HHS Public Access

Author manuscript

J Comput Graph Stat. Author manuscript; available in PMC 2022 January 01.

Published in final edited form as:

J Comput Graph Stat. 2021 ; 30(3): 780–793. doi:10.1080/10618600.2020.1853550.

Additive Functional Cox Model

Erjia Cui,
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, USA

Ciprian M. Crainiceanu,
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, USA

Andrew Leroux
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, USA; Department of Biostatistics and Bioinformatics, University of Colorado, Anschutz Medical Campus, USA

Abstract

We propose the Additive Functional Cox Model to flexibly quantify the association between functional covariates and time to event data. The model extends the linear functional proportional hazards model by allowing the association between the functional covariate and log hazard to vary non-linearly in both the functional domain and the value of the functional covariate. Additionally, we introduce critical transformations of the functional covariate which address the weak model identifiability in areas of information sparsity and discuss their impact on interpretation and inference. We also introduce a novel estimation procedure that accounts for identifiability constraints directly during model fitting. Methods are applied to the National Health and Nutrition Examination Survey (NHANES) 2003-2006 accelerometry data and quantify new and interpretable circadian patterns of physical activity that are associated with all-cause mortality. We also introduce a simple and novel simulation framework for generating survival data with functional predictors which resemble the observed data. The accompanying inferential R software is fast, open source and publicly available. Our data application and simulations are fully reproducible through the accompanying vignette.

Keywords

accelerometry; functional data; survival analysis; wearable devices

Corresponding author Erjia Cui ecui1@jhmi.edu.

Conflict of Interest: Dr. Crainiceanu is consulting with Bayer and Johnson and Johnson on methods development for wearable devices in clinical trials. The details of the contracts are disclosed through the Johns Hopkins University eDisclose system and have no direct or apparent relationship with this paper.

SUPPLEMENTARY MATERIALS

vignette_afcm.Rmd The vignette (designed for rnhanesdata package) containing code and instructions to reproduce all results shown in the paper. (.Rmd file)

vignette_afcm.html The HTML version of the vignette. (.html file)

process_data.R The R code to process and organize raw NHANES data into a data frame. Please follow the instructions in the vignette and do not run it separately. (.R file)

supplementary_materials.pdf Additional discussion about estimability and identifiability, comparison of methods that impose identifiability constraints, sample R code for simulating survival data, and additional simulation results. (.pdf file)

1 Introduction

We introduce a class of nonparametric additive functional Cox regression models for quantifying the association between a time to event outcome and functional covariates. This expands the rich literature on survival analysis by allowing for one or multiple functional covariates. It also expands the sparser literature on functional data analysis with survival outcomes by allowing a more flexible association between the functional covariate and time-to-event outcome. The approach is fully reproducible, fast, is implemented in R (R Core Team 2019), and can be used with minimal effort on personal laptops. Our work is motivated by the study of the association between time to death and physical activity (PA). PA has long been known to confer health benefits (Cooper et al. 2017) and has been associated with reduced risk of mortality (Matthews et al. 2016; Schmid et al. 2015). However, until the relatively recent development and adoption of wearable accelerometers, researchers relied on crude, inaccurate, and biased measures obtained from self-report questionnaires (Sallis and Saelens 2000; Silsbury et al. 2015). In contrast, accelerometers offer an unintrusive, continuous, and unbiased alternative to objectively measure PA over the course of several days, weeks, or even months. For these reasons they have been deployed in many large epidemiologic studies; see, for example, Bai et al. (2016); Doherty et al. (2017); Schrack et al. (2014); Troiano et al. (2008).

Here we are interested in quantifying the effect of timing and volume of PA on all-cause mortality in the National Health and Nutrition Examination Survey (NHANES). NHANES is a nationally representative study conducted by the Centers for Disease Control (CDC) to assess the health and nutritional status of adults and children in the United States. Participants were selected for inclusion according to the CDC sample design (Mirel et al. 2013) and assigned a survey weight based on the proportion the individual represents in the US population. Broadly, the NHANES data can be divided into three main categories: (1) questionnaire data, including responses to demographic, socioeconomic, dietary and health-related questions; (2) examination and laboratory data, including results of medical, dental, physiological measurements and laboratory tests; (3) accelerometer-measured PA. The processed NHANES 2003-2004 and 2005-2006 data are available in the R package rnhanesdata (Leroux et al. 2019a).

Specifically, the high resolution PA was measured by hip-worn accelerometers in the NHANES 2003-2004 and 2005-2006 waves. Each eligible participant was asked to wear the device for 7 consecutive days, and data were summarized in minute-level activity counts (a proprietary measure of PA intensity level). The minute-level activity counts are then transformed as $AC \rightarrow \log(1+AC)$ resulting in the log-transformed activity counts (LAC), which reduces the severe skewness of original data and is an appropriate measure of PA volume of lower levels of physical activity which have been adopted in the physical activity research literature (Varma et al. 2017, 2018). A sample of recorded minute-level LAC for one individual in NHANES is shown in the upper-left of Figure 1. Data are displayed on rows, where each row corresponds to a day of the week, where higher values correspond to more intense PA.

There are many different approaches for compressing and using these high dimensional accelerometry data. The most popular is to calculate a daily average (or sum) of LAC and then average these means (or totals) across days. This is illustrated in the right panel of Figure 1 by the horizontal arrows labeled “Take the average of each day” and by the arrow labeled “Take the average of daily averages”. The bottom panel illustrates a less aggressive summarising approach, where LAC are averaged at each time point across days. We will use this approach to create our functional covariates, which loses information on day-to-day variability in PA but retains substantially more information than traditional averaging over time of days and day. The pre-processing steps used to create the functional covariates is described in detail in Section 3.1.

Because NHANES can be linked to the National Death Index (NDI) released by National Center for Health Statistics (NCHI), it provides an unique opportunity to investigate the association between accelerometer-based PA measurements and time to death in a nationally representative sample. Figure 2 provides the intuition behind the problem and describes the data structure. PA data measured as minute-level LAC averaged over available days are shown for six study participants as a function of time of day. For each individual the data contain sociodemographic factors (age, race, employment status, education attainment, poverty-income ratio), health factors (self-reported overall health, smoking status, alcohol consumption, body mass index, mobility difficulty), and disease indicators (diabetes, coronary heart disease, congestive heart failure, stroke, cancer, systolic blood pressure, total cholesterol). For each study participant we display only their age, though much more additional information is available. The right panel in Figure 2 displays the mortality information. For example, the first study participant, who was 83 years old at the time the PA data were collected was deceased 2 years later (red horizontal line with a × symbol at the end to indicate a death event). The fourth study participant was 70 years old when the PA data were collected and was still alive 9.08 years later, the last time data were available for this individual (black horizontal line with a • symbol at the end).

In the NHANES 2003-2004 and 2005-2006 study, accelerometry data were collected from a total of 14631 study participants. For the purpose of this analysis, we exclude participants who: (1) were younger than 50 years of age, or 85 and older at the time they wore the accelerometer (10859 participants); (2) had fewer than 3 days of data with at least 10 hours of estimated wear time or were deemed by NHANES to have poor quality data (517 participants); (3) had missing covariates of interest, including age, employment status, educational attainment, poverty-income ratio, body mass index, self-reported overall health, coronary heart disease, congestive heart failure, stroke, cancer, diabetes, smoking status and alcohol consumption (436 participants); or (4) had missing mortality information (3 participants). The final data contained 2816 participants with 659 deaths in the first 10 years after the time PA data were collected. Individuals with observed mortality beyond 10 years are administratively censored at 10 years in our application. Surprisingly, there are few published methods for analyzing this type of data. In particular, Gellar et al. (2015), Qu et al. (2016) and Kong et al. (2018) proposed different versions of the “linear functional Cox model”, which included a linear functional term of the form $\int_S X_i(s)\beta(s)ds$ in the log-hazard expression to capture the effect of the functional covariate $\{X_i(s) : s \in S\}$. In practice we only observe $X_i(s)$ at a finite number of points. In our example, $X_i(s)$ is the

smoothed minute-level average log-transformed activity count (smoothed LAC) for study participant i at time s of day, and the domain S is midnight to midnight. We introduce three important methodological innovations: (1) extending the linear functional form to $\int_S F\{s, X_i(s)\}ds$, where $F(\cdot,\cdot)$ is an unspecified smooth function, as done by McLean et al. (2014) for generalized linear models; (2) introducing a flexible class of transformation functions for $X_i(\cdot)$ to account for the complexity of the NHANES accelerometry data; and (3) providing necessary assumptions and constraints to ensure the estimability and identifiability of the functional coefficient. We implement our method using easy-to-use software and provide a vignette which provides a detailed introduction of our model estimation procedure.

The remainder of the paper is organized as follows. Section 2 introduces the model and functional data transformations. Section 3 provides the results of the model applied to NHANES and interpretations. Section 4 proposes a simulation framework for both functional covariates and survival data. Section 5 summarizes the major findings and provides conclusions.

2 Methods

2.1 Model Setup

Motivated by the data structure illustrated in Figure 2, we model the log hazard function for $i = 1, \ldots, N$ study participants in the presence of independent right censoring. Denote the mortality event time as T_i and censoring time as C_i. We observe $Y_i = \min(T_i, C_i)$ and the event indicator $\Delta_i = I(T_i \leq C_i)$ for each study participant, where $I(\cdot)$ is the indicator function. The censoring time, C_i, is assumed to be independent of the event time, T_i, conditional on covariates. Suppose that at baseline we observe for each study participant p scalar covariates $\boldsymbol{Z}_i \in \mathbb{R}^p$, and a functional covariate $\boldsymbol{X}_i = \{X_i(s) : s \in S\}$. The framework extends to multiple functional predictors, but we use single functional predictor for presentation purposes. We assume that $\boldsymbol{X}_i$ takes values on a compact interval, and denote the partial information in the functional covariate up to s as $X_i^{\mathscr{P}}(s) = \{X_i(u) : u \leq s\}$. Hereafter we refer to this partial information as the "history" of the functional covariate, though this "history" is distinct from the notion of time as it relates to the survival process. Although the functional domain in our application is time of day, in other applications it may be, for example, space or some other argument. Using this notation we propose the following additive functional Cox model

$$\log \lambda_i(t \mid \boldsymbol{Z}_i, \boldsymbol{X}_i) = \log \lambda_0(t) + \boldsymbol{Z}_i^T \boldsymbol{\beta} + \int_{\mathcal{S}} F\{s, h_{is}[X_i^{\mathscr{P}}(s)]\}ds, \quad (1)$$

where $F(\cdot,\cdot)$ is an unspecified bivariate twice differentiable function; see McLean et al. (2014) for a similar approach in the context of outcomes from the exponential family. We discuss the identifiability of $F\{s, h_{is}[X_i^{\mathscr{P}}(s)]\}$ in Section 2.3.

This formulation allows the hazard function to vary smoothly with respect to both the functional domain and the value of the functional covariate, relaxing the assumption of linearity in the linear functional Cox model. We will show that this is important in our application, where activity during the day and night have opposite effects on the hazard

of mortality. Another innovation is to allow for a known subject- and domain-specific transformation, $h_{is}[X_i^{\mathscr{P}}(s)]$, of the partial history of the functional covariate up to time s of the functional domain, $X_i^{\mathscr{P}}(s)$. The main reason for considering transformations in the NHANES accelerometry data is that its structure is highly complex and exhibits substantial skewness, missingness, and heterogeneity within- and between-study participants. In addition, transformations of the functional data can be used to improve the estimability of $F(\cdot,\cdot)$; see our discussion in Section 2.3.

2.2 Transformations of the Functional Covariate

We consider two types of transformations, one that is domain-specific, $h_s(\cdot)$, and one that is subject/domain-specific, $h_{is}(\cdot)$. The difference is that the second type of transformation depends on the subject, i, in addition to the domain, s.

2.2.1 Domain-specific Transformations—The NHANES study activity data (minute-level LAC) shown in the left panel of Figure 3 indicates that during the night (1AM-4AM), PA measurements are much smaller than during the day. Therefore, estimating the function $F(\cdot,\cdot)$ on the entire rectangular domain [0,24]×[0,8] is nearly impossible. Here 24 stands for the number of hours in a day and 8 stands for an upper bound on the LAC. Indeed, there is basically no data in the [1,6]×[5,8] sub-domain. Therefore, estimates will be entirely driven by extrapolation of the smooth function $F(\cdot,\cdot)$ that borrows information from regions that are too far away to provide meaningful information. Fundamentally, the problem is that the function $F(\cdot,\cdot)$ cannot be well estimated in areas where there is little or no data. This is a limitation of the model and the primary motivation for our emphasis on transformation functions. The middle panel in Figure 3 displays the same data after smoothing each individual curve. Results indicate that the data sparsity becomes even more serious in certain parts of the domain of $F(\cdot,\cdot)$. Below we propose two classes of domain-specific transformations to address this issue.

Quantile transformation.: The first domain-specific transformation is the "quantile transformation", which takes the form

$$h_{is}[X_i^{\mathscr{P}}(s)] = h_s[X_i(s)] = \mathrm{P}(X(s) \leq X_i(s) \mid s)\,. \tag{2}$$

Here $\{X(s) : s \in S\}$ is the stochastic process on the functional domain S and $X_i(s)$ is the observed functional realization for the ith study participant. As a result, the functional covariate at each $s \in S$ is transformed to the cumulative distribution function (cdf) conditional on s. The right panel in Figure 3 displays the NHANES data after being smoothed and quantile-transformed. In contrast to the original data, these transformed data cover well its range, [0,24]×[0,1]. The difference is that the interpretation of $\widehat{F}(\cdot,\cdot)$ changes because the first argument is the relative, not absolute, size of the intensity of physical activity for an individual at a given time of day. More precisely, $\widehat{F}(s,\tau)$ is the effect of being in the τth quantile of the functional covariate (physical activity intensity) at time $s \in \mathscr{S}$. Results can be interpreted on the original scale of physical activity intensity $h_{is}^{-1}(\tau)$, but

interpretation of results should only be conducted in regions of the domain with sufficient data density.

A similar approach was proposed by McLean et al. (2014), who used the empirical cdf separately for each observed $s \in \mathcal{S}$, $\mathrm{P}(X(s) \leq X_i(s) \mid s) = N^{-1} \sum_{j=1}^{N} I(X_j(s) \leq X_i(s))$ to estimate the marginal cdf. When the functional covariate is irregularly sampled or measured with error we propose a complementary approach using additive quantile regression, which assumes smoothness of the quantiles of $X(s)$ across the functional domain. Specifically, consider the model

$$\mu_\tau(s) = f_{0,\tau}(s), \tag{3}$$

where τ is the quantile to be estimated, $\mu_\tau(s) = \inf[X_i(s): \mathrm{P}(X(s) \leq X_i(s) \mid (s) \geq \tau]$ is the τth quantile of $X(\cdot)$ given s, and $f_{0,\tau}(s)$ is a smooth function of s. Computationally stable estimation of $f_{0,\tau}(s)$ can be done via penalized splines (Fasiolo et al. 2017, 2019). Quantile regression estimates the inverse cdf and requires a separate model fit for each quantile of interest. Therefore, our estimator for (2) involves separate regression models for $\tau \in \boldsymbol{\tau}_0$ where $\boldsymbol{\tau}_0$ is a fine grid on (0, 1). Given these model fits, the quantile transformation is obtained by $\mathrm{P}(X(s) \leq X_i(s) \mid s) = \sup\{\tau: X_i(s) \leq \hat{\mu}_\tau(s), \tau \in \boldsymbol{\tau}_0\}$. While the empirical cdf approach may suffice in many applications, the proposed estimator can be extended to conditioning on subject-specific features, as discussed in Section 2.2.2.

Domain-specific standardization.: The second domain-specific transformation subtracts the domain specific mean and divides by the domain specific standard deviation:

$$h_{is}[X_i^{\mathcal{P}}(s)] = h_s[X_i(s)] = \frac{X_i(s) - E[X(s) \mid s]}{\sqrt{\mathrm{Var}[X(s) \mid s]}}, \tag{4}$$

where $E[X(s)|s]$ and $\mathrm{Var}[X(s)|s]$ can be estimated using their empirical estimators. After this transformation the interpretation of $\hat{F}(s, x)$ is the effect of being x standard deviations from the population mean at each $s \in \mathcal{S}$. Unlike quantile transformation, the domain standardization approach is more sensitive to skewness and may not cover the domain of $F(\cdot,\cdot)$ well.

2.2.2 Subject-specific Transformations—

We also consider transformations that depend on subject-specific characteristics. In our application we will use such transformations to conduct age-specific standardization of PA profiles. This will allow to assess the predictive power of PA on mortality independent of the natural decline of PA with age. Suppose that $\boldsymbol{U}_i$ is a q-dimensional vector of subject-specific characteristics and we want to extend the quantile transformation introduced in Section 2.2.1 to account for $\boldsymbol{U}_i$. In this extended setting, $X(s)$ is defined as the stochastic process on the functional domain S that also depends on $\boldsymbol{U}_i$.

Subject-specific quantile transformation.: Consider the subject/domain-specific transformation

$$h_{is}[X_i^{\mathscr{P}}(s)] = h_{is}[X_i(s)] = \mathrm{P}(X(s) \leq X_i(s) \mid s, \boldsymbol{U}_i)\,. \tag{5}$$

We propose to extend model (3) to the more general additive quantile regression model

$$\mu_\tau(s \mid \boldsymbol{U}_i) = \sum_{j=1}^{q} f_{j,\tau}(U_{ij}, s)\,. \tag{6}$$

The functions $f_{j,\tau}(\cdot,s)$ are smooth functions of each covariate and the functional domain $s \in S$. While the model may seem involved, it can be easily estimated by existing software; see, for example, the qgam package (Fasiolo et al. 2019) in R. Estimating (5) follows the same procedure described for the domain-specific quantile transformation. First, we estimate separate models for $\tau \in \boldsymbol{\tau}_0$ where $\boldsymbol{\tau}_0$ is a fine grid in (0, 1). Then, given these model fits, we estimate $\mathrm{P}(X(s) \leq X_i(s) \mid s, \boldsymbol{U}_i) = \sup\{\tau : X_i(s \mid \boldsymbol{U}_i) \leq \hat{\mu}_\tau(s \mid \boldsymbol{U}_i), \tau \in \boldsymbol{\tau}_0\}$. Note that extending the empirical cdf ideas to account for subject-specific covariates, $\boldsymbol{U}_i$, would be difficult, especially if the number of covariates is large.

Subject-specific standardization.: The second subject/domain-specific transformation is

$$h_{is}[X_i^{\mathscr{P}}(s)] = h_{is}[X_i(s)] = \frac{X_i(s) - E[X(s) \mid s, \boldsymbol{U}_i]}{\sqrt{\mathrm{Var}[X(s) \mid s, \boldsymbol{U}_i]}}\,. \tag{7}$$

As with the subject- and domain-specific quantile transformation, this transformation will likely involve some modeling of the first and second moments of $X(s)$ conditional on s and $\boldsymbol{U}_i$. Separate additive regression models for $E[X(s)|s,\boldsymbol{U}_i]$ and $E[X^2(s)|s,\boldsymbol{U}_i]$ with linear predictors of the same form as Model (6) could be used.

History of the functional domain.: The third subject/domain-specific transformation is

$$h_{is}[X_i^{\mathscr{P}}(s)] = \int_0^s X_i(u)du\,. \tag{8}$$

Just as with the other transformations, the interpretation of $F(\cdot,\cdot)$ changes compared to using the original functional covariates. For example, in the NHANES study $F(\cdot,\cdot)$ becomes "the effect of volume and timing of cumulative PA".

2.2.3 Choosing a transformation function—Choosing a transformation function for any given application is an open and important problem. We propose to choose the transformation function based on interpretability of results, ability to cover the domain of interest, and predictive performance. In our application predictive performance was roughly comparable for models with or without transformations, so the first two criteria took precedence. We also strongly suggest to display density plots and identify regions of

the space where there with sparse or no data. Model coefficients should not be interpreted in these areas, as little is known about extrapolation of complex nonparametric smoothers.

It could be tempting to jointly model the transformation function $h_{is}(\cdot)$ and $F\{s, h_{is}[X_i^{\mathcal{P}}(s)]\}$, though the exact procedure for doing so is not currently available. Such an approach would require the building of custom software and could substantially increase the computational complexity of the associated algorithms. To preserve computational efficiency and interpretability we do not pursue this idea, though this could be an important area for future research.

2.3 Identifiability

Wood (2017) discussed the necessity of adding constraints on the smooth functions to ensure the identifiability of additive models. Specifically, the constraint

$$\sum_{i=1}^{N} \int_{\mathcal{S}} F\{s, h_{is}[X_i^{\mathcal{P}}(s)]\} ds = 0\,, \tag{9}$$

is imposed by default when fitting an additive model using the R mgcv package. However, this constraint is not sufficient to ensure identifiability of the additive functional Cox model. For example, for any bivariate smooth function $F(s, x)$ let $g(s)$ be a function such that $\int_{\mathcal{S}} g(s)ds = 0$. If we define the function $F^*(s,x) = F(s,x) + g(s)$ then

$$\int_{\mathcal{S}} F^*(s,x)ds = \int_{\mathcal{S}} F(s,x) + g(s)ds = \int_{\mathcal{S}} F(s,x)ds + \int_{\mathcal{S}} g(s)ds = \int_{\mathcal{S}} F(s,x)ds \tag{10}$$

.

Therefore, the integrals are the same, but $F^*(s,x) \neq F(s,x)$ as long as $g(s) \neq 0$. Müller et al. (2013) proved that $F(s, x)$ is identifiable up to a function that does not depending on x. However, this result applies only in regions of the domain covered by $\{s, X_i(s)\}$. Hence, the identifiability condition is not sufficient to ensure that the model is estimable in areas of the domain sparsely covered or not covered by $\{s, X_i(s)\}$. The domain covered by $\{s, X_i(s)\}$ is often different from and much smaller than the rectangular domain defined by the minimum and maximum of s and $X_i(s)$ for all s and i. We refer to this as the “rectangular domain”.

We will show that this distinction is crucial in our application, where the functional coefficient is estimable only in a sub-region of the rectangular domain. This suggests that, when possible, transformations of the functional covariate should be considered to improve the coverage of the rectangular domain. This is particularly important as automatic nonparametric smoothers tend to work well on rectangular domains; see the supplementary materials for a more detailed discussion.

Suppose enough observations are available in the functional parametric region of interest. To address identifiability over the estimable domain we impose the additional identifiability constraints

$$\sum_{i=1}^{N} F\{s, h_{is}[X_i^{\mathscr{P}}(s)]\} = 0, \text{ for each } s \in \mathcal{S} \ . \tag{11}$$

These constraints restrict $F(s, x)$ at each $s \in S$ to have a unique form within the range of $h_{is}[X_i^{\mathscr{P}}(s)]$, thus ensuring identifiability over the area of interest. This restriction can be implemented directly in our software. The simulation results in Section 4 confirm that this approach provides a reasonable solution; see implementation details in Section 2.4.

2.4 Estimation and Inference

2.4.1 Penalized Spline Smoothing—Penalized splines smoothing (Ruppert et al. 2003; Wood 2017; Wood et al. 2016) and its connection with mixed effects modeling provide a powerful inferential platform for nonparametric regression modeling. Thus, pairing penalized spline smoothing and functional modeling (Goldsmith et al. 2011, 2012; Greven and Scheipl 2017; Scheipl et al. 2015) provides a modern, easy to implement, extendable framework for data analysis. Here we follow this principle and provide only the essential modeling details, as we consider penalized splines to be a mainstream inferential approach. Other methods include regressing on the functional principal component scores; see, for example, Müller and Yao (2008). While this approach leads to comparable predictive performance, the estimation of the functional parameter is highly sensitive to the choice of the number of principal components. For the bivariate case McLean et al. (2014) suggested using the tensor products of two univariate P-splines to model $F(\cdot,\cdot)$

$$F(s, x) = \sum_{j=1}^{K_S} \sum_{k=1}^{K_X} \theta_{j,k} B_j(s) B_k(x) \,, \tag{12}$$

where $B_j(\cdot)$ and $B_k(\cdot)$ are two univariate splines on the domains of s and x, respectively. The parameters $\{\theta_{j,k} : j = 1, 2, \ldots, K_s; k = 1, 2, \ldots, K_x\}$ are the spline coefficients. We use cyclic cubic regression splines for the functional domain s, and cubic regression splines for the functional covariate domain x. Given the excellent mgcv software in R this can be implemented as (users of mgcv will find this easy to understand):

```
ti (x, s, bs=c("cr", "cc"), k=c(Kx, Ks), mc=c (TRUE, FALSE))
```

The mc parameter specifies marginal centering constraints to the functional covariate domain, which coincides with the identifiability constraints (11) discussed in Section 2.3. The cyclic cubic regression splines are used for the functional domain, s, to account for the periodicity of PA as both $s = 0$ and $s = 24$ hours indicate midnight in our notation.

2.4.2 Estimation—Using the tensor product notation, the additive functional Cox model can be rewritten as

$$\begin{aligned}\log \lambda_i(t \mid \boldsymbol{Z}_i, \boldsymbol{X}_i) &= \log \lambda_0(t) + \boldsymbol{Z}_i^T \boldsymbol{\beta} + \int_{\mathcal{S}} F\{s, h_{is}[X_i^{\mathscr{P}}(s)]\} ds \\ &= \log \lambda_0(t) + \boldsymbol{Z}_i^T \boldsymbol{\beta} + \sum_{j=1}^{K_s} \sum_{k=1}^{K_x} \theta_{j,k} \int_{\mathcal{S}} B_j(s) B_k\{h_{is}[X_i^{\mathscr{P}}(s)]\} ds \\ &= \log \lambda_0(t) + \boldsymbol{Z}_i^T \boldsymbol{\beta} + \boldsymbol{V}_i^T \boldsymbol{\theta} \\ &= \log \lambda_0(t) + \boldsymbol{W}_i^T \boldsymbol{\gamma}\end{aligned} \tag{13}$$

.

Here $\boldsymbol{W}_i^T = (\boldsymbol{Z}_i^T, \boldsymbol{V}_i^T)$ and $\boldsymbol{\gamma}^T = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)$, where $\boldsymbol{\theta}$ is the vector with entries $\theta_{j,k}$ and $\boldsymbol{V}_i$ is the vector with entries $\int_{\mathcal{S}} B_j(s) B_k\{h_{is}[X_i^{\mathscr{P}}(s)]\} ds$, and both vectors $\boldsymbol{\theta}$ and $\boldsymbol{V}_i$ are organized in the same order of the indices $j = 1, 2, \ldots, K_s$; $k = 1, 2, \ldots, K_x$. The parameters $\boldsymbol{\gamma}$ are estimated by maximizing the penalized partial log likelihood, where the penalty is induced on the $\boldsymbol{\theta}$ parameters (the vector of parameters of the bivariate spline function) using standard quadratic penalties that depend on the vector of smoothing parameters λ. Selection of λ is discussed in Section 2.4.3. The penalized partial log likelihood has the following form

$$l_p(\boldsymbol{\gamma} \mid \lambda) = l(\boldsymbol{\gamma}) - \lambda J(\boldsymbol{\theta}) = \sum_{i=1}^{N} \delta_i [\boldsymbol{W}_i^T \boldsymbol{\gamma} - \log \sum_{Y_j \geq Y_i} e^{\boldsymbol{W}_i^T \boldsymbol{\gamma}}] - \lambda J(\boldsymbol{\theta}) . \tag{14}$$

For every fixed smoothing parameter λ, the estimator of the regression coefficients is obtained by $\hat{\boldsymbol{\gamma}}(\lambda) = \operatorname{argmin}_{\boldsymbol{\gamma}} - l_p(\boldsymbol{\gamma} \mid \lambda)$ using the Newton-Raphson algorithm. Detailed information on this approach can be found in Wood et al. (2016) supplementary materials G. Following ideas in Wood et al. (2016) we use cubic spline penalties. The practical implication of this approach is that it is easy to implement in the gam function of the mgcv package. For example, suppose that the functional covariates are observed on an equally-spaced grid $\{s_1, \ldots, s_m\}$ of the functional domain. The integral in equation (13) is approximated through weighted numerical summation of functional observations, where the weights are the increments between each neighboring pair and are stored in the vector l. In the case with only one scalar covariate, z, if the event indicator δ_i and observed survival time Y_i are stored in the variables delta and Y, respectively, the code is simply

```
fit <- gam(Y ~ z + ti(x, s, by = l, bs=c("cr", "cc"), k=c(Kx, Ks),
mc=c(TRUE, FALSE)), weights = delta, data, family = cox.ph())
```

The detailed procedure of fitting the model, extracting estimates on a fine grid, and visualizing the results is provided in the vignette in the supplementary materials. We would like to underline the simplicity of the code. This was possible because of the careful and novel methodological work and is an important contribution. Indeed, it is only through the use of powerful, reproducible, inferential code that functional methods can become popular after publication in highly specialized journals.

An alternative approach to Cox regression is to use estimation of the nonparametric proportional hazard model; see, for example, Lin et al. (2016) and Hiabu et al. (2017). However, here we focus on generalizing the Cox proportional hazard model.

2.4.3 Smoothing Parameter Selection—An important problem is the selection of the smoothing parameter λ. Several selection criteria have been proposed, including GCV (Gu 2013), AIC (Hurvich et al. 1998), EPIC (Shinohara et al. 2011) and REML (Ruppert et al. 2003). In the context of functional Cox regression, Gellar et al. (2015) proposed using a criteria based on AIC. Here we follow the estimation procedure described in Wood et al. (2016), which involves maximizing the Laplace approximation of the marginal likelihood of the smoothing parameter.

2.4.4 Statistical Inference—In addition to estimating the model parameters, $\boldsymbol{\gamma}$, the corresponding Hessian matrix $\boldsymbol{H}$ is also estimated; see Wood (2017) supplementary material G for details. Several estimators of the covariance matrix have been proposed in the literature, including a "sandwich estimator" $\boldsymbol{H}^{-1}\boldsymbol{G}\boldsymbol{H}^{-1}$ proposed by Gray (1992), and a "pseudo standard error" $\boldsymbol{H}^{-1}$ proposed by Verweij and Van Houwelingen (1994). Here $\boldsymbol{G}$ denotes the corresponding Hessian matrix without a penalty term. Therneau et al. (2003) recommended to perform significance tests on the estimator $\boldsymbol{H}^{-1}$ instead of $\boldsymbol{H}^{-1}\boldsymbol{G}\boldsymbol{H}^{-1}$. Because the structure of the problem does not change fundamentally, the inference for our model follows a similar procedure with that introduced by McLean et al. (2014) for the functional generalized additive model.

3 Applications

The additive functional Cox model was motivated by studying the association between the high-resolution physical activity measures and time to death. We present results using different transformations of the functional covariate, and compare their interpretation and predictive performance with those of traditional approaches.

3.1 NHANES

As discussed in Section 1, NHANES contains a large number of individual characteristics together with physical activity data measured by hip-worn accelerometry. Data are linked to mortality outcomes and are available, for example, through the rnhanesdata package in R. For more details on organizing and analyzing NHANES physical activity data see Leroux et al. (2019b). In our application, the functional covariate is the smoothed minute-level average LAC over available days, referred to as "smoothed LAC" below. We now describe the pre-processing procedure for creating the functional covariate (smoothed LAC). Denote the minute level activity counts $AC_{ij}(s)$ for subject $i = 1, \dots, N$, and day $j = 1, \dots, 7$, for minute $s = 1, \dots, 1440$. To account for subject non-compliance with study wear-time protocols, we use the default estimated wear/non-wear at every minute available in the rnhanesdata package, which were created using established algorithms (Troiano et al. 2008). Denote wear/non-wear indicators by $W_{ij}(s)$, where 0 and 1 correspond to estimated non-wear and wear, respectively. Next step introduces an indicator variable G_{ij}, which encodes a "good" day of accelerometry data as 1 and is defined as a day with at least 10 hours of estimated

wear time. More precisely, $G_{ij} = 1(\sum_{s=1}^{1440} W_{ij}(s) \geq 600)$, where $1(\cdot)$ is the indicator function. The index set for all good days for subject i is denoted by $J_i^* = \{j : G_{ij} = 1\}$. To create the functional predictor the daily activity counts are transformed as $LAC_{ij}(s) = g(AC_{ij}(s))$ where $g(y) = \log(1+y)$. This $g(\cdot)$ is introduced when building our functional predictors and is conceptually completely separated from the transformation function $h_{is}(\cdot)$ in the model. These log activity profiles are averaged across all the "good" accelerometry days for study participant i: $LAC_i(s) = |J_i^*|^{-1} \sum_{j \in J_i^*} LAC_{ij}(s)$. These individual profiles are then smoothed using FPCA. Therefore, we start our NHANES application with these smoothed LAC, denoted by $LAC_i'(s) = X_i(s) = \sum_{k=1}^{K} \tilde{\xi}_{ik} \hat{\phi}_k(s)$ where $\tilde{\xi}_{ik}$ are the predicted scores and $\hat{\phi}_k(s)$ are the estimated eigenfunctions obtained from functional principal component analysis (FPCA) (Xiao et al. 2016a). These steps are all pre-processing steps and are conceptually distinct from the subject-domain transformation function $h_{is}(\cdot)$.

Survival time is measured in months from accelerometer wear and, for the purpose of this study, all survival times are censored at 10 years. Among the 2816 study participants who met the inclusion criteria, 2157 (76.6%) survived for more than 10 years from the time when accelerometry data were collected. We adjust for sociodemographic factors (age, race, employment status, education, poverty-income ratio), health factors (self-reported overall health, smoking status, alcohol consumption, body mass index, mobility difficulty), and disease indicators (diabetes, coronary heart disease, congestive heart failure, stroke, cancer, systolic blood pressure, total cholesterol).

3.2 Application Results

3.2.1 Estimated Functional Surface $\widehat{F}(\cdot, \cdot)$—All models are fit using the R code described in Section 2. A vignette to reproduce the analysis is provided in the supplementary materials. Different transformations were used on the functional covariate, including identity transformation, quantile transformation, and subject-specific quantile transformation. We focus on the density of different transformed functional covariates and its connection with model estimation and interpretation.

To illustrate the complexity of the problem, Figure 4 displays density plots for the observations $\{s, X_i(s)\}$, where $X_i(s)$ is a generic notation for the LAC before or after transformation. First and second row correspond to individuals who were deceased within and alive for 10 years, respectively. First column: unsmoothed LAC. Second column: smoothed LAC. Third column: quantile-transformed smoothed LAC. The rectangular domain was partitioned into small sub-rectangles and the number of points $\{s, X_i(s)\}$ was counted in each sub-rectangle and plotted. For example, for unsmoothed and smoothed LAC the [0,24]×[0,8] rectangle domain was partitioned into 24 × 20 = 480 equal size rectangles, where each rectangle corresponds to one hour and an increment of 0.4 on the log(1+AC) scale. A similar partition (into 480 equal size sub-rectangles) was done for the quantile-transformed data, though the domain in this case was [0,24]×[0,1], because the quantile transformed data spans the [0,1] domain, whereas the original LAC data spans the

[0,8] domain. The number in each block decreases from red (largest) to blue (smallest). Color scales are different across plots.

The panels for unsmoothed LAC (left panels) show that data are extremely sparse in the sub-domain corresponding to high activity counts during the night; see the dark blue in the top-left region of the grid. This illustrates the estimability principle that we have discussed in this paper. Indeed, the regions [0,6]×[3,8] and [6,24]×[7,8] contain little or no data, despite the fact that we have a relatively large sample size (2816 study participants). An additional concern is that between 12AM to 6AM, when most people sleep, the density of observations is highly concentrated around zero counts. Thus, in this case, imposing the identifiability condition in Müller et al. (2013) is necessary but insufficient to ensure that we obtain meaningful estimates in these regions. In fact, we expect similar results even if the sample size were 100 times larger. The panels for smoothed LAC (second column) show that the problem is further exacerbated by smoothing. In contrast, the panels for quantile-transformed smoothed LAC (right panels in Figure 4) show a much better coverage of the rectangle domain [0,24]×[0,1]. This suggests that the quantile transformation could be an effective approach for addressing the estimability problem over the entire domain.

The estimates using smoothed LAC before and after transformations are shown in Figure 5, where each plot is visualized as a function of both the functional domain and the value of the functional covariate. The value of $\widehat{F}(\cdot, \cdot)$ decreases from red (highest) to white to blue (lowest), where a higher value corresponds to a higher hazard of death. The top-left panel in Figure 5 provides the functional surface estimates for the smoothed LAC ($h_{is}(x) = x$). A superficial look at the results could indicate that low activity intensity is associated with a higher hazard of mortality at any time of a day. This seems unreasonable, as a vast scientific literature exists on the benefits of restful sleep. We believe that this result is due to spurious extrapolation in regions of the functional domain with sparse or no data; compare these results with the data density panels in the second column of Figure 4.

To further explore whether this is, indeed the case, we conducted a stratified analysis by separating the time of day into night (12am to 8am) and day (8am to 12am). Results of this analysis are shown in the two side-by-side panels on the bottom in the first column (titled “Night” and “Daytime”). The interpretation of these results is that, higher activity during the night and lower activity during the day are associated with a higher hazard of mortality. However, even in this case the results during the night continue to be affected by data sparsity (left-upper and right-upper corners of the panel labeled “Night”). Another problem is that when conducting stratified analyses, the y-axis (which corresponds to the value of smoothed LAC) changes for each strata ([0,1.98] for night and [0,4.44] for day), which makes interpretation of results more difficult. The boundary value of the functional covariate is set to the 90th percentile of smoothed LAC for each time period to ensure good domain coverage. The lower boundary value at night is due to the lower LAC during the night. Moreover, the choice of threshold of 8am for night/day transition is debatable and 7am could provide a better transition point. Deciding which transition threshold to use is not obvious in practice, which further reduces the appeal of the stratified analysis.

For all these reasons we considered quantile transformations of smoothed LAC. As shown on the right panels of Figure 4, the improved coverage on the grid indicates that the functional surface is more likely to be estimable on the [0,24]×[0,1] grid of the transformed data. The top-right panel of Figure 5 indicates that lower relative activity during the day and higher relative activity during the night are associated with a higher hazard of mortality. Specifically, being below the 30th percentile of smoothed LAC in the population during daytime (9am to 9pm) is associated with a higher hazard of mortality. In contrast with the results in the top-left panel, this plot indicates that a lower relative LAC (less than the 35th percentile of smoothed LAC in the population) during the night (12am to 8am) is associated with a lower hazard of mortality. These results agree with those obtained from the stratified analysis.

While both approaches yield similar interpretable results, we favor the use of quantile transformation because: (1) the quantile transformation automatically unifies different scales of functional observations across the domain; (2) stratified analysis requires manual choice of the threshold, while the y-axis of the domain may be different; (3) results using quantile transformations are interpretable and translatable, whereas stratified analyses are based on quantities that are difficult to use for providing physical activity guidance; (4) the quantile transformation is easier to implement; and (5) the quantile transformation has a long and successful history in genomics analyses.

Building on the success of the quantile transformation, we have further applied the age-specific quantile transformation, where age is the subject-specific characteristic. This eliminates the effects of age on the individual quantile, as older individuals tend to have lower levels of activity. The result is illustrated on the bottom-right panel of Figure 5. The plot indicates that even after using age-specific quantile transformations, the pattern of the effect of diurnal and nocturnal activity intensity on the hazard of mortality remains relatively unchanged. Results indicate that individuals who are above the 60th percentile of activity during the night and below the 35th percentile during the day in their corresponding age group are at increased risk of mortality, irrespective of age.

3.2.2 Predictive Performance—Cross-validated Harrell's C-index (Harrell Jr et al. 1982, 1984, 1996) and Brier score (Brier 1950) are used as measures of predictive performance. Across models, the non-functional covariates are kept the same, allowing for a comparison of different approaches for modelling the association between activity and mortality while adjusting for common confounders. The results of 10-fold cross validation are shown in Table 1. Two functional models, additive functional Cox model ("AFCM") and linear functional Cox model ("LFCM"), are implemented as the comparison. For each functional model, we evaluate the predictive performance using three forms of LAC including unsmoothed, smoothed, and quantile-transformed smoothed. In addition, the non-functional Cox proportional hazard model ("Cox PHM") is implemented as the baseline model, where the the average smoothed LAC over the entire day is used as a scalar predictor.

The predictive performance of our models (with or without transformed data) is better than that of the linear functional Cox model and non-functional model, though differences are small. Among the additive functional Cox models the difference in predictive performance

is marginal. This may be due to the fact that the test and training datasets share the regions where the functional parameters are well estimated, irrespective of the transformation used. This indicates that using prediction measures may not be sufficient to differentiate between models that use raw or transformed data or among different types of transformations. However, interpretation of results is substantially improved by the quantile transformation and agrees with stratified analyses by time of day, as shown in Figure 5.

4 Simulation Study

4.1 Simulation Framework

For simplicity, we consider the case with only one functional covariate $\boldsymbol{X}_i$ and no scalar covariate. Consider the case when $h_{is}[X_i^{\mathcal{P}}(s)] = h_{is}[X_i(s)]$ and denote by $\eta_i = \int_{\mathcal{S}} F\{s, h_{is}[X_i(s)\}ds$. The model introduced in Section 2 can be simplified as

$$\log \lambda_i(t \mid \boldsymbol{X}_i) = \log \lambda_0(t) + \int_{\mathcal{S}} F\{s, h_{is}[X_i(s)]\}ds = \log \lambda_0(t) + \eta_i \ . \tag{15}$$

Functional covariates are simulated using functional principal component analysis (FPCA) (Ramsay 2004) applied to the NHANES data. Survival data are simulated using either the estimated $F(\cdot,\cdot)$ based on the NHANES data or pre-specified forms of $F(\cdot,\cdot)$ in combination with simulated functional covariates and estimated cumulative baseline hazards.

4.1.1 Simulating Functional Covariates—

FPCA has been widely used to smooth functional data by restricting the projection to the first M principal components of the Karhunen-Loève expansion (Karhunen 1947; Loeve 1978). If we denote by $\mu(s) = E[X(s)]$, then the subject-specific functional predictors can be expanded as $X_i(s) \approx \mu(s) + \sum_{j=1}^{M} \sqrt{\lambda_j}\xi_{ij}\psi_j(s)$. Here $\lambda_1 \geq \ldots \geq \lambda_M$ and $\psi_1(\cdot),\ldots,\psi_M(\cdot)$ are the first M eigenvalues and eigenfunctions, respectively. The scores are derived by $\xi_{ij} = \frac{1}{\sqrt{\lambda_j}}\int X_i(t)\psi_j(t)dt$ and $E(\xi_{ij}) = 0$, $E(\xi_{ij}\xi_{ik}) = I(j = k)$, which is equal to 1 if $j = k$ and 0 otherwise. The functional covariates $\widetilde{X}_i(s)$ are simulated as $\widetilde{X}_i(s) = \hat{\mu}(s) + \sum_{j=1}^{M} \sqrt{\hat{\lambda}_j}e_{ij}\hat{\psi}_j(s)$, where e_{ij} are i.i.d. $N(0, 1)$ random variables. The mean, $\hat{\mu}(s)$, eigenvalues, $\hat{\lambda}_j$, and eigenfunctions, $\hat{\psi}_j(s)$, are estimated using FPCA on the NHANES data. This was done using the R function fpca.face (Xiao et al. 2016b) in the refund package (Crainiceanu et al. 2012).

In our simulation, the functional covariates $\widetilde{X}_i(s)$ are generated by applying FPCA to the smoothed LAC, the functional covariates $X_i(s)$ of NHANES application. We then impose quantile transformation h_{is} on simulated functional covariates to reduce data sparsity observed on the middle panels of Figure 4. See R code in the supplementary materials for implementation details.

4.1.2 Simulating Survival Data—

Simulating survival data with non-pathological properties that mimic the NHANES data was one of the most difficult tasks addressed by this paper. We propose to use the estimated survival function, which proved to be

both practical and realistic. While methods for estimating survival times under parametric assumptions on the distribution of survival times exist (Austin 2012; Bender et al. 2005), we have been unable to adapt these methods to NHANES. Part of the problem is that small changes on the modeling assumptions can lead to substantial changes in the distribution of survival times. Moreover, we could not find a general set of recommendations on how to choose parameters, especially in the context of functional predictors.

Thus, we are taking a different approach and use the gam function in R package mgcv to estimate the cumulative baseline hazard $\widetilde{\Lambda}_0(t) = \int_0^t \lambda_0(u)du$ from the fitted model based on the NHANES data, where certain constraints are imposed to ensure non-negative and non-decreasing estimates; see R code in the supplementary materials for details. We use two simulation approaches to derive the estimated linear predictor $\widetilde{\eta}_i$ based on: (1) the surface estimated from NHANES; and (2) several pre-specified functional forms of $F(\cdot,\cdot)$. The estimated survival function is calculated as $\widetilde{S}_i(t) = \exp\{-e^{\widetilde{\eta}_i}\widetilde{\Lambda}_0(t)\}$, and the simulated survival time $\widetilde{T}_i$ is obtained using the relationship between the density and the survival function. The censoring times $\widetilde{C}_i$ are simulated from the empirical distribution of censoring times in the NHANES data to control the censoring rate.

In summary, the simulation procedure has the following steps: (1) derive the estimated cumulative baseline hazard function $\widetilde{\Lambda}_0(t)$; (2) derive the estimated linear predictor $\widetilde{\eta}_i$; (3) derive the estimated survival function $\widetilde{S}_i(t)$; (4) simulate survival time $\widetilde{T}_i$ from $\widetilde{S}_i(t)$; and (5) simulate censoring time $\widetilde{C}_i$ from the empirical distribution of censoring times in NHANES. The R code for this simulation approach is provided in the supplementary materials.

4.2 Simulation Results

As discussed in Section 4.1.1, we simulate functional covariates using FPCA on the NHANES data. We use two choices of $F(\cdot,\cdot)$, one based on NHANES and one based on pre-specified functional forms to evaluate model performance from different perspectives.

4.2.1 The Functional Surface Estimated from NHANES—We simulate survival and functional data that mimic real NHANES data with different sample sizes in the first simulation. The "true" $F(\cdot,\cdot)$ is set as the estimator using the quantile-transformed smoothed LAC in NHANES. In this section we show the model fitting performance using the correctly specified quantile transformation, while additional results using the misspecified identity transformation are included in the supplementary materials. The functional domain is rescaled to [0,1] for notation convenience. We focus on the estimation accuracy of the surface $F(s, x)$ and cumulative baseline hazard $\Lambda_0(t)$ under different sample sizes. The surface is estimated on the grid $\mathcal{S} \times \mathcal{X} = [0, 1] \times [0, 1]$ with 100 equally-spaced points in each dimension. Thus, the estimated surface is a 100 × 100 dimensional matrix where the value in each cell represents the estimated $\hat{F}(\cdot, \cdot)$ at that point in the domain. The cumulative baseline hazard function is estimated on the interval [0,10] on a 1000 dimensional equally-spaced grid of points.

The estimated surface based on quantile-transformed smoothed LAC of all $N = 2816$ NHANES participants is shown in the top-left panel of Figure 6, serving as the baseline for comparing the estimation performance of simulated data with different sample sizes. This plot is different from the application results in Section 3 since no other covariates are included. In simulations we used three sample sizes $N = 1000, 2000, 5000$. The sample size N controls the amount of information, in general, and the data density on the functional grid in particular. For each N, we performed 100 simulations and the average of the estimated surfaces are shown in Figure 6. A sample of randomly selected estimates from 100 simulations are included in the supplementary materials.

As sample size increases, the average estimated surfaces are getting closer to the baseline functional surface; see panels from left to right in Figure 6. These results provide a first check that the new simulation framework is reasonable and produces datasets with similar characteristics with the original NHANES. Moreover, the estimation method provides, at least on average, reasonable estimators of the target functional predictor surface. To better quantify how well surfaces and cumulative baseline hazards are estimated, the integrated square error (ISE) is calculated for each simulated data set. For surfaces ISE is defined as $\mathrm{ISE}(\widehat{F}(s,x)) = \int_{\mathcal{S}}\int_{\mathcal{X}}(\widehat{F}(s,x) - F(s,x))^2 dxds$, where $F(s, x)$ refers to the baseline functional term estimated from the real data and used in simulations. The bottom panel of Figure 6 displays the distribution of ISE as a function of sample size. Results illustrate a large decrease in ISE as sample size increases. More precisely, the median ISE when $N = 5000$ is less than a third the median ISE for $N = 1000$. The ISE for the cumulative baseline hazard functions is defined analogously and we show their distributions under different sample sizes in the supplementary materials. Further decompositions of ISE into integrated squared bias (denoted by "bias^2") and average variance (denoted by "variance") for both surfaces and cumulative baseline hazards are reported in Table 2. Results suggest that both bias and variance decrease as sample size increases. In addition, the estimation procedure is fast even for large sample sizes. Indeed, it took only ~ 2 minutes to obtain one fit with 5000 study participants on a regular laptop (2.7GHz dual-core Intel Core i5 processor), as shown in the right column of Table 2.

4.2.2 Pre-specified Functional Forms of $F(\cdot,\cdot)$—

We also considered pre-specified functional forms for $F(s, x)$, while keeping the simulation of the functional covariates the same. We considered the following functional forms for $F(s, x)$: (1) $F(s, x) = 2x$, which scales linearly with respect to x, and remains constant across s; (2) $F(s, x) = xs$, which scales linearly with respect to both x and s; (3) $F(s, x) = x^3 s$, which scales linearly with respect to s, but is nonlinear with respect to x; and (4) $F(s,x) = \sin xs$, which is nonlinear with respect to both x and s. The term $\int X_i(s)\beta(s)ds$ in the linear functional Cox model corresponds to $\beta(s) = 2$ in the first scenario and $\beta(s) = s$ in the second. However, the linear functional Cox model is misspecified for the last two scenarios. For each $F(s, x)$, we perform 100 simulations with sample size $N = 5000$ and derive the average estimated functional surfaces from each model. To reduce the linear approximation effect of nonlinear functions within small regions, for example $f(s, x) = xs$ and $g(s, x) = x^3 s$ are very close for x and s between 0 and 1, and to comply with the necessary identifiability constraints, the grid is

modified to $[0, 2]\times[-1,1]$ for all $F(s, x)$. The simulated functional covariates are rescaled to the same range to ensure good data coverage.

Figure 7 displays the true surfaces (first row) and the average estimated surfaces based on the linear (second row) and additive (third row) functional Cox model. The estimated surfaces and cumulative baseline hazard functions from a sample of randomly selected simulations are provided in the supplementary materials. The color scale is the same within each $F(s, x)$, but varies across different functions, as they have different ranges. The first two columns correspond to functions $F(s, x)$ that are linear in x. Both the linear and additive functional Cox models estimate the true surfaces well, at least when comparing the average surfaces. The ISE distributions shown in the last row indicate that the linear model performs slightly better, probably because of the higher complexity of the functional additive model. These results are expected and reassuring, indicating that the additive functional Cox model performs well when the true model is linear. The last two columns correspond to functions $F(s, x)$ that are nonlinear in x. In both scenarios the additive functional Cox model substantially outperforms the linear Cox model. This can be observed both from the comparison of the average of estimated surfaces (first three rows) and from the distributions of ISE (last row).

5 Discussion

The major contribution of our paper is the introduction of the nonparametric additive functional Cox model. This allows to quantify complex associations between a time to event outcome and functional covariates. This approach is crucial in the NHANES application where activity intensity during the night and day has different implications for the hazard of mortality. The technical argument is to use an unspecified bivariate function $F(s, x)$ that depends on the functional domain, $\mathcal{S}$, and the transformed functional covariates $h_{is}[X_i^{\mathcal{P}}(s)]$, where necessary constraints are imposed to ensure the model identifiability.

Another important contribution is to introduce a class of transformations of functional covariates, which can alleviate problems related to data sparsity in particular areas of the domain of the $F(\cdot,\cdot)$ function and substantially improve the model estimability. We have discussed several types of domain-specific transformations and extended the idea to subject-specific transformation. While the interpretation of results changes with the transformation, this provides a flexible approach for exploring the type of association between the functional predictors and time to event.

Our model was motivated by the NHANES study, where we identified highly interpretable patterns of association between daily trajectories of physical activity and the hazard of mortality. The prediction performance of the proposed model also improved slightly relative to the linear functional Cox model. Important advantages of the model are that it can be implemented using existing software, implementation is very fast even for large datasets, and reproducible code is provided with this paper.

We also introduced the first approach for realistic simulations of survival data for Cox models with functional predictors. Detailed R simulation code is provided in the

supplementary materials and the associated vignette. Simulations indicate that the additive functional Cox model performs almost as well as the linear functional Cox model when the function is linear and much better when it is not. A vignette is provided in the rnhanesdata package introducing and implementing all our work.

Our approach shows that complex functional models can be fit quickly and efficiently using state of the art software. However, our work has also opened several exciting avenues of research including establishing the theoretical properties of the estimation approach and exploring additional functional transformations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This work was supported by the National Institute of Neurological Disorders and Stroke under Grant Number R01 NS060910; and the National Institute on Aging under Grant Number T32 AG000247.

References

Austin PC (2012). Generating survival times to simulate cox proportional hazards models with time-varying covariates. Statistics in medicine, 31(29):3946–3958. [PubMed: 22763916]

Bai J, Di C, Xiao L, Evenson KR, LaCroix AZ, Crainiceanu CM, and Buchner DM (2016). An activity index for raw accelerometry data and its comparison with other activity metrics. PloS one, 11(8):e0160644. [PubMed: 27513333]

Bender R, Augustin T, and Blettner M (2005). Generating survival times to simulate cox proportional hazards models. Statistics in medicine, 24(11):1713–1723. [PubMed: 15724232]

Brier GW (1950). Verification of forecasts expressed in terms of probability. Monthly weather review, 78(1): 1–3.

Cooper R, Huang L, Hardy R, Crainiceanu A, Harris T, Schrack JA, Crainiceanu C, and Kuh D (2017). Obesity history and daily patterns of physical activity at age 60–64 years: findings from the mrc national survey of health and development. Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences, 72(10): 1424–1430.

Crainiceanu C, Reiss P, Goldsmith J, Huang L, Huo L, Scheipl F, Greven S, Harezlak J, Kundu M, and Zhao Y (2012). refund: Regression with functional data. R package version 0.1-6

Doherty A, Jackson D, Hammerla N, Plötz T, Olivier P, Granat MH, White T, Van Hees VT, Trenell MI, Owen CG, et al. (2017). Large scale population assessment of physical activity using wrist worn accelerometers: the uk biobank study. PloS one, 12(2):e0169649. [PubMed: 28146576]

Fasiolo M, Goude Y, Nedellec R, and Wood SN (2017). Fast calibrated additive quantile regression.

Fasiolo M, Goude Y, Nedellec R, and Wood SN (2019). qgam: quantile non-parametric additive models.

Gellar JE, Colantuoni E, Needham DM, and Crainiceanu CM (2015). Cox regression models with functional covariates for survival data. Statistical modelling, 15(3):256–278. [PubMed: 26441487]

Goldsmith J, Bobb J, Crainiceanu CM, Caffo B, and Reich D (2011). Penalized functional regression. Journal of Computational and Graphical Statistics, 20(4):830–851. [PubMed: 22368438]

Goldsmith J, Crainiceanu CM, Caffo B, and Reich D (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. Journal of the Royal Statistical Society: Series C (Applied Statistics), 61(3):453–469.

Gray RJ (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. Journal of the American Statistical Association, 87(420):942–951.

Greven S and Scheipl F (2017). A general framework for functional regression modelling. Statistical Modelling, 17(1-2):1–35.

Gu C (2013). Smoothing spline ANOVA models, volume 297. Springer Science & Business Media.

Harrell FE Jr, Califf RM, Pryor DB, Lee KL, and Rosati RA (1982). Evaluating the yield of medical tests. Jama, 247(18):2543–2546. [PubMed: 7069920]

Harrell FE Jr, Lee KL, Califf RM, Pryor DB, and Rosati RA (1984). Regression modelling strategies for improved prognostic prediction. Statistics in medicine, 3(2): 143–152. [PubMed: 6463451]

Harrell FE Jr, Lee KL, and Mark DB (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine, 15(4):361–387. [PubMed: 8668867]

Hiabu M, Mammen E, Martinez-Miranda MD, and Nielsen JP (2017). Smooth backfitting of proportional hazards–a new approach projecting survival data. arXiv preprint arXiv: 1707.04622

Hurvich CM, Simonoff JS, and Tsai C-L (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 60(2):271–293.

Karhunen K (1947). Über lineare Methoden in der Wahrscheinlichkeitsrechnung, volume 37. Sana.

Kong D, Ibrahim JG, Lee E, and Zhu H (2018). Flcrm: Functional linear cox regression model. Biometrics, 74(1):109–117. [PubMed: 28863246]

Leroux A, Crainiceanu C, Smirnova E, and Cao Q (2019a). rnhanesdata: Nhanes accelerometry data pipeline.

Leroux A, Di J, Smirnova E, Mcguffey EJ, Cao Q, Bayatmokhtari E, Tabacu L, Zipunnikov V, Urbanek JK, and Crainiceanu C (2019b). Organizing and analyzing the activity data in nhanes. Statistics in Biosciences, 11(2):262–287. [PubMed: 32047572]

Lin H, He Y, and Huang J (2016). A global partial likelihood estimation in the additive cox proportional hazards model. Journal of Statistical Planning and inference, 169:71–87.

Loeve M (1978). Probability theory. ii, volume 46 of. Graduate Texts in Mathematics.

Matthews CE, Keadle SK, Troiano RP, Kahle L, Koster A, Brychta R, Van Domelen D, Caserotti P, Chen KY, Harris TB, et al. (2016). Accelerometer-measured dose-response for physical activity, sedentary time, and mortality in us adults. The American journal of clinical nutrition, 104(5):1424–1432. [PubMed: 27707702]

McLean MW, Hooker G, Staicu A-M, Scheipl F, and Ruppert D (2014). Functional generalized additive models. Journal of Computational and Graphical Statistics, 23(1):249–269. [PubMed: 24729671]

Mirel L, Mohadjer L, Dohrmann S, Clark J, Burt V, Johnson C, and Curtin L (2013). National health and nutrition examination survey: estimation procedures, 2007–2010. Vital and health statistics. Series 2, Data evaluation and methods research, 159:1–17.

Müller H-G, Wu Y, and Yao F (2013). Continuously additive models for nonlinear functional regression. Biometrika, 100(3):607–622.

Müller H-G and Yao F (2008). Functional additive models. Journal of the American Statistical Association, 103(484): 1534–1544.

Qu S, Wang J-L, Wang X, et al. (2016). Optimal estimation for the functional cox model. The Annals of Statistics, 44(4): 1708–1738.

R Core Team (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay JO (2004). Functional data analysis. Encyclopedia of Statistical Sciences, 4.

Ruppert D, Wand MP, and Carroll RJ (2003). Frontmatter, pages i–vi. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Sallis JF and Saelens BE (2000). Assessment of physical activity by self-report: status, limitations, and future directions. Research quarterly for exercise and sport, 71(sup2): 1–14.

Scheipl F, Staicu A-M, and Greven S (2015). Functional additive mixed models. Journal of Computational and Graphical Statistics, 24(2):477–501. [PubMed: 26347592]

Schmid D, Ricci C, and Leitzmann MF (2015). Associations of objectively assessed physical activity and sedentary time with all-cause mortality in us adults: the nhanes study. PloS one, 10(3):e0119591. [PubMed: 25768112]

Schrack JA, Zipunnikov V, Goldsmith J, Bai J, Simonsick EM, Crainiceanu C, and Ferrucci L (2014). Assessing the "physical cliff': detailed quantification of age-related differences in daily patterns of physical activity. Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences, 69(8):973–979.

Shinohara RT, Crainiceanu CM, Caffo BS, and Reich DS (2011). Longitudinal analysis of spatiotemporal processes: a case study of dynamic contrast-enhanced magnetic resonance imaging in multiple sclerosis.

Silsbury Z, Goldsmith R, and Rushton A (2015). Systematic review of the measurement properties of self-report physical activity questionnaires in healthy adult populations. BMJ open, 5(9):e008430.

Therneau TM, Grambsch PM, and Pankratz VS (2003). Penalized survival models and frailty. Journal of computational and graphical statistics, 12(1):156–175.

Troiano RP, Berrigan D, Dodd KW, Masse LC, Tilert T, and McDowell M (2008). Physical activity in the united states measured by accelerometer. Medicine & Science in Sports & Exercise, 40(1): 181–188. [PubMed: 18091006]

Varma VR, Dey D, Leroux A, Di J, Urbanek J, Xiao L, and Zipunnikov V (2017). Re-evaluating the effect of age on physical activity over the lifespan. Preventive medicine, 101:102–108. [PubMed: 28579498]

Varma VR, Dey D, Leroux A, Di J, Urbanek J, Xiao L, and Zipunnikov V (2018). Total volume of physical activity: Tac, tlac or tac (λ). Preventive medicine, 106:233. [PubMed: 29080825]

Verweij PJ and Van Houwelingen HC (1994). Penalized likelihood in cox regression. Statistics in medicine, 13(23–24):2427–2436. [PubMed: 7701144]

Wood SN (2017). Generalized additive models: an introduction with R. Chapman and Hall/CRC.

Wood SN, Pya N, and Säfken B (2016). Smoothing parameter and model selection for general smooth models. Journal of the American Statistical Association, 111(516): 1548–1563.

Xiao L, Zipunnikov V, Ruppert D, and Crainiceanu C (2016a). Fast covariance estimation for high-dimensional functional data. Statistics and computing, 26(1-2):409–421. [PubMed: 26903705]

Xiao L, Zipunnikov V, Ruppert D, and Crainiceanu C (2016b). Fast covariance estimation for high-dimensional functional data. Statistics and Computing, 26:409–421. [PubMed: 26903705]

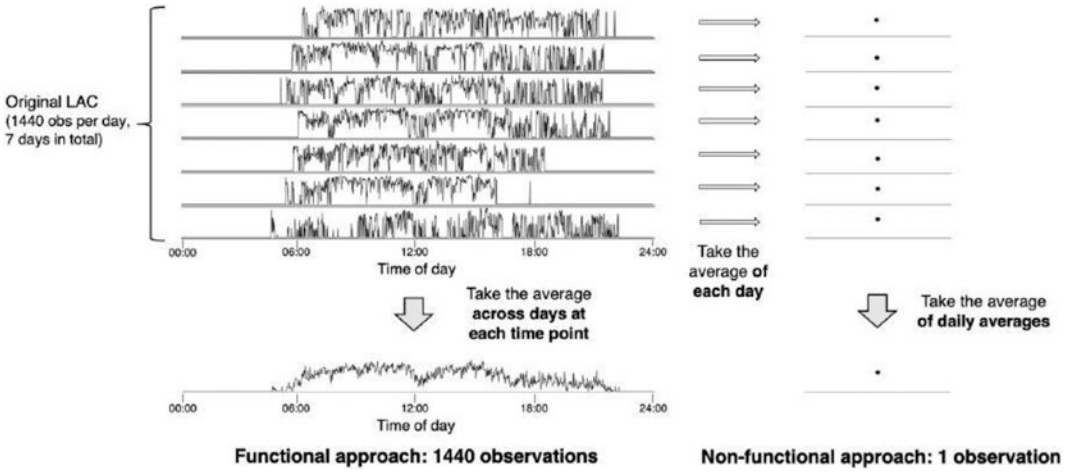

Fig. 1.

A sample of minute-level LAC for one individual in NHANES shown in the upper left and two summarising approaches. The right panel illustrates the traditional summarising approach, which calculates a daily average (or sum) of LAC and then averages these means (or totals) across days. The bottom panel illustrates a less aggressive summarising approach, where LAC are averaged at each time point across days.

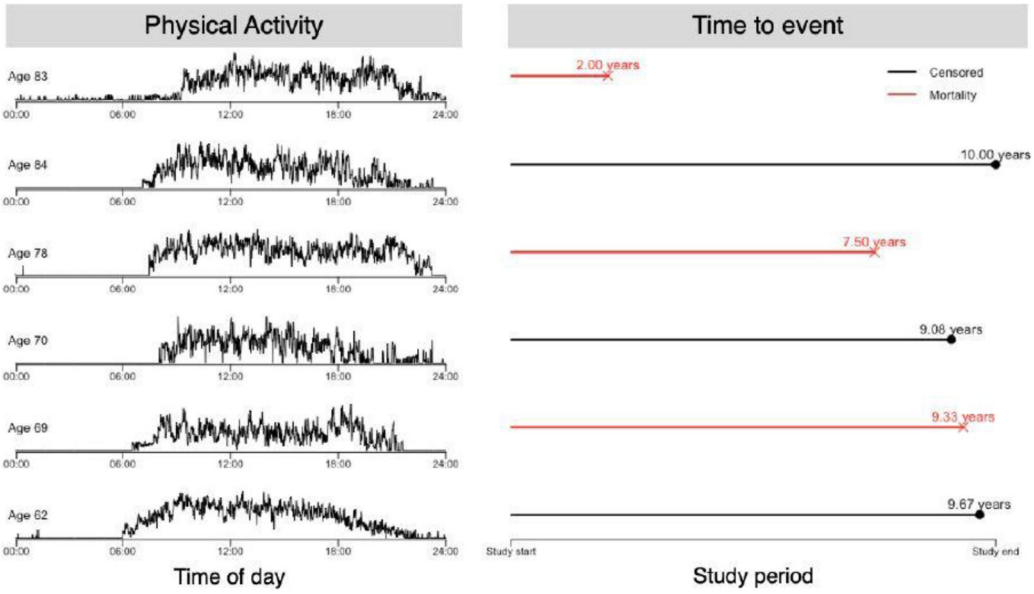

Fig. 2.

Physical activity and survival data of six study participants in NHANES. Each function represents the minute-level average LAC over the available days of valid data for that study participant. The age of the study participant is shown together with their mortality status (red for dead, black for alive) and the follow up time.

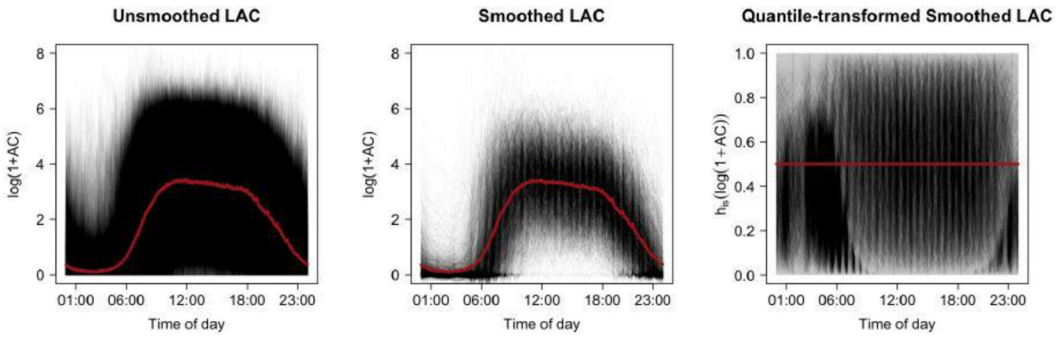

Fig. 3.

Distribution of the transformed minute-level LAC of all selected participants in the NHANES study, including unsmoothed (left), smoothed (middle), and smoothed + quantile transformation (right). The white top-left regions in the two left panels indicate the lack of high activity counts during the night.

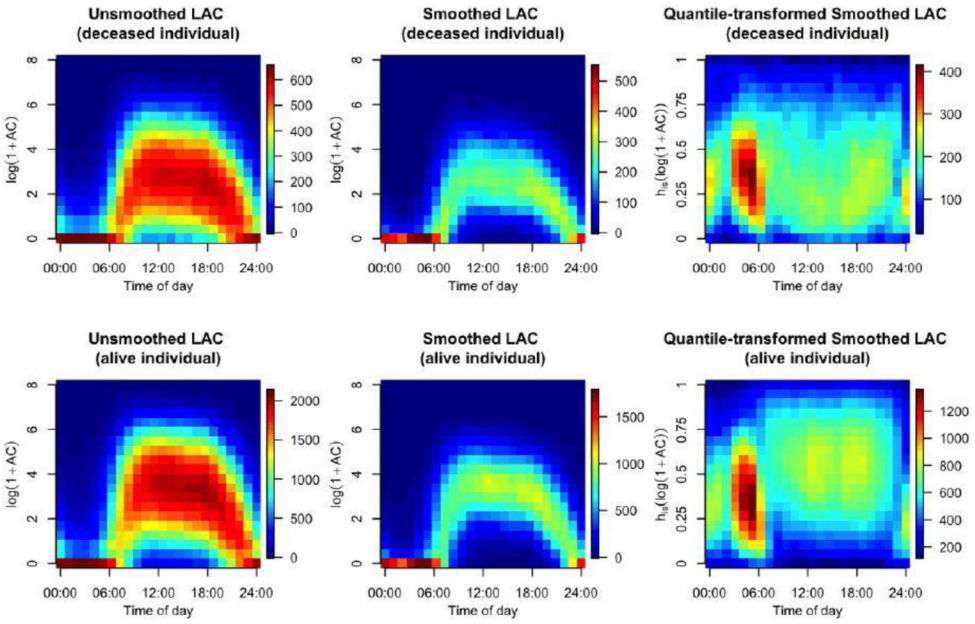

Fig. 4.

Density plots for the log-transformed activity counts (LAC) before or after transformation. First and second row correspond to individuals who were deceased within and alive for at least 10 years, respectively. First column: unsmoothed LAC. Second column: smoothed LAC. Third column: quantile-transformed smoothed LAC. The rectangular domain was partitioned into small sub-rectangles and the number of $\{s, X_i(s)\}$ was counted in each sub-rectangle and plotted. The number in each block decreases from red (largest) to blue (smallest). Color scales are different across plots.

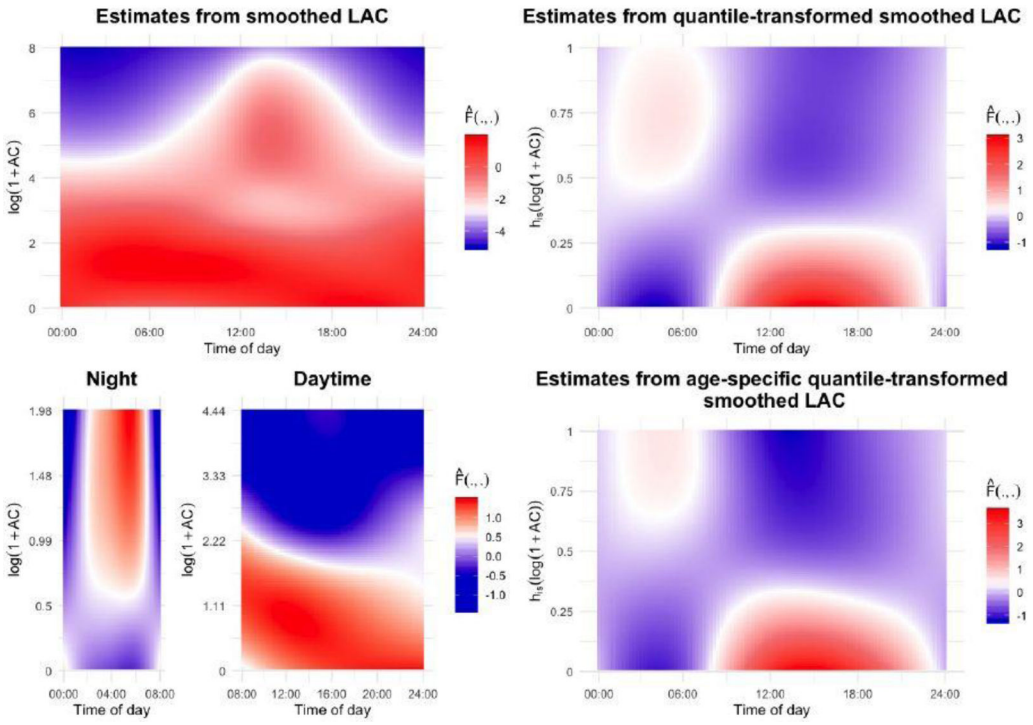

Fig. 5.

Estimated surfaces using the additive functional Cox model from untransformed (top-left), quantile-transformed (top-right), and age-specific quantile-transformed (bottom-right) smoothed LAC. The bottom-left panels show the estimates from smoothed LAC when stratifying the analysis by night (12am to 8am) and day (8am to 12am). For each time period of the stratified analysis, the functional covariate region of interest is set at the 90th percentile of the functional covariate values to ensure good coverage of data. The value of $\widehat{F}(\cdot, \cdot)$ decreases from red (highest) to blue (lowest hazard of death). Color scales are different across plots.

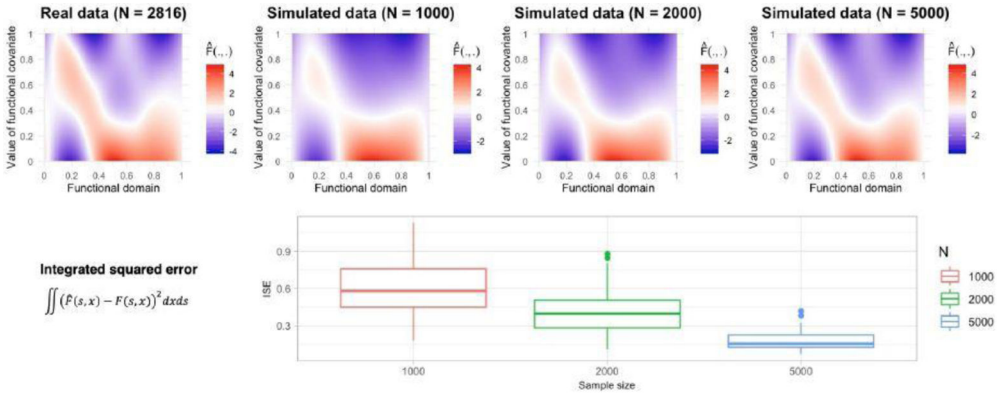

Fig. 6.

Estimated surface in NHANES, $\hat{F}(\cdot, \cdot)$, (first panel in the top row), which was used as true surface in simulations. Average estimated surfaces based on 100 simulations for N = 1000, 2000, 5000 (second, third, and fourth panel in the top row). Red, white, and blue correspond to highest, median, and lowest hazard of mortality. For each N, the distribution of the integrated squared error (ISE) is shown in the second row.

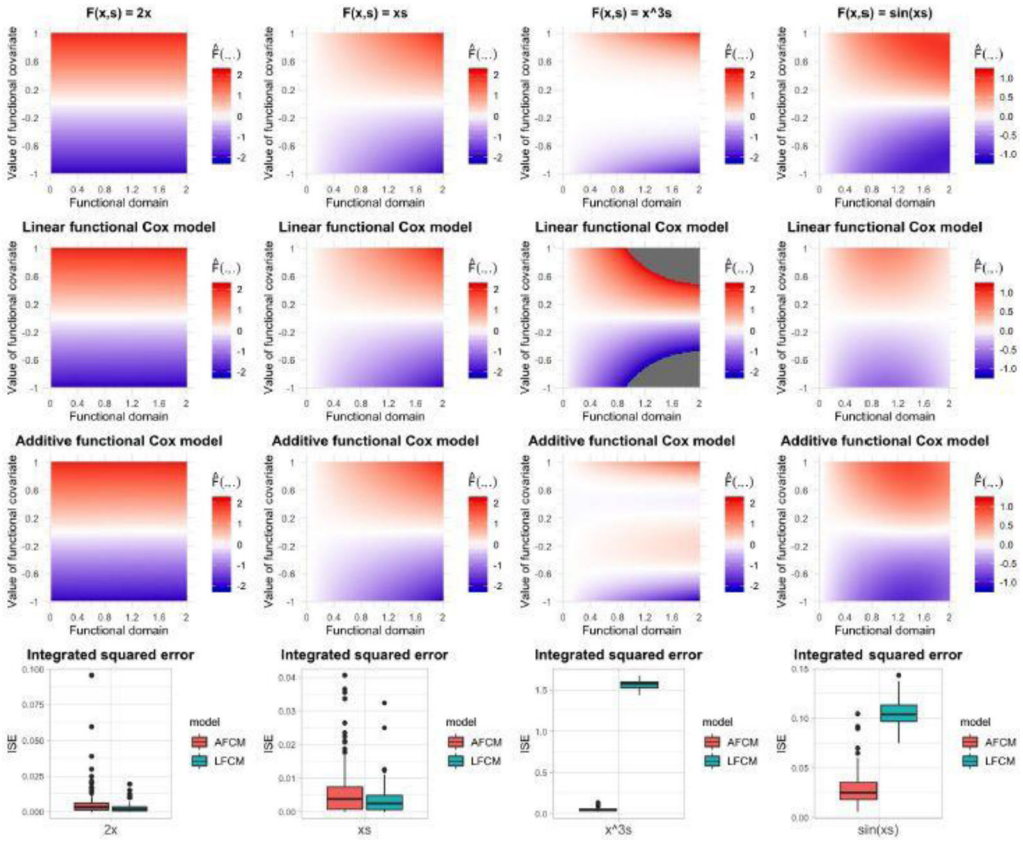

Fig. 7.

True surface (first row) and average estimated surfaces based on 100 simulations with sample size $N = 5000$ (second and third row). The second row corresponds to the linear functional Cox model and the third row corresponds to the additive functional Cox model. The fourth row displays the integrated squared error for the additive (red) and linear (blue) functional Cox models. Each column corresponds to a specific functional form of $F(\cdot,\cdot)$.

Table 1

The average 10-fold cross-validated Harrell's C-index and Brier score of all combinations of model and physical activity measures. "AFCM" denotes the additive functional Cox model, "LFCM" denotes the linear functional Cox model, and "Cox PHM" denotes the standard Cox proportional hazard model using the average activity as predictor.

Model	LAC	Harrell's C-index	Brier score
AFCM	unsmoothed	0.795	0.0751
	smoothed	0.795	0.0751
	smoothed + quantile	0.793	0.0753
LFCM	unsmoothed	0.791	0.0754
	smoothed	0.791	0.0754
	smoothed + quantile	0.791	0.0753
Cox PHM		0.791	0.0758

Table 2

The integrated squared bias and average variance for the estimated surface $\widehat{F}(\cdot, \cdot)$ and cumulative baseline hazard function $\widehat{\Lambda}_0(\cdot)$ based on 100 simulations with different sample sizes $N = 1000, 2000, 5000$. The average computing time per simulation is shown on the right column.

Sample size	$\widehat{F}(\cdot, \cdot)$		$\widehat{\Lambda}_0(\cdot)(\times 10^{-4})$		**Average Comp. Time (sec.)**
	bias^2	**variance**	**bias^2**	**variance**	
$N = 1000$	0.303	0.292	0.011	1.061	21.07
$N = 2000$	0.150	0.211	0.009	0.598	46.24
$N = 5000$	0.042	0.138	0.002	0.203	126.20