weakly associated with mortality. On the other hand, $\beta_1(s)$ is relatively easy to estimate, but is much less strongly associated with survival. This result is due to the fact that the sine wave contains an equal amount of area that is positive and negative, causing subjects with relatively flat $X_i(s)$ to not be clearly identified as either “high-risk” or “low-risk” for mortality. Only subjects who display a clear increasing

or decreasing trajectory (high to low or low to high) in their predictor functions will be easily separable.

As expected, estimation of the coefficient functions is more difficult with low sample size and in the incomplete case, resulting in higher AMSE measurements. However, the model still seems to be useful in these cases. Interestingly, lower sample sizes and incomplete data do not seem to cause a very large drop in the C-Index measurements, indicating that these more challenging scenarios do not cause the model to lose much in terms of predictive ability. In order to assess whether differences in AMSE between the incomplete and full data scenarios were due the missing data in the incomplete case or the FPCA step that this data requires, we fit the same models to a version of the full data that was pre-smoothed by FPCA. The results (supplemental material) show very similar performance to those corresponding to the unsmoothed full dataset, indicating that it is the lack of information due to the missing data, and not the FPCA step, that causes the decreased model performance.

Coverage probabilities of the confidence intervals are shown in Table 1. The two model-based confidence intervals (V1 and V2) perform well in moderate-to-large sample size ($N \geq 200$) with no missing data, for $\beta_1(s)$ and $\beta_2(s)$. However, when $N$ is small or missing data is present, both tend to underestimate the variance of the estimates. This underestimation can be quite severe, with coverages as low as 25% in the most difficult scenarios. V1 tends to be more conservative than V2. The two bootstrap-based confidence intervals maintain coverage probabilities above 95%, except under $\beta_3(s)$. However, in scenarios that are easier to estimate (full dataset, large $N$, and $\beta_2(s)$) these coverages seem to be overly conservative, especially B-V. Under $\beta_3(s)$, which is the most difficult coefficient function to estimate, the bootstrap-based confidence intervals perform well when there is no missing data, but perform much more poorly on the incomplete datasets. Interestingly, this problem is exacerbated in larger sample sizes. We suspect that coverage may be improved by taking more than 100 bootstrap samples, which is what was chosen for these simulations. In all scenarios, coverage probabilities were highest when using the AIC criterion to optimize smoothness, and lowest with EPIC. These results are emphasized when examining plots of the pointwise coverage probability as a function of the domain $s$ (supplemental material).

## 5 Effect of SOFA Score on Post-ICU Mortality

### 5.1 Data description

The Improving Care of Acute Lung Injury Patients (ICAP) study (Needham et al., 2006) is a prospective cohort study that investigates the long-term outcomes of patients who suffer from acute lung injury/acute respiratory distress syndrome (ALI/ARDS). ALI/ARDS is a severe condition characterized by inflammation of the lung tissue. It can be triggered by a wide variety of causes, including pneumonia, sepsis, or trauma. Treatment consists of supported care, including mechanical ventilation in the intensive care unit (ICU), until the patient's condition stabilizes. Short-term mortality from ARDS can exceed 40% (Zambon and Vincent, 2008), and those that do survive are at increased risk for physical, cognitive, and emotional impairments, as well as death.

The ICAP study enrolled 520 subjects, with 237 (46%) dying in the ICU. We are concerned with long-term survival among the 283 survivors, once they are discharged from the hospital. Out of the 283 survivors, 16 subjects (5.7%) did not consent to follow-up, their mortality status was unknown after hospital discharge, and they were excluded from the analysis. Thus, our analysis is based on the remaining 267 subjects. All patients in the ICAP study who are discharged alive from the hospital and consented to follow-up were followed for up to two years from their date of enrollment in the study. If the subject died, mortality information was recorded to the nearest day, based on family report or publicly available records.

In the ICAP study, data recording starting upon enrollment in the study, and then daily thereafter during the patient's ICU stay. One measurement recorded daily during each subject's ICU stay is the Sequential Organ Failure Assessment (SOFA) score, which is a composite score of a patient's overall organ function status in the ICU. It consists of six physiological components (respiratory, cardiovascular, coagulation, liver, renal, and neurological), each measured on a scale of 0–4, with higher scores indicative of poorer organ function. The total SOFA score is the sum of these six subscores, ranging from 0–24. We consider each subject's history of measured SOFA scores, over time, to be a functional covariate, $X_i(u)$, where $u$ is the ICU day. These functions are depicted in the left panel of Figure 2 as a lasagna plot (Swihart et al., 2010).

### 5.2 Analysis plan

Our goal will be to estimate the association between post-hospital mortality and a patient's SOFA function. In addition to the SOFA function, our model also includes three non-functional covariates, which are meant to control for a subject's baseline risk of post-hospital mortality: age, gender, and Charlson co-morbidity index (Charlson et al., 1987). We consider “time zero”, the first day the subject is eligible for our event of interest, to be the day the subject is discharged from the hospital following ALI/ARDS, and subjects are censored at two years following their ALI diagnosis.

There are two features of the data that raise compilations in our analysis. The first complication occurs because some subjects are discharged from the ICU (due to an improvement in their condition) to a hospital ward, only to be readmitted to the ICU later during their hospitalization. Since SOFA measurements are only recorded in the ICU, these subjects will have gaps in their SOFA functions (indicated by gray space in Figure 2). Of the 267 subjects, 20 (7.5%) had gaps of this type, for a total of 4.4% of missing patient days. Based on clinical advice, we complete these gaps using a last observation carried forward (LOCF) imputation, though we test the sensitivity and sensibility of this approach by using alternative imputation approaches and by comparing to the complete case only analysis. As differences were found to be minimal, we present results for the LOCF imputation only.

The second, more significant complication is that each subject remains in the ICU for a different length of time, $U_i$. Since we use time as the domain of our functional covariate, this means that each function, $X_i(u)$, will be measured over a different domain, $[0, U_i]$. The distribution of $U_i$ up to 35 days can be seen in Figure 2, and overall ranges from as short as 2 days to as long as 157 days. Gellar et al. (2014) have proposed methods for accounting for

this type of data in the context of classical functional regression using domain-dependent functional parameters. These approaches could be incorporated here, but we focus on a more traditional approach here to keep presentation simple. More precisely, we apply the subject-specific domain transformation $s := g_i(u) = u/U_i$ to each function. This allows us to define new SOFA functions, $\tilde{X}_i(s) = X_i(sU_i)$, that are each defined over [0, 1]. The new SOFA functions defined over the $s$ domain are a linearly compressed version of the original SOFA functions (Figure 2, right panel), and $s$, has the interpretation of being the proportion of the way through one's ICU stay that the measurement was taken. For example, $\tilde{X}_i(0.5)$ represents a subject's SOFA score half way through his ICU stay, and $\tilde{X}_i(1)$ is the SOFA score on the subject's last day in the ICU. We evaluate $\tilde{X}_i(s)$ at $J = \max_i(U_i) = 157$ time points for each function.

Standardizing each subject's SOFA curve to a common domain may cause us to lose some potentially valuable information; specifically, we lose information regarding the original domain width, $U_i$. However, we note (Figure 3) that the average domain-standardized SOFA trajectory tends to be markably consistent across different strata of $U_i$. In particular, we do not observe any strong patterns in the functions across these strata, and $U_i$ does not appear to affect the difference in curves between those who do and do not experience the event. These observations support our decision to standardize the SOFA functions to a common domain. In addition, since $U_i$ itself could be a strong predictor of long-term mortality, we incorporate it into our model using a smooth effect on the log scale. The resulting model is

$$log\, h_i(t;\gamma,\beta(\cdot)) = log\, h_0(t) + \boldsymbol{Z}_i\gamma + \int_0^1 \tilde{X}_i(s)\,\beta(s)\,ds + f\{log(U_i)\} \quad (5.1)$$

where the the scalar covariates $\boldsymbol{Z}_i$ include subject age, gender, and Charlson Comorbidity Index. The Charlson Index (Charlson et al., 1987) is a commonly used measure of baseline health, with each existing clinical conditions assigned a score from 1–6 based on severity. In ICAP, total Charlson scores range from 0 to 15. A P-spline basis is used to approximate both the functional coefficient $\beta(\cdot)$ and the additive term $f(\cdot)$.

### 5.3 Results

We plot the estimated additive function $\hat{f}\{log(U_i)\}$ and coefficient function $\hat{\beta}(s)$ based on (5.1) in Figure 4, under each of the three optimization criteria. Overall, we see very little functional association between one's SOFA function and time to death after hospital discharge, as the 95% confidence interval covers the horizontal line $\beta(s) = 0$ throughout the entire domain $s$. This implies that the integral term of the model, $\int_0^1 \tilde{X}_i(s)\,\beta(s)\,du$, will be close to zero for all subjects, so $\tilde{X}_i(s)$ offers little contribution towards one's hazard of death. There seems to be a positive association with length of stay for lengths of stay less than 5 days, but not for those greater than or equal to 5. On the other hand, two of the non-functional covariates, age and Charlson co-morbidity index, are highly associated with an increased hazard of death. According to the model optimized using $AIC_c$ and the confidence interval calculated from the quantiles of the bootstrap distribution, we found that one's hazard of mortality increases by 5% (95% CI 4%–9%, $p < 0.001$) for every 1-year increase

in age, and it increases by 12% (95% CI 6% – 36%, $p = 0.001$) for every 1-point increase in Charlson Index.

We see that the EPIC criterion imposes a stronger degree of smoothness in both $\hat{f}(\cdot)$ and $\hat{\beta}(\cdot)$ than the other two smoothing criteria, to the extent that very little functional signal can be detected for either estimate. The widely-varying bootstrap confidence intervals, especially for $\beta(s)$, is likely due to a flat likelihood, without much information contributed by $X_i(s)$. By imposing a higher degree of smoothness, the EPIC criterion results in more reasonable intervals.

In order to compare the three sets of estimates, we calculate the *N*-fold cross-validated C-Index for each one (Table 2, top 3 rows), as a measure of how well each model would predict mortality when applied to an independent dataset. The model fit with EPIC had the highest predictive ability. This observation reinforces the conclusion that there is not a very strong functional relationship between SOFA score in the ICU and long-term mortality among patients surviving their hospitalization, as EPIC favors estimates closer to the zero line.

### 5.4 Follow-up Models

While often functional regression models are viewed as the final part of the inference, here we use them as exploratory tools. Indeed, a more careful examination of Figure 4 reveals some potentially important patterns; in this section we investigate those patterns further.

First, we note that the plots for $\hat{f}\{log(LOS)\}$ obtained using the AIC and $AIC_c$ criteria indicate that the function may be appropriately modeled by a linear spline with one knot at 5 days. There is a visually striking effect after 100 days, as well, but as only three survivors had a length of stay longer than 100 days, we have decided to ignore it. Additionally, we note a number of regions over the *s*-domain of $\hat{\beta}(s)$ with relatively large magnitude, even if the point-wise confidence intervals cover 0. These regions occur around $s = 0, 0.15, 0.3$, and 1.

Using these observations as a guide, we fit a series of follow-up parametric models, and investigate their predictive ability via the *N*-fold cross-validated C-Index (Table 2). These models differed in their treatment of the functional predictor and the $U_i$ variable. We also investigated whether or not incorporating the age covariate as a smooth term improved model fit, but in each case it did not so we only present results that treat age linearly. Interestingly, we have found that parameterizing the additive and functional effects resulted in superior predictive performance. In particular, we observed that removing the SOFA scores from the model completely resulted in a higher cross-validated C-Index than including them as functional effects; this seems to indicate that there is not a strong functional relationship between SOFA score and long-term mortality in this subset of patients.

The best-performing model included the scalar covariates $Z_i$, a linear spline for $\log(U_i)$, and the mean SOFA score over the region $s \in [0.85, 1]$ (M4 in Table 2). The mean SOFA over $s \in [0, 0.05]$ (M1) also shows a potentially important association. Investigating the effects in

these regions further, Figure 5 depicts the Kaplan-Meier estimates of the survival curve, stratified by high vs. low values of M1 and M4. Based on these findings, we think that more exploration of functional approaches followed by aggressive thresholding of functional parameter estimates may actually lead to improved prediction and interpretation. This is likely to be the case in applications with a small to moderate number of subjects and weak functional effects.

## 6 Discussion

We develop new methodology to account for functional covariates in a Cox proportional hazards model. We use a spline basis to approximate the functional coefficient, and estimation is accomplished by maximizing the penalized partial likelihood, with the degree of smoothness determined by optimizing one of three presented information criteria. The model is flexible and modular, in that it can be easily extended to incorporate a number of advances both in the fields of survival analysis and in functional data analysis. We have developed easy to use and computationally efficient software to implement this model.

We demonstrate through simulations that this model does a good job of estimating the true coefficient function even in cases with a moderate amount of missing data, except when the true coefficient function is especially complicated. Model-based estimates of the standard error resulted in pointwise confidence intervals that tended to be too narrow, and we therefore recommend bootstrap-based confidence intervals as a more conservative alternative.

We applied this model to estimate the functional association between SOFA score and post-hospital mortality among patients with ARDS. We found that this association is quite close to zero, throughout the domain of our functional predictor. Despite the null result, this observation is potentially quite clinically meaningful. It tells us that, among this patient population, one's survival after leaving the hospital does not appear to depend heavily on patterns of organ failure in the ICU, after accounting for age, gender, comorbidity status, and ICU length of stay. It was previously hypothesized by our collaborators that we may observe different patterns in mortality based on one's SOFA pattern. This analysis does not support that hypothesis.

Another possibility is that this model is mis-specified. Recall that we modified our original SOFA functions $X(u)$ by collapsing them to $\tilde{X}(s)$. While this conveniently allowed us to avoid the problem of each function falling on a different domain, $[0, U_i]$, it may have caused us to lose important information that was present in the original functions. A more appropriate model would allow for the coefficient function to change with $U_i$, resulting in a variable domain functional regression model with bivariate coefficient function $\beta(u, U_i)$, similar to that proposed by Gellar et al. (2014). Another alternative is to allow for a time-varying effect of the SOFA functions, which would involve replacing the coefficient function $\beta(s)$ in (5.1) with the bivariate coefficient function $\beta(s, t)$. We do not explore these models in this paper, and leave them to future work.

Even though our analysis suggested a very weak functional relationship between SOFA and mortality, we were able to use our functional estimates to guide the design of simpler models that parameterize this association. These simpler models proved to demonstrate better predictive accuracy than the full functional approach, as measured by the cross-validated C-Index. This process shows the strength of functional regression techniques as an exploratory tool for understanding the relationship between a functional predictor and an outcome. In some cases the full functional model may be most appropriate, but more often than not this association may be simplified to a more parsimonious and interpretable relationship.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

We want to thank Professors Terry M. Therneau of the Mayo Clinic, Simon N. Wood of the University of Bath, and Fabian Scheipl of the University of Munich for very useful comments and help with software implementation. This project was supported by NIH grants 2T32ES012871 from the National Institute of Environmental Health Sciences and R01HL123407 from the National Heart, Lung, and Blood Institute of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## References

Belitz, C.; Brezger, A.; Kneib, T.; Lang, S.; Umlauf, N. BayesX: Software for Bayesian Inference in Structured Additive Regression Models. 2013. URL http://www.bayesx.org/

Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. Statistics in medicine. 2005; 24(11):1713–23. ISSN 0277-6715. doi: 10.1002/sim.2059. URL http://www.ncbi.nlm.nih.gov/pubmed/15724232. [PubMed: 15724232]

Buckley J, James I. Linear regression with censored data. Biometrika. 1979; 66(3):429–436.

Burnham, K.; Anderson, D. Model selection and multimodel inference: a practical information-theoretic approach. 2nd edition. Springer Verlag; 2002. URL http://books.google.com/books?hl=en&lr=&id=fT1Iu-h6E-oC&oi=fnd&pg=PR7&dq=Model+Selection+and+Multimodel+Inference:+A+Practical+Information-Theoretic+Approach&ots=tdAob1TDo6&sig=zkkzCorOehBPeCX0xhtuEnGVlfQ

Cai T, Hyndman RJ, Wand MP. Mixed Model-Based Hazard Estimation. Journal of Computational and Graphical Statistics. 2002; 11(4):784–798. ISSN 1061-8600. doi: 10.1198/106186002862. URL http://www.tandfonline.com/doi/abs/10.1198/106186002862.

Cai T, Betensky R. a. Hazard regression for interval-censored data with penalized spline. Biometrics. 2003; 59(3):570–9. ISSN 0006-341X. URL http://www.ncbi.nlm.nih.gov/pubmed/14601758. [PubMed: 14601758]

Cardot H, Ferraty F, Sarda P. Spline Estimators for the Functional Linear Model. Statistica Sinica. 2003; 13:571–591. URL http://www3.stat.sinica.edu.tw/statistica/oldpdf/A13n31.pdf.

Cardot H, Sarda P. Estimation in generalized linear models for functional data via penalized likelihood. Journal of Multivariate Analysis. 2005; 92(1):24–41. ISSN 0047259X. doi: 10.1016/j.jmva.2003.08.008. URL http://linkinghub.elsevier.com/retrieve/pii/S0047259X03001489.

Cardot H, Ferraty F, Sarda P. Functional linear model. Statistics & Probability Letters. 1999; 45(1):11–22. ISSN 01677152. doi: 10.1016/S0167-7152(99)00036-X. URL http://linkinghub.elsevier.com/retrieve/pii/S016771529900036X.

Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal Of Chronic Diseases. 1987; 40(5):373–383. URL http://www.ncbi.nlm.nih.gov/pubmed/3558716. [PubMed: 3558716]

Chiou J, Müller H. Modeling Hazard Rates as Functional Data for the Analysis of Cohort Lifetables and Mortality Forecasting. Journal of the American Statistical Association. 2009; 104:572–585. URL http://amstat.tandfonline.com/doi/full/10.1198/jasa.2009.0023.

Cox D. Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B (Methodological). 1972; 34(2):187–220. URL http://www.jstor.org/stable/10.2307/2985181.

Crainiceanu CM, Staicu A-M, Ray S, Punjabi N. Bootstrap-based inference on the difference in the means of two correlated functional processes. Statistics in medicine. 2012a; 31(26):3223–40. ISSN 1097-0258. doi: 10.1002/sim.5439. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3966027&tool=pmcentrez&rendertype=abstract. [PubMed: 22855258]

Crainiceanu, C.; Reiss, P.; Goldsmith, J.; Huang, L.; Huo, L.; Scheipl, F.; Greven, S.; Harezlak, J.; Kundu, M.; Zhao, Y. R package version 0.1-6. 2012b. refund: Regression with Functional Data.

Eilers PHC, Marx BD. Flexible smoothing with B -splines and penalties. Statistical Science. 1996; 11(2):89–121. ISSN 0883-4237. doi: 10.1214/ss/1038425655. URL http://projecteuclid.org/Dienst/getRecord?id=euclid.ss/1038425655/.

Ferraty, F., editor. Recent Advances in Functional Data Analysis and Related Topics. Springer Verlag; Berlin: 2011.

Ferraty, F.; Vieu, P. Nonparametric Functional Data Analysis: Theory and Practice. Springer Verlag; 2006. URL http://books.google.com/books?hl=en&lr=&id=lMy6WPFZYFcC&oi=fnd&pg=PR8&dq=Nonparametric+Functional+Data+Analysis:+Theory+and+Practice&ots=CFiTl-0bVw&sig=ItS8h-u11R_NZXHfkuNi3g-gHpQ

Gellar, JE.; Colantuoni, E.; Needham, DM.; Crainiceanu, CM. Variable-Domain Functional Regression for Modeling ICU Data (in-press). Journal of the American Statistical Association. 2014. URL http://biostats.bepress.com/jhubiostat/paper261

Goldsmith J, Bobb J, Crainiceanu CM, Caffo B, Reich D. Penalized Functional Regression. Journal of Computational and Graphical Statistics. 2011; 20(4):830–851. doi: 10.1198/jcgs.2010.10007. [PubMed: 22368438]

Grambsch P, Therneau T. Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. Biometrika. 1994; 81(3):515–526. URL http://biomet.oxfordjournals.org/content/81/3/515.short.

Gray R. Flexible Methods for Analyzing Survival Data Using Splines, With Applications to Breast Cancer Prognosis. Journal of the American Statistical Association. 1992; 87(420):942–951. URL http://amstat.tandfonline.com/doi/full/10.1080/01621459.1992.10476248.

Harezlak, J.; Randolph, TW. Structured Penalties for Generalized Functional Linear Models (GFLM). In: Ferraty, F., editor. Recent Advances in Functional Data Analysis and Related Topics, Contributions to Statistics. Vol. chapter 25. Springer Verlag, Heidelberg; 2011. p. 161-167.ISBN 978-3-7908-2735-4. doi: 10.1007/978-3-7908-2736-1. URL http://link.springer.com/10.1007/978-3-7908-2736-1

Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine. 1996; 15:361–387. [PubMed: 8668867]

Hurvich CM, Simonoff JS, Tsai CL. Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion. Journal of the Royal Statistical Society. Series B (Methodological). 1998; 60(2):271–293.

Ibrahim JG, Chu H, Chen LM. Basic concepts and methods for joint models of longitudinal and survival data. Journal of clinical oncology : offcial journal of the American Society of Clinical Oncology. 2010; 28(16):2796–801. ISSN 1527-7755. doi: 10.1200/JCO.2009.25.0654. URL http://www.ncbi.nlm.nih.gov/pubmed/20439643.

James GM. Generalized linear models with functional predictors. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2002; 64(3):411–432. ISSN 13697412. doi: 10.1111/1467-9868.00342. URL http://doi.wiley.com/10.1111/1467-9868.00342.

James GM, Wang J, Zhu J. Functional linear regression thats interpretable. The Annals of Statistics. 2009; 37(5A):2083–2108. ISSN 0090-5364. doi: 10.1214/08-AOS641. URL http://projecteuclid.org/euclid.aos/1247663749.

Kneib T, Fahrmeir L. A Mixed Model Approach for Geoadditive Hazard Regression. Scandinavian Journal of Statistics. 2007; 34(1):207–228. ISSN 0303-6898. doi: 10.1111/j.1467-9469.2006.00524.x. URL http://doi.wiley.com/10.1111/j.1467-9469.2006.00524.x.

Marx B, Eilers P. Generalized Linear Regression on Sampled Signals and Curves: A P-Spline Approach. Technometrics. 1999; 41(1):1–13. URL http://amstat.tandfonline.com/doi/abs/10.1080/00401706.1999.10485591.

McCullagh, P.; Nelder, JA. Generalized linear models. Vol. ume 37. Chapman & Hall/CRC; 1989.

McLean, MW.; Hooker, G.; Staicu, A-M.; Scheipl, F.; Ruppert, D. Functional Generalized Additive Models. Journal of Computational and Graphical Statistics. May. 2012 2013130322113058003. ISSN 1061-8600. doi: 10.1080/10618600.2012.729985. URL http://www.tandfonline.com/doi/abs/10.1080/10618600.2012.729985

Miller R, Halpern J. Regression with censored data. Biometrika. 1982; 69(3):521–531.

Müller H-G, Stadtmüller U. Generalized functional linear models. The Annals of Statistics. 2005; 33(2):774–805. ISSN 0090-5364. doi: 10.1214/009053604000001156. URL http://projecteuclid.org/Dienst/getRecord?id=euclid.aos/1117114336/.

Müller H, Zhang Y. Time-Varying Functional Regression for Predicting Remaining Lifetime Distributions from Longitudinal Trajectories. Biometrics. 2005; 61(4):1064–1075. doi: 10.1111/j.1541-0420.2005.00378.x. Time-Varying. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1541-0420.2005.00378.x/full. [PubMed: 16401280]

Needham DM, Dennison CR, Dowdy DW, Mendez-Tellez P. a. Ciesla N, Desai SV, Sevransky J, Shanholtz C, Scharfstein D, Herridge MS, Pronovost PJ. Study protocol: The Improving Care of Acute Lung Injury Patients (ICAP) study. Critical care (London, England). 2006; 10(1):R9. ISSN 1466-609X. doi: 10.1186/cc3948. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid= 1550857&tool=pmcentrez&rendertype=abstract.

O'sullivan F, Yandell BS, Raynor WJJ. Automatic Smoothing of Regression Functions in Generalized Linear Models. Journal of the American Statistical Association. 1986; 81(393):96–103. URL http: //amstat.tandfonline.com/doi/abs/10.1080/01621459.1986.10478243.

O'Sullivan F. Nonparametric estimation of relative risk using splines and cross-validation. SIAM Journal on Scientific and Statistical Computing. 1988; 9(3):531–542. URL http://epubs.siam.org/doi/pdf/10.1137/0909035.

R Development Core Team. R: A Language and Environment for Statistical Computing. 2011. ISSN 16000706. URL http://www.r-project.org

Ramsay, JO.; Silverman, B. Functional data analysis. Springer; New York: 2005.

Ramsay, JO.; Hooker, G.; Graves, S. Functional data analysis with R and MATLAB. Springer Verlag; 2009.

Reiss PT, Ogden RT. Functional Principal Component Regression and Functional Partial Least Squares. Journal of the American Statistical Association. 2007; 102(479):984–996. ISSN 0162-1459. doi: 10.1198/016214507000000527. URL http://www.tandfonline.com/doi/abs/10.1198/016214507000000527.

Rizopoulos, D. Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. Chapman & Hall/CRC; Boca Raton, FL: 2012. URL http://books.google.com/books?hl=en&lr=&id=xotIpb2duaMC&oi= fnd&pg=PP1&dq=Joint+Models+for+Longitudinal+and+Time-to-Event+Data:+With+Applications+in+R&ots=ODo7KlbUmd&sig=fsARbJp2SJbuUxcWx_GOjRUHrUQ

Ruppert, D.; Wand, MP.; Carroll, RJ. Semiparametric regression. Vol. ume 12. Cambridge University Press; 2003.

Schmee J, Hahn G. A Simple Method for Regression Analysis with Censored Data. Technometrics. 1979; 21(4):417–432. URL http://www.tandfonline.com/doi/full/10.1080/00401706.1979.10489811.

Shinohara, R.; Crainiceanu, C.; Caffo, B.; Reich, D. Longitudinal Analysis of Spatiotemporal Processes: A Case Study of Dynamic Contrast- Enhanced Magnetic Resonance Imaging in Multiple Sclerosis. Technical Report September 2011. 2011. URL http://biostats.bepress.com/jhubiostat/paper231/

Staniswalis J, Lee J. Nonpametric Regression Analysis of Longitudinal Data. Journal of the American Statistical Association. 1998; 93(444):1403–1418.

Strasak A, Lang S, Kneib T, Brant L. Use of Penalized Splines in Extended Cox-Type Additive Hazard Regression to Flexibly Estimate the Effect of Time-varying Serum Uric Acid on Risk of Cancer Incidence: A Prospective, Population- Based Study in 78,850 Men. Annals of Epidemiology. 2009; 19(1):15–24. doi: 10.1016/j. annepidem.2008.08.009.Use. URL http://www.sciencedirect.com/science/article/pii/S1047279708001695. [PubMed: 18835524]

Swihart BJ, Caffo B, James BD, Strand M, Schwartz BS, Punjabi NM. Lasagna plots: a saucy alternative to spaghetti plots. Epidemiology (Cambridge, Mass.). 2010; 21(5):621–5. ISSN 1531-5487. doi: 10.1097/EDE.0b013e3181e5b06a. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2937254&tool= pmcentrez&rendertype=abstract.

Therneau, TM. coxme: Mixed E ects Cox Models. 2012.

Therneau, TM. A Package for Survival Analysis in S. 2014.

Therneau, TM.; Grambsch, PM. Technical report, Division of Biostatistics. Mayo Clinic; Rochester, MN: 1998. Penalized Cox models and Frailty; p. 1-58.

Therneau TM, Grambsch PM, Pankratz VS. Penalized Survival Models and Frailty. Journal of Computational and Graphical Statistics. 2003; 12(1):156–175. ISSN 1061-8600. doi: 10.1198/1061860031365. URL http://www.tandfonline.com/doi/abs/10.1198/1061860031365.

Tsiatis A, Davidian M. Joint Modeling of Longitudinal and Time-to-Event Data: An Overview. Statistica Sinica. 2004; 14(2004):809–834. URL http://www3.stat.sinica.edu.tw/statistica/oldpdf/A14n39.pdf.

van Houwelingen, H.; Putter, H. Dynamic Prediction in Clinical Survival Analysis. Chapman & Hall/CRC; Boca Raton, FL: 2012. URL http://dl.acm.org/citation.cfm?id=2161914

Verweij PJ, Van Houwelingen HC. Cross-validation in survival analysis. Statistics in medicine. 1993; 12(24):2305–14. ISSN 0277-6715. URL http://www.ncbi.nlm.nih.gov/pubmed/8134734. [PubMed: 8134734]

Verweij P, Houwelingen HV. Penalized Likelihood in Cox Regression. Statistics in Medicine. 1994; 13:2427–2436. URL http://onlinelibrary.wiley.com/doi/10.1002/sim.4780132307/abstract. [PubMed: 7701144]

Yao F, Müller H-G, Clifford AJ, Dueker SR, Follett J, Lin Y, Buchholz B. a. Vogel JS. Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. Biometrics. 2003; 59(3):676–85. ISSN 0006-341X. URL http://www.ncbi.nlm.nih.gov/pubmed/14601769. [PubMed: 14601769]

Zambon M, Vincent J-L. Mortality rates for patients with acute lung injury/ARDS have decreased over time. Chest. 2008; 133(5):1120–7. ISSN 0012-3692. doi: 10.1378/chest.07-2134. URL http://www.ncbi.nlm.nih.gov/pubmed/18263687. [PubMed: 18263687]

Zucker D, Karr A. Nonparametric Survival Analysis with Time-Dependent Covariate E ects: A Penalized Partial Likelihood Approach. The Annals of Statistics. 1990; 18(1):329–353. URL http://www.jstor.org/stable/10.2307/2241546.

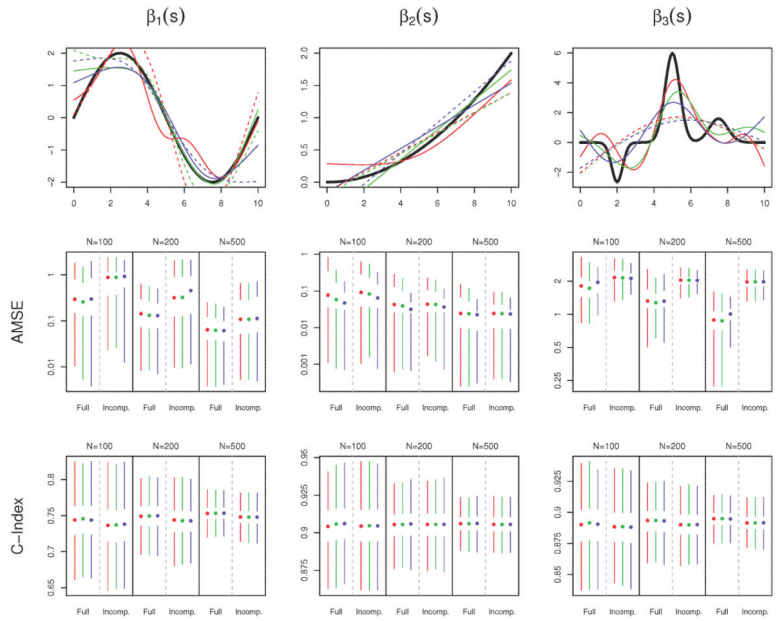


**Figure 1.**

Simulation results. In all plots, color indicates the method for optimizing the smoothing parameter: red (AIC), green ($AIC_C$), or blue (EPIC). The top row displays the true coefficient functions (black), as well as the estimates with median AMSE when $N = 200$. The estimates based on the full data are given by the solid lines, and those based on the incomplete dataset are dashed. The second and third rows contain Tufte box plots of the distributions of AMSE and cross-validated concordance probability (C-Index) respectively, over the 1000 simulated datasets, stratified by sample size and missingness. The median value is indicated by a dot, the interquartile range by white space around the dot, and the smallest and largest non-outlying values by the endpoints of the colored bars. Outliers are defined to be data points more than 1.5 times the interquartile range from the nearest quartile. Lower AMSE and higher C-Index are indicative of better model performance.

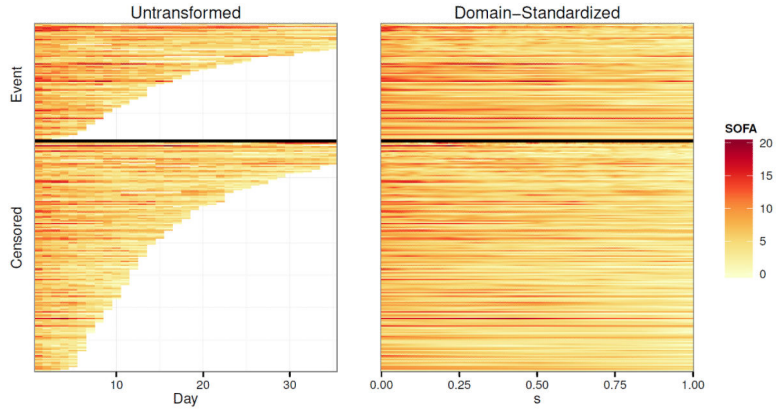


**Figure 2.**

SOFA functions, before and after domain-standardization, depicted as lasagna plots. Each row corresponds to a subject, with color indicating the SOFA score at each time point. Subjects are ordered by domain width of the untransformed functions, $U_i$, within each outcome category (event vs. censored), in both plots.

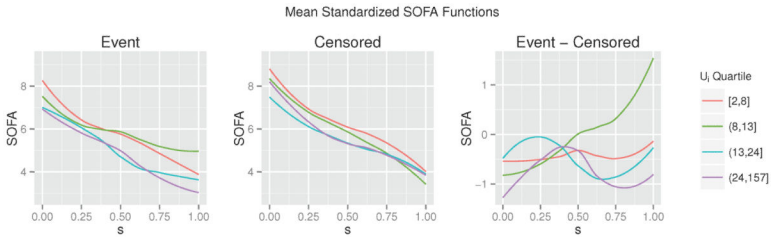


**Figure 3.**

Mean domain-standardized SOFA functions, stratified by $U_i$ and event status, as well as the difference in mean functions between those who did and did not experience the event. All curves are smoothed using a lowess smoother.

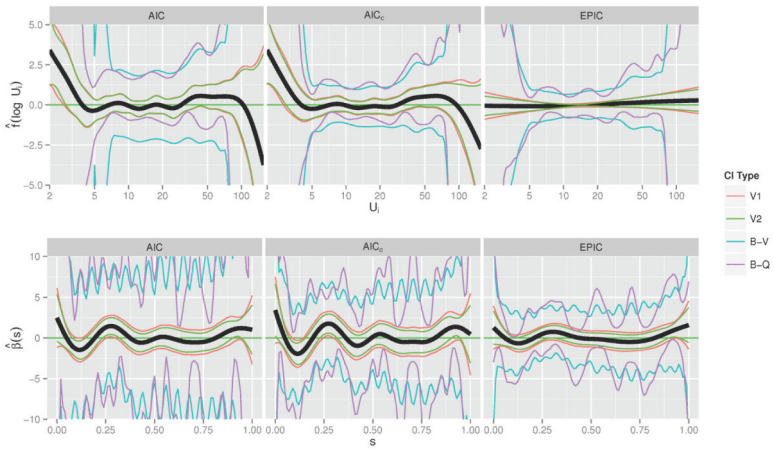


**Figure 4.**

Estimated associations between one's log hazard of mortality and their ICU length of stay (top row of figures) and the standardized SOFA functions (bottom row), under each of the three methods for optimizing λ. Pointwise 95% confidence intervals are formed by one of four methods. The first two (V1 and V2) are from model-based estimates of the variance. The second two are based on 10,000 bootstrapped samples of the dataset, one (B-V) which uses the pointwise bootstrap variance, and the other (B-Q) which uses the pointwise quantiles.

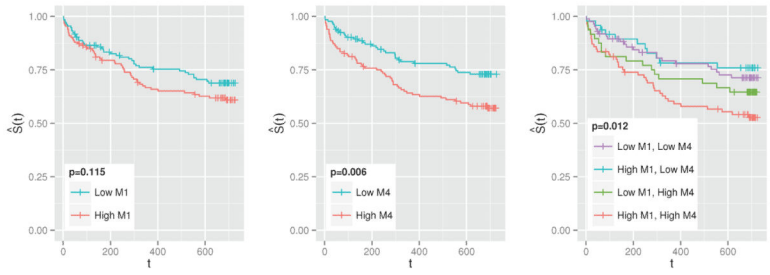


**Figure 5.**

Kaplan-Meier estimates of the survival function, stratified by whether the subject has a higher or lower than median value of M1 (left), M4 (middle), and M1 and M4 (right). M1 is the mean SOFA score in the first 5% of one's ICU stay, whereas M4 is in the last 15%. p-values correspond to the log-rank test for the equivalence of the survival functions. “+” indicates censoring.

**Table 1**

Mean coverage probability of each of the four pointwise 95% confidence intervals (averaged across s), under each scenario. VI and V2 are Wald-type confidence intervals based on the model-based estimates of the variance $V_1$ and $V_2$, defined in Section 2.4. B-V is a Wald-type confidence interval based on the variance of the bootstrap estimates, and B-Q is constructed from the 2.5% and 97.5% quantiles of the boostrap distribution.

| N | Dataset | Opt. Method | $\beta_1(s)$ | | | | $\beta_2(s)$ | | | | $\beta_3(s)$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | V1 | V2 | B-V | B-Q | V1 | V2 | B-V | B-Q | V1 | V2 | B-V | B-Q |
| 100 | Full | AIC | 90 | 83 | 100 | 99 | 88 | 82 | 100 | 99 | 76 | 64 | 100 | 99 |
| | | $AIC_C$ | 87 | 81 | 100 | 98 | 86 | 82 | 100 | 98 | 70 | 58 | 99 | 95 |
| | | EPIC | 68 | 63 | 96 | 94 | 82 | 80 | 100 | 96 | 56 | 46 | 93 | 86 |
| | Inomplete | AIC | 63 | 54 | 98 | 98 | 88 | 81 | 100 | 98 | 66 | 41 | 85 | 82 |
| | | $AIC_C$ | 60 | 52 | 98 | 98 | 87 | 80 | 100 | 98 | 64 | 40 | 85 | 81 |
| | | EPIC | 50 | 44 | 96 | 96 | 84 | 79 | 100 | 97 | 55 | 36 | 82 | 77 |
| 200 | Full | AIC | 95 | 89 | 100 | 98 | 86 | 81 | 100 | 99 | 80 | 68 | 100 | 97 |
| | | $AIC_C$ | 95 | 88 | 100 | 98 | 86 | 81 | 100 | 98 | 76 | 64 | 99 | 95 |
| | | EPIC | 89 | 80 | 97 | 94 | 79 | 76 | 100 | 96 | 61 | 50 | 86 | 82 |
| | Inomplete | AIC | 78 | 65 | 95 | 98 | 86 | 80 | 99 | 98 | 64 | 37 | 73 | 71 |
| | | $AIC_C$ | 76 | 64 | 95 | 98 | 85 | 79 | 100 | 98 | 63 | 36 | 72 | 70 |
| | | EPIC | 65 | 56 | 91 | 96 | 78 | 75 | 100 | 97 | 53 | 33 | 69 | 67 |
| 500 | Full | AIC | 97 | 91 | 100 | 98 | 88 | 83 | 100 | 99 | 83 | 72 | 99 | 96 |
| | | $AIC_C$ | 97 | 92 | 100 | 98 | 88 | 82 | 100 | 99 | 82 | 71 | 99 | 95 |
| | | EPIC | 95 | 87 | 98 | 94 | 79 | 75 | 99 | 96 | 63 | 51 | 80 | 81 |
| | Inomplete | AIC | 92 | 68 | 93 | 97 | 87 | 79 | 99 | 98 | 66 | 27 | 62 | 63 |
| | | $AIC_C$ | 92 | 68 | 93 | 97 | 87 | 79 | 99 | 98 | 66 | 27 | 61 | 63 |
| | | EPIC | 88 | 65 | 90 | 96 | 80 | 74 | 98 | 96 | 56 | 25 | 59 | 60 |

**Table 2**

Predictive ability of each model, measured by the *N*-fold cross-validated C-Index. All models are adjusted for the scalar covariates age, gender, and CCI, and they differ in the way they model the length of stay $U_i$ and the SOFA function $X_i(s)$, as well as the method for optimizing any smoothing parameters. M1–M4 refer to the mean SOFA scores in the regions [0,0.05], [0.05,0.2], [0.2,0.4], and [0.85,1] of the *s*-domain, respectively. S1–S3 refer to the slopes of a regression line through the SOFA scores in the regions [0,0.15], [0.15,0.3], and [0.3,0.45].

| $U_i$ | SOFA | λ-Opt | C-Index |
|---|---|---|---|
| P-Spline | Functional Effect | AIC | 0.700 |
| P-Spline | Functional Effect | $AIC_C$ | 0.714 |
| P-Spline | Functional Effect | EPIC | 0.715 |
| Linear Spline | Functional Effect | AIC | 0.702 |
| Linear Spline | Functional Effect | $AIC_C$ | 0.730 |
| Linear Spline | Functional Effect | EPIC | 0.733 |
| Linear Spline | Ml + M2 + M3 + M4 | | 0.737 |
| Linear Spline | Ml + M2 + M4 | | 0.738 |
| Linear Spline | Ml + M4 | | 0.736 |
| Linear Spline | Ml | | 0.735 |
| Linear Spline | M4 | | 0.739 |
| Linear Spline | S1 + S2 + S3 | | 0.729 |
| Linear Spline | None | | 0.733 |
| None | None | | 0.739 |