Functional Regression Methods for Densely-Sampled Biomarkers in the ICU

by

Jonathan E. Gellar

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

May, 2015

© Jonathan E. Gellar 2015

All rights reserved

Abstract

This thesis develops methods for modeling longitudinal predictors by treating them as functional covariates in regression models. First, I introduce Variable-Domain Functional Regression, which extends the generalized functional linear model by allowing for functional covariates that have subject-specific domain widths. I then propose a blueprint for the inclusion of baseline functional predictors in Cox proportional hazards models. Finally, I propose the Historical Cox Model, which introduces a new way of modeling time-varying covariates in survival models by including them as historical functional terms. Methods were motivated by and applied to a study of association between daily measures of the Intensive Care Unit (ICU) Sequential Organ Failure Assessment (SOFA) score and mortality, and are generally applicable to a large number of new studies that record a continuous variables over time.

ABSTRACT

Chair: Eliseo Guallar, M.D., Dr.P.H., JHSPH Epidemiology

Advisor: Ciprian M. Crainiceanu, Ph.D., JHSPH Biostatistics

Readers: Elizabeth Colantuoni, Ph.D., JHSPH Biostatistics

Dale M. Needham, F.C.A., M.D., Ph.D., JHU School of Medicine

Adam Spira, Ph.D., JHSPH Mental Health

Vadim Zipunnikov, Ph.D., JHSPH Biostatistics

Acknowledgments

I would like to express my deepest appreciation for my thesis advisor, Ciprian Crainiceanu. Without his experience, knowledge, and guidance, none of this work would have been possible. I could always count on Ciprian to give fair and honest advice in any situation, while also supporting and encouraging any decision I make, both personally and professionally. Above all, I cherish the friendship we have formed over the past five years, and I look forward to our relationship continuing for years to come.

I would also like to thank two other members of my thesis committee who have also served as mentors, one from a clinical perspective and one from a statistical one. Dale Needham was one of the first people I met at Johns Hopkins, and he first introduced me to clinical research. His hard work and dedication to his research are a big part of what inspired me to pursue my own research career. Elizabeth Colantuoni was my advisor during my time in the ScM program, and she has been a close collaborator and friend ever since. She has been an invaluable resource, both personally and professionally. In fact, her idea to investigate the use of functional regression techniques to model daily ICU variables is what inspired my entire dissertation.

I have had the pleasure of working with a number of amazing collaborators at Johns Hopkins. I am especially grateful for Michael Rosenbloom, Mei-Cheng Wang, Zheyu Wang, and Daniel Scharfstein, each of whom have given me many hours of stimulating discussion, both related to this thesis and to other projects. I am also indebted to Fabian Scheipl for his invaluable contributions towards the software implementation of these methods, and for inviting me to spend time at LMU Munich this past Novemeber. I would like to acknowledge Victor Dinglas and the Outcomes After Critical Illness and Surgery (OASIS) group, Dan Hanley and the Brain Injury OutcomeS (BIOS) team, Rick Thompson, Gayane Yenokyan, and the rest of the members of the Johns Hopkins Biostatistics Center, and Danny Reich and the Translational Neuroradiology Unit at NINDS. Each member of these groups has contributed to my well-rounded experience as a graduate student.

My graduate experience would not have been complete without the support and friendship of my fellow students. In particular I would like to thank my PhD cohort including Francis Abreu, Parichoy Pal Choudhury, Alyssa Frazee, and Yenny Webb-Vargas, as well as Benjamin Althouse, Rayman Huang, John Muschellli, and Elizabeth Sweeney, who have each contributed in some way towards my research. I have formed friendships and collaborations at Johns Hopkins that I hope will last the rest of my life.

On a personal note, I would like to thank my parents, Sheri and Jim for always believing in my abilities and encouraging my passion for math and science, and my sister-in-law and brother-in-law, Fan-fan Yu and Tom Bateson, for helping steer my career towards public health and biostatistics. And most importantly, to my wonderful wife, Rin-rin, for her patience as I tacked on year after year to my education. The past five years have been a roller coaster ride in every possible way, and I feel so lucky to have had her love, companionship, and support throughout it all.

Dedication

To Feilin and Zylan, who have made the past three years the craziest, most exhausting, and most rewarding time of my life.

Contents

Abstract **ii**

Acknowledgments **iv**

List of Tables **xii**

List of Figures **xiv**

1 Introduction **1**

2 Variable-Domain Functional Regression **5**

2.1 Introduction 5

2.2 Motivating Example 8

2.2.1 Data Description 8

2.2.2 Visualizing the Data 11

2.2.3 Approach 16

2.3 Variable-Domain Functional Regression 18

2.3.1 Model Specification 18
2.3.2 Estimation 19
2.3.3 Computational Issues 21
2.4 Expanded Class of Variable-Domain Models 23
2.4.1 Lagged Time 23
2.4.2 Domain-Standardization 24
2.4.3 Parametric Interactions 26
2.5 Simulation Studies 27
2.5.1 Simulation Design 27
2.5.2 Evaluation criteria 31
2.5.3 Simulation results 32
2.6 Application to ICAP Data 37
2.6.1 Model Specification 37
2.6.2 Model Performance 38
2.6.3 Estimated Coefficient Functions 41
2.7 Discussion 48

3 Cox Regression with Functional Predictors **52**

3.1 Introduction 52
3.2 Cox Model with Functional Covariates 57
3.2.1 Proportional Hazards Model 57
3.2.2 Estimation via Penalized Partial Likelihood 59

3.2.3 Optimization of the smoothing parameter 61
3.2.4 Inference 62
3.3 Extensions 63
3.3.1 Additional penalized model covariates 63
3.3.2 Full Likelihood Approach 64
3.3.3 Missing and unequally-spaced data 65
3.4 Simulation Study 65
3.4.1 Simulation design 65
3.4.2 Evaluation criteria 67
3.4.3 Simulation results 68
3.5 Effect of SOFA Score on Post-ICU Mortality 72
3.5.1 Data description 72
3.5.2 Analysis plan 74
3.5.3 Results 77
3.5.4 Follow-up Models 80
3.6 Discussion 82

4 The Historical Cox Model **85**

4.1 Introduction 85
4.2 The Historical Cox Model 88
4.2.1 Model 88
4.2.2 Estimation 91

4.2.3 Software Implementation 94
4.3 Extensions 95
4.3.1 Partial Integration Limits 95
4.3.2 Alternative Basis Parameterizations 96
4.3.3 Competing Risks 97
4.4 Simulations 98
4.4.1 Simulation 1: Estimation of $\beta(s, t)$ 100
4.4.2 Simulation 2: Performance Under Concurrent and Lagged Data-Generating Models 102
4.5 ICU Mortality among ARDS Patients 105
4.5.1 Data Description 105
4.5.2 Model Specification 107
4.5.3 Results: Historical Models 108
4.6 Discussion 112

5 Final Remarks **115**

Curriculum Vitae **134**

List of Tables

2.1 Summary statistics regarding the distribution of lengths of stay, within-subject mean SOFA score, and ICU gaps in the ICAP data, stratified by outcome. An ICU gap occurs when a subject is discharged from the ICU to a hospital ward, but later readmitted to the ICU prior to hospital discharge. Mean SOFA Score refers to the average SOFA score observed for each subject. 14

2.2 Median cross-validated AUC for all models applied to the simulated data, for the case when N=200, T_i is skewed, measurement error is present, and the outcome is binary. Models include seven functional models, five simple logistic regressions on the indicated summary statistic, and nine more complicated parametric or semi-parametric functions of the within-subject mean ($\bar{X}_i$) and the domain width (T_i). . . 33

2.3 AUC statistics for each model, applied to the binary outcomes of in-hospital mortality and physical function in the ICAP dataset. Results are presented as ${}_LX_R$, where X is the estimate, and (L, R) are the lower and upper bounds, respectively, of a 95% confidence interval based on 1000 bootstrapped samples. Both in-sample and cross-validated AUC statistics are presented. Cross validation is N-fold for the estimates, and 10-fold for the bootstrapped confidence intervals. All models are adjusted for age, gender, Charlson comorbidity index, and the log length of stay. Log length of stay is included as a smooth term for mortality and as a linear term for physical impairment. 39

3.1 Mean coverage probability of each of the four pointwise 95% confidence intervals (averaged across s), under each scenario. V1 and V2 are Wald-type confidence intervals based on the model-based estimates of the variance V_1 and V_2, defined in Section 3.2.4. 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. 71

3.2 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 $[0.3, 0.45]$. 81

4.1 C-Index for each historical model applied to the SOFA data, using the EPIC optimization criterion. C-Index values are expressed as ${}_LX_U$, where X is the C-Index, and (L, U) are the lower and upper bound of the 95% confidence interval based on 1000 bootstrap samples. p-values correspond to the null hypothesis that the C-Index for each indicated model is less than or equal to the C-Index for the concurrent model, which were ${}_{0.781}0.812_{0.844}$ for Death and ${}_{0.743}0.776_{0.807}$ for hospital discharge. α_0 indicates the range of integration for models fit with partial limits. 109

List of Figures

2.1 Exploratory plots. In plots (a), (b), and (c), NS = non-survivors, S:IPF = survivors with impaired physical function, S:UPF = survivors with unimpaired physical function, and S:N/A = survivors not assessed for physical function. The first two panels display the first 35 days of SOFA data as (a) a spaghetti plot and (b) a lasagna plot. Both subjects are separated into four groups, based on their values for the two outcomes (in-hospital mortality and physical impairment). In the spaghetti plot, color indicates outcome category. Four subjects are highlighted, and lines are used to connect adjacent measurements on the same subject, with gaps representing days where SOFA information was not available. In the lasagna plot, rows correspond to individual subjects, and darker colors are indicative of higher SOFA scores, i.e., poorer health. (c) Density estimates of the length of stay, stratified by the two outcomes, multiplied by the number of subjects in each stratum. (d) Mean SOFA functions that have been linearly compressed to a common domain, stratified by both outcome and ICU length of stay, for both mortality and physical function. Each LOS stratum contains approximately one quarter of the subjects for each outcome. 13

2.2 Simulation results for the case when $N = 200$, T_i is skewed, measurement error is present, and the outcome is binary. The top row depicts a heat map of the true coefficient functions. The second and third rows depict the root average mean squared error (rAMSE) of $\hat{\beta}(t, T_i)$ and 10-fold cross-validated area under the ROC curve (AUC), respectively, for each of the seven models. Smaller rAMSE and larger AUC indicate better model performance. Results are presented as Tufte box plots, with the median represented by a dot, the interquartile range by the white space around the dot, and the smallest and largest non-outlying points by the endpoints of the lines. Outliers are defined as values not within 1.5 times the interquartile range of the nearest quartile. Arrows indicate lines that extend outside the plotting range. 35

2.3 Heat maps of the estimate with median AMSE across the 1000 simulated datasets, for each model and coefficient function, in the case where $N = 200$, T_i is skewed, measurement error is included, and the outcome is binary. The range for each plot is from -6 (blue) to 6 (red), with values outside of this range indicated by white space. The numbers in the lower-right corner of each plotting area are the rAMSE statistics for each estimate. 36

2.4 Estimated coefficient functions for the association between daily SOFA score and in-hospital mortality in the ICAP dataset. Each column corresponds to one of our six functional models. In the top row of plots, estimates are depicted as $\hat{\beta}(t, T_0)$ for 10 evenly-spaced values of T_0. AUC statistics subject to 10-fold cross-validation are also provided. The bottom row displays the corresponding pointwise Z-scores, $\hat{\beta}(t, T_0)/SE(\hat{\beta}(t, T_0))$, as a function of t. The value of T_0 is indicated both by color and by the support of each function. The zero line is indicated with a horizontal dotted line, and dashed lines correspond to Z-scores of ±1.96. 42

2.5 Estimated coefficient functions for the association between daily SOFA score and impaired physical function in the ICAP dataset, presented similarly to Figure 2.4. 43

2.6 Trajectories of four hypothetical patients, along with their predicted probabilities of outcome according to each model. Top row corresponds to the mortality outcome, and bottom corresponds to impaired physical function. 46

3.1 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. 69

3.2 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. 74

3.3 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. 77

3.4 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. . . 78

3.5 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. 82

4.1 Illustration of how $\beta(s,t)$ acts on the covariate functions $X_i(t)$, at $t = t_0 = 75$. Both the coefficient function and covariate functions used in this figure are fabricated. (a) Heat map of a bivariate $\beta(s,t)$. For this illustration, we set $\beta(s,t) = \sin(2\pi s/100 - 2\pi t/100 - 3\pi/2)$. $\beta(s,t_0)$ can be viewed as a single horizontal slice from this heat map. (b) Two examples of potential covariate functions, $X_i(t)$. The weight function $\beta(s,t_0)$ is overlaid in red over each covariate function. (c) The weight function is multiplied by each covariate history $X_i^{\mathcal{H}(t_0)}(s)$, pointwise, to create the weighted covariate functions $X_i(s)\beta(s,t_0)$, $s \in [0,t_0]$. This function is then integrated over its domain and divided by t_0, with the result interpreted as the total contribution of $X_i^{\mathcal{H}(t_0)}(s)$ to $h_i(t_0)$. . . 89

4.2 Results for Simulation 1, for EPIC estimation under effect size $M = 2$. Each row of plots corresponds to a different true coefficient function, which is depicted in the first column as a heat map. The second column displays the median-performing estimate (as measured by AMSE) over 1000 datasets for the scenario where $N = 200$ and no domain transformation is used; this is done to provide context to the AMSE values. The third and fourth columns present Tufte box plots of the distributions of the AMSE and C-Index, respectively, stratified by sample size and domain transformation. For these plots, the median of each distribution is indicated by the center dot, the interquartile range by the white space around the dot, and the "whiskers" by the colored bars. Lower values of AMSE and higher values of C-Index are indicative of better model performance. 101

4.3 Results for Simulation 2, for AIC estimation with the $s - t$ domain transformation. δ_0 indicates the true lag of the model ($\delta_0 = 0$ indicates a concurrent effect), and α_0 indicates the maximum integration width ($\delta_0 = \infty$ corresponds to the full historical Cox model). The left panel of the figure depicts a heat map of the median-performing estimate (as measured by C-Index) for each combination of α_0 and δ_0, for the scenario when $N = 200$. Note that the heat maps are transformed to the lagged domain $(s - t)$ on the x-axis, to avoid excessive white space in the plots. The right panel of the figure displays the distribution of C-Index across the 1000 iterations of the simulation, for each model, true lag, and sample size. See Figure 4.2 for a description of Tufte box plots. The "Lagged" model assumes the true value δ_0 is known. . . . 104

4.4 SOFA scores in the ICAP dataset, stratified by event type. Left: lasagna plot of the first 35 days of SOFA scores, with each row corresponds to a subject and the SOFA score is indicated by color. Top right: spaghetti plot of each subject's last 10 (or fewer if $Y_i < 10$) SOFA scores, with a lowess smooth overlaid. Bottom right: density plot of the distribution of ICU length of stay, Y_i, on the log scale. . . 106

4.5 Estimation results for the historical Cox models fit to the SOFA data, using the s/t transformation and EPIC optimization criterion, for $t \leq 50$. Top left: Estimates of $\beta(s, t)$ for the full historical Cox model ($\alpha_0 = \infty$), with corresponding C-Index values listed. Top right: Estimates for the historical Cox models with partial integration limits, with the lagged scale $s - t$ used for the x-axis to prevent excessive white space. C-Indices for these fits (and others) appear in Table 4.1. Bottom: Time-varying AUC curves for the four historical models, as well as the concurrent time-varying covariate model. 110

Chapter 1

Introduction

Studies which collect longitudinal data have traditionally tracked subjects at a relatively small number of time points, often at sparse intervals over a long follow-up period. In recent years the face of medical research has changed. With the advent of wearable computing devices and federal mandates promoting the implementation of electronic health record systems, physicians and researchers are able to track a much denser collection of information over time to give a more complete picture of a patient's health status. This movement reflects larger trends in our society towards increased data collection and monitoring, and is aided by advances in our ability to collect, store, and transfer data.

The increased frequency with which this data is obtained should allow us to better understand disease processes and be able to personalize treatments towards individual patients as their condition changes over time. However, methods for modeling longitudinal data have in many ways failed to fully take advantage of the rich information contained in these datasets. A common assumption in models for longitudinal data, including survival models, is that the effect of a particular covariate is *concurrent*, i.e., the only value of the predictor that effects the outcome at any time is the last recorded value. Though this assumption may be relaxed, it usually involves specifying a rigid form for the way in which past values of the covariate can impact the outcome, such as a fixed time lag for the covariate.

In this thesis, I present a number of techniques for flexibly including longitudinally-measured predictors in regression models, with the aim of allowing one's entire trajectory of that covariate to impact the outcome. These methods incorporate approaches from the field of functional data analysis (FDA), which like longitudinal data analysis, is concerned with data consisting of repeated measurements over some domain (e.g., time). However, FDA approaches have largely ignored many of the challenges imposed by longitudinal data, including various types of missingness and censoring (Rice, 2004).

My research was inspired by questions arising from the ICAP study (Needham et al., 2006), a longitudinal cohort study that tracks patients with acute respiratory distress syndrome while they are in the intensive care unit (ICU), and for up to five years thereafter. While hospitalized in the ICU, various measurements of health are collected daily and summarized into the SOFA score, a composite biomarker that measures one's overall organ function. Our focus will be on relating these SOFA scores to various outcomes, in particular mortality.

This thesis consists of three journal articles, organized into chapters. Each of these chapters addresses a different challenge related to the inclusion of longitudinal predictors in regression models. The first article is concerned with including longitudinal predictors in generalized linear models when each subject's predictor is measured for a different length of time. The remaining two articles are concerned with time-to-event models. The former discusses the inclusion of a fixed width, baseline functional predictor in the Cox proportional hazards model (Cox, 1972), while the latter proposes a new way for treating time-varying covariates as functional effects in a Cox model. I now give a short overview of each of these three chapters.

Chapter 2 introduces the variable-domain functional regression model, which extends the generalized functional linear model by allowing for functional covariates that have subject-specific domain widths. The fundamental idea is to consider a bivariate functional parameter that depends both on the functional argument and on the width of the functional predictor domain. Both parametric and nonparametric models are introduced to fit the functional coefficient. The nonparametric model is theoretically and practically invariant to functional support transformation, or support registration. The paper has been published in a recent issue of the *Journal of the American Statistical Association* (Gellar et al., 2014).

Chapter 3 extends the Cox proportional hazards model to cases when the exposure is a densely sampled functional process, measured at baseline. This is done by combining penalized functional regression approaches 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. It may be extended to allow for multiple functional predictors, time varying coefficients, and missing or unequally-spaced data. This article is in press at *Statistical Modeling* (Gellar et al., 2015).

Chapter 4 introduces the historical Cox model, which further extends the Cox proportional hazards model to account for densely sampled time-varying covariates as historical functional terms. This approach allows the hazard function at any time t to depend not only on the current value of the time-varying covariate, but also on all previous values. The fundamental idea is to assume a bivariate coefficient function $\beta(s,t)$ that estimates a weight function that is applied to the full or partial covariate history up to t, and is allowed to change with t. Estimation is performed by maximizing the penalized partial likelihood, using a likelihood-based information criterion to optimize the smoothing parameter. The final version of this paper is in progress.

Chapter 2

Variable-Domain Functional Regression

2.1 Introduction

We study the relationship between a scalar response and a functional predictor, when the functional predictor falls on a fine grid with a different length for each subject. Such data are most commonly encountered when the domain variable is time, and each subject is followed for a different length of time. We refer to this type of data as variable-domain functional data. In particular, we were motivated by covariates collected in an inpatient hospital setting, where measurements are recorded daily (or at another fixed interval) for as long as the subject remains in the hospital. Examples of such measurements include measures of patient status, nutritional intake, and medication dosing. We are interested in understanding how the functional covariate affects an outcome collected at the end of hospitalization, or afterward.

The feature of subject-specific functional domains is not limited to the inpatient hospital setting. In sleep studies, subjects are connected to an electroencephalography (EEG) machine, which records electric activity for as long as the patient is asleep. Each subject sleeps for a different length of time, and one goal may be to relate these electrical signals of varying lengths to a subject-specific outcome or condition. In studies on aging, each subject lives for a different length of time, so the amount of available data varies by subject.

Traditional approaches to analyzing variable-domain functional data fall into two categories. The first consists of collapsing the trajectory of values into a summary statistic that can be used in a regression model. Common statistics include the mean, median, or maximum value, or the sum of available data. Alternatively, the slope from a linear regression, or other ad hoc statistics may be used (Dinglas et al., 2011; Sakr et al., 2012). These approaches ignore the functional nature of the data, and are inefficient as they throw away much of the available information. Additionally, the choice of summary statistic is often arbitrary, and not driven by the data.

The second common approach to modeling variable-domain functional data is to register each function to a common domain, and then apply existing functional regression techniques (Goldsmith et al., 2011). In certain contexts this is a perfectly reasonable approach. However, it might be less appropriate for data in which the between-subject variability in the width of the domain is extreme, or when the original time domain is informative. For example, in the ICU data described above, the subject-specific lengths of stay range from a single day to over 100 days. It does not seem natural to align functions to a common domain when they differ in width by orders of magnitude. Similar problems occur in sleep data, where registering shorter and longer sleep intervals to the same domain would fundamentally affect the observed sleep architecture.

In response to these problems we introduce a class of statistical models that incorporate the functional covariate and account explicitly for varying domains across subjects. *We assume that the primary analysis goal is to retrospectively explore the association between a functional covariate with subject-specific domain and a scalar outcome.* The novel aspect of our modeling approach is to allow the functional coefficient to vary, smoothly, according to the domain width. We refer to this type of regression as variable-domain functional regression (VDFR). Our approach is fast, flexible, and easy to interpret.

The remainder of this paper is organized as follows. In the next section, we describe our data in more detail and provide the necessary scientific context. Section 2.3 introduces the VDFR model, and describes one approach to estimating the associated parameters. In Section 2.4, we present a number of re-parametrizations of the VDFR to create a useful expanded class of models. Section 2.5 presents the results of a detailed simulation study, and we apply our model to the ICU data in Section 2.6.

We conclude with a discussion of what we have learned about regression on functions with variable domains.

2.2 Motivating Example

2.2.1 Data Description

The primary data for this analysis was taken from the Improving Care of Acute Lung Injury Patients (ICAP) study (Needham et al., 2006). Acute lung injury (ALI), also known as acute respiratory distress syndrome (ARDS), is a severe lung condition characterized by inflammation of the lung tissue (primary causes: pneumonia or sepsis). Patients with ALI/ARDS require mechanical ventilation in the intensive care unit (ICU), and experience high rates of mortality (Ware and Matthay, 2000). ICAP is a multi-site, prospective cohort study that enrolled 520 subjects with ALI/ARDS, 283 (54%) of which survived their hospitalization. Data for each patient are collected at baseline (enrollment into the study), daily while in the ICU, at hospital discharge or death, and among survivors, at seven follow-up points over five years.

Organ failure is measured by the Sequential Organ Failure Assessment (SOFA) score. The SOFA score is divided into six physiological components (respiratory, coagulation, liver, cardiovascular, central nervous system, and renal). Each component is assessed on a scale of 0-4 based on a set of physiological criteria, with larger values indicating poorer function. In cases where a physiological measurement is recorded repeatedly during the day, the worst 24-hour score is used. The component scores are then summed for a total SOFA score ranging from 0-24. Although it can only take integer values, we treat it as a continuous measure. The SOFA score is meant to be an overall measure of organ function, and is commonly used to track the physiological status of patients while in the ICU.

Thus, the observed data consist of $\{Y_i, Z_i, X_i(t_{ij}) : 0 \leq t_{ij} \leq T_i\}$, where i is the index for subject and j is the index for observation time, $j = 0, 1, \ldots, J_i$ and $\{t_{ij}\}$ are (not necessarily consecutive) integers with $t_{iJ_i} = T_i$ for all i. In this notation, $X_i(t_{ij})$ are the SOFA scores, recorded daily in the ICU, T_i is the length of stay in the ICU, Z_i are non-functional covariates, and Y_i is an outcome, recorded at the end of hospitalization, or afterwards. We assume that $\{X_i(t_{ij})\}$ are sampled from an underlying stochastic process $\{X_i(t) : t \in \mathcal{T}_i\}$, where $\mathcal{T}_i$ is an interval on the real line.

Our analysis will focus on two binary outcomes: in-hospital mortality, and physical impairment at hospital discharge among ICU survivors. For the mortality outcome, one possible approach would be to model the time-to-event process for the two competing events, death and hospital discharge, and treat SOFA as a time-varying covariate in a proportional cause-specific hazards model (Cox, 1972; Holt, 1978). This approach could be extended to treat the SOFA scores as a longitudinal outcome in a joint model for the longitudinal and survival processes. Indeed, joint models for longitudinal and survival data have been the focus of intense research over the past two decades (Brown and Ibrahim, 2003; Hanson et al., 2011; Ibrahim et al., 2001, 2010; Rizopoulos, 2012; Tsiatis and Davidian, 2004; Tsiatis et al., 1995; Wang and Taylor, 2001; Yu et al., 2004). An advantage of this modeling strategy is that it would allow for dynamic prediction of mortality, i.e., the ability to estimate whether or not a person will survive their ICU stay while they are still in the hospital (Garre et al., 2008; Proust-Lima and Taylor, 2009; Rizopoulos, 2011; Yu et al., 2008).

While this would certainly be a clinically important goal, it is not the focus of our analysis. Instead, our scientific problem is different: given a group of patients who died in the ICU and a group who survived, each with a different length of stay, how can we compare their within-ICU health trajectories? To accomplish this objective, we treat the outcome as a binary indicator of mortality, and we condition on each subject's entire SOFA curve (including its domain length, T_i). Since we need to wait until the end of one's hospitalization in order for T_i to be known, our methods will not be useful for dynamic prediction of mortality. Instead, our analysis is a retrospective analysis that aims to identify the precise features of one's SOFA curve that differ between survivors and non-survivors. This allows us to better understand how patterns of dynamic organ failure differ between these two groups, and provides a way to quantify these differences.

Our second outcome is physical function at hospital discharge, measured using the Activities of Daily Living (ADL) scale (Katz et al., 1963). This questionnaire consists of six tasks, and for each one the subject indicates whether they can accomplish the activity independently, or that they require assistance. ADL information is available at both baseline and at hospital discharge, and at both time points the number of dependencies (i.e., total activities for which the subject requires assistance) are calculated. In order to isolate the effect of one's hospital experience on physical function, the baseline number of dependencies is subtracted from the number of dependencies at discharge, and this number is dichotomized at ≥ 3. Thus, the outcome of interest will be whether or not the subject required assistance with three or more tasks than they did at baseline, a condition we refer to as "physical impairment." The subjects who had 4 or more dependencies at baseline were removed from this analysis, as they were not eligible to experience the outcome. Of the 283 hospital survivors, 34 did not consent to followup, 1 was missing baseline ADL data, and 17 were ineligible for the outcome, resulting in a sample size of 231. Since this outcome is not available until hospital discharge, which typically occurs a few days or weeks after ICU discharge, the model may be treated as a predictive model.

2.2.2 Visualizing the Data

Exploratory plots of the data are presented in Figure 2.1. Plots (a) and (b) contain two depictions of the first 35 days of SOFA data. Both plots are stratified by the two outcomes: in-hospital mortality, and impaired physical function. Subjects are aligned according to the day of their onset of ALI/ARDS, which also corresponds to the first recorded SOFA measurement; this time point is indicated as day 0. We highlight four individual subjects in the spaghetti plot (Figure 2.1a) to provide some context. The patient indicated by the purple line entered the ICU with a moderate SOFA score of 11, but his health steadily declined (as indicated by an increasing SOFA score) until his death on the 11th day. The blue subject, on the other hand, started with a more severe initial score of 14, but his health rapidly improved, and he was discharged alive from the ICU on the fifth day without impaired physical function. The black and red subjects are examples of subjects with gaps in their curves. This occurs when a subject is discharged from the ICU to a hospital ward and later readmitted to the ICU; SOFA is not collected in the ward. The black subject entered the ICU with a score of 17, but improved enough to be discharged from the ICU to a hospital ward on his 12th day. However, he was re-admitted to the ICU four days later, and rapidly deteriorated until dying on the 24th day from baseline. The red subject was discharged from the ICU on his 10th day, was re-admitted 5 days later, and eventually was discharged a final time from the ICU on his 35th day.

At least one gap similar to those observed in the black and red highlighted curves occurs in 33 of the 520 subjects (6%), causing 364 of the 8879 potential patient days (4%) to be missing (Table 2.1). The missingness is potentially informative, as patients are healthier when outside of the ICU than inside it, but models that account for the missing data mechanism are outside of the scope of this paper. Instead, since our method requires dense and equally spaced data, we impute these days using last observation carried forward (LOCF) based on advice from clinical experts.

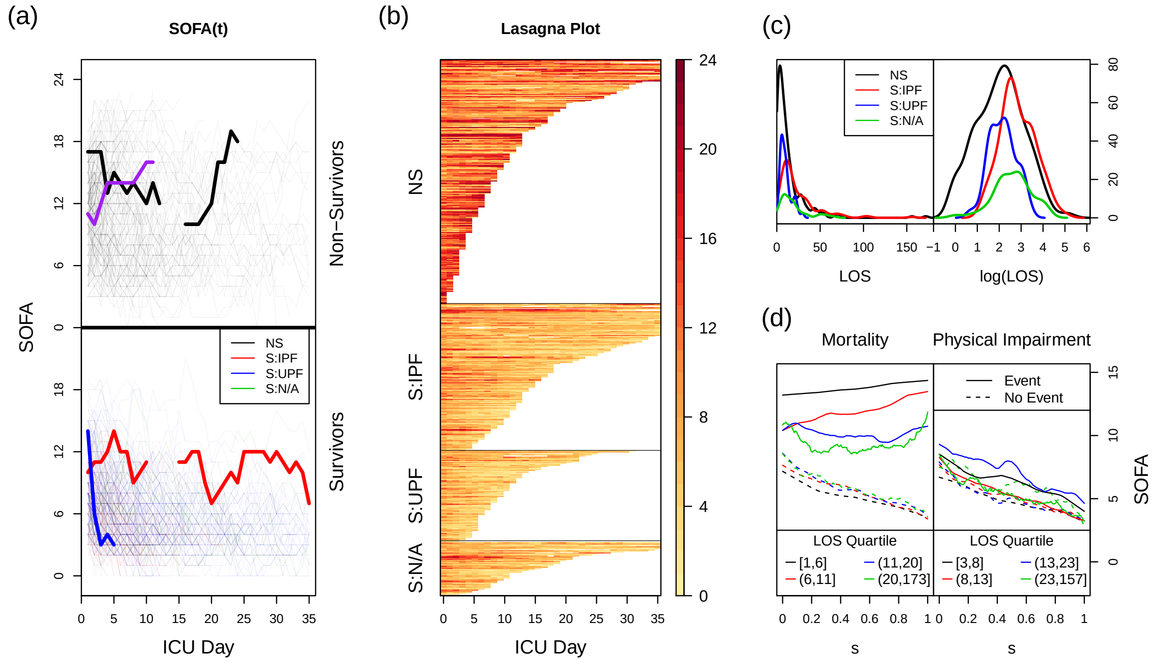


Figure 2.1: Exploratory plots. In plots (a), (b), and (c), NS = non-survivors, S:IPF = survivors with impaired physical function, S:UPF = survivors with unimpaired physical function, and S:N/A = survivors not assessed for physical function. The first two panels display the first 35 days of SOFA data as (a) a spaghetti plot and (b) a lasagna plot. Both subjects are separated into four groups, based on their values for the two outcomes (in-hospital mortality and physical impairment). In the spaghetti plot, color indicates outcome category. Four subjects are highlighted, and lines are used to connect adjacent measurements on the same subject, with gaps representing days where SOFA information was not available. In the lasagna plot, rows correspond to individual subjects, and darker colors are indicative of higher SOFA scores, i.e., poorer health. (c) Density estimates of the length of stay, stratified by the two outcomes, multiplied by the number of subjects in each stratum. (d) Mean SOFA functions that have been linearly compressed to a common domain, stratified by both outcome and ICU length of stay, for both mortality and physical function. Each LOS stratum contains approximately one quarter of the subjects for each outcome.

Table 2.1: Summary statistics regarding the distribution of lengths of stay, within-subject mean SOFA score, and ICU gaps in the ICAP data, stratified by outcome. An ICU gap occurs when a subject is discharged from the ICU to a hospital ward, but later readmitted to the ICU prior to hospital discharge. Mean SOFA Score refers to the average SOFA score observed for each subject.

		Mortality		**Physical Function**	
	All Subjects (N=520)	**Non-survivors** (N=237)	**Survivors** (N=283)	**Impaired** (N=142)	**Unimpaired** (N=89)
Length of Stay:					
Mean (SD)	17.1 (19.0)	14.2 (19.3)	19.5 (18.4)	24.6 (22.3)	11.2 (6.4)
Median (IQR)	11.0 (6.0, 20.0)	8.0 (4.0, 17.0)	13.0 (8.5, 23.0)	16.5 (11.0, 32.0)	10.0 (6.0, 13.0)
Range	(1, 173)	(1, 173)	(2, 157)	(4, 157)	(3, 31)
Mean SOFA Score:					
Mean (SD)	8.5 (4.6)	11.9 (4.3)	5.6 (2.5)	6.0 (2.6)	4.9 (2.0)
Median (IQR)	7.2 (4.6, 11.6)	12.0 (8.3, 14.9)	5.0 (3.7, 7.2)	5.3 (4.0, 7.4)	4.6 (3.4, 6.1)
Range	(1.2,22.0)	(2.7,22.0)	(1.2,14.8)	(1.9,14.8)	(1.6,11.2)
ICU Gaps:					
Subjects Affected (%)	33 (6%)	12 (5%)	21 (7%)	13 (9%)	2 (2%)
Patient Days Affected	364/8879 (4%)	132/3362 (4%)	232/5517 (4%)	174/3494 (5%)	17/1001 (2%)

We conducted a number of sensitivity analyses, such as excluding the 33 subjects whose data contained gaps, and results remained relatively unchanged (supplemental material). This leads us to believe that any bias introduced by the LOCF imputation has minimal effect on our results.

Density estimates for the ICU length of stay, T_i, are displayed in Figure 2.1c, with summary statistics presented in Table 2.1. We see that subject-to-subject variability in terms of length of stay is quite extreme. There are several subjects for whom only a single SOFA measurement is available (all of whom died on that day), while others remained in the ICU for over 100 days. The median length of stay is 11 days, with survivors tending to remain in ICU longer than non-survivors. Accounting for this heterogeneity in the length of stay will be a key challenge that our method must address.

Some trends in the data are easy to see; for example, higher SOFA scores, shorter times in the ICU, and greater SOFA variability (both within and between subjects) are more common among the non-survivors than survivors. Indeed, a simple logistic regression of in-hospital mortality on each subject's mean SOFA score performs well in terms of discrimination, resulting in a cross-validated area under the receiver operating characteristic curve (AUC) of 0.89. However, the question remains whether we could do better by considering the entire SOFA curve, without collapsing the values into a single summary statistic such as the mean. It is much more difficult to visually identify patterns that differentiate SOFA scores between impaired and unimpaired physical function, except that those with impaired physical function tend to remain in the ICU for longer than those who are unimpaired.

An alternative way of exploring SOFA trends across different lengths of stay is displayed in Figure 2.1d. Here, individual SOFA functions have been linearly compressed to a common domain from 0 to 1, a procedure that we refer to as domain-standardization (this will be discussed in more detail in Section 2.4.2). We then plot the mean SOFA function for each outcome category, stratified into four groups by length of stay. For mortality, we see a clear separation in the mean functions of the survivors (dashed lines) as compared to those who died (solid lines). Interestingly, the mean function of the survivors is quite consistent regardless of T_i. We do see differences in the mean function of the non-survivors according to T_i, with the functions decreasing and becoming more "U-shaped" as T_i increases. We do not see nearly as strong of a differentiation between those with (solid lines) and without (dashed lines) impaired physical function. In fact, the mean functions for those with and without physical impairment are virtually indistinguishable from each other in the $(8, 13]$ and $(23, 157]$ strata. In the other two strata there is a tendency for those who had impaired physical function on hospital discharge to have elevated SOFA scores. We do not observe a strong pattern in these functions as T_i increases.

2.2.3 Approach

Our goal is to explore the data in order to understand how patterns of SOFA scores differ among subjects with different levels of the two outcomes, in-hospital mortality and physical impairment. We are investigating regression procedures that take each subject's entire set of covariates, $\{X_i(t_{ij}), Z_i, T_i\}, t_{ij} \in [0, 1, \ldots, T_i]$, and produce a single number that is most predictive of outcome; e.g., the log odds of mortality. In particular, this procedure must be flexible enough to account for a functional covariate of varying length. Note that by conditioning on the domain width T_i, our model cannot be used to dynamically predict when the curve will terminate (e.g., when a subject will die). Instead, our focus is a retrospective analysis that explores differences in SOFA patterns and how they relate to each outcome.

For potential solutions, we incorporate ideas from the field of functional regression, which we briefly describe here. Standard functional regression models focus on the association between a scalar outcome and a functional covariate of fixed width (i.e., functional domain). Suppose that $\{Y_i\}$ are a set of scalar outcomes, $\{X_i(t)\}$ are functional covariates all defined on the interval $[0, T]$, and $\{\boldsymbol{Z}_i\}$ are non-functional covariates, where $i \in \{1, 2, \ldots, N\}$. Then the generalized functional linear model (GFLM) to relate a functional covariate to a scalar outcome is

$$g(\mu_i) = \alpha + \boldsymbol{Z}_i \boldsymbol{\gamma} + \int_0^T X_i(t)\beta(t)\, dt \tag{2.1}$$

where Y_i follows an exponential family distribution with mean μ_i, and $g(\cdot)$ is a link function.The functional parameter $\beta(t)$ represents the optimal way of weighting each $X_i(t)$ across the domain $t \in [0, T]$, to obtain the total contribution of $X_i(t)$ towards $g(\mu_i)$. $\beta(t)$ is typically constrained to be smooth across the domain t.

Model (2.1) has been studied extensively (Cardot et al., 1999; Cardot and Sarda, 2005; James, 2002; James et al., 2009; Marx and Eilers, 1999; Müller and Stadtmüller, 2005; Ramsay and Silverman, 2005; Reiss and Ogden, 2007). Incorporating non-Gaussian outcomes, producing confidence intervals, and incorporating multiple noisy and heterogeneous functional predictors has proven to be difficult. Using a penalized likelihood approach and the connection with mixed effects models, Goldsmith et al. (2011) introduced penalized functional regression (pfr), a simple fitting approach that solved these outstanding problems. The method is implemented in the namesake function `pfr()` deployed in the R (R Development Core Team, 2014) package `refund` (Crainiceanu et al., 2012).

All these fundamental contributions have only considered the case when subject-specific functions have the same fixed domain. We now propose a new model that relates variable-domain functions to a scalar outcome.

2.3 Variable-Domain Functional Regression

2.3.1 Model Specification

We propose the following model to regress a scalar outcome on a function with subject-specific domain, which we refer to as variable-domain functional regression (VDFR):

$$g(\mu_i) = \alpha + \boldsymbol{Z}_i\boldsymbol{\gamma} + \frac{1}{T_i}\int_0^{T_i} X_i(t)\beta(t, T_i)\, dt \tag{2.2}$$

The model contains two important modifications from (2.1). The first is that the bounds of integration, previously fixed to be from 0 to T, are now subject-specific. The second is to replace the univariate coefficient function $\beta(t)$ with the bivariate coefficient function $\beta(t, T_i)$. We now describe this bivariate coefficient function in more detail to provide intuition. For any fixed domain width T_0, $\beta(t, T_0)$ is a univariate function of length T_0, defined over the t-domain. This function serves as the optimal weight function for $X_i(t)$ to express its contribution towards $g(\mu_i)$, just as $\beta(t)$ did in (2.1). However, one typically would not want to assume that the weight function for a subject who remained in the ICU for 5 days, for example, would be the same as that for a subject who stayed in the ICU for 20 days. The bivariate coefficient function allows these weights to change as the width of the domain changes. We require that these weights change smoothly in both the t and T_i directions.

The VDFR model is similar in spirit to the varying-coefficient model (Hastie and Tibshirani, 1993), also referred to as a continuous-by-continuous interaction model (Ruppert et al., 2003). The interaction describes the way in which one variable (i.e. the domain width, T_i) modifies the association between the outcome and our covariate of interest $X_i(t)$. Varying-coefficient models have previously been extended to the functional regression setting by Wu et al. (2010), who allow for a coefficient function that changes with any fixed covariate Z_i. The unique feature of our model is that the fixed covariate that we interact with $X_i(t)$ is the domain width, T_i, and the integration only occurs over that domain width. Note that the term $\frac{1}{T_i}$ which appears in front of the integral sign is unnecessary, as it could easily be absorbed by the nonparametric coefficient function $\beta(t, T_i)$. Its inclusion causes the estimate of the coefficient function to have similar magnitude across different levels of T_i.

2.3.2 Estimation

The domain of the coefficient function $\beta(t, T_i)$ is $\{t, T_i : 0 \leq t \leq T_i \leq \max_i T_i\}$, which is a triangular or trapezoidal surface. Most common functional regression methods use a B-spline basis to approximate the coefficient function, but a tensor-product B-spline basis is defined over a rectangular surface and is thus not appropriate for variable-domain data. Instead, we use a thin plate regression spline basis (Wood, 2003), which adapts well to the non-rectangular regression surface covered by the data. A potential disadvantage of such a basis choice is that each basis function is symmetric in all directions (isotropic). In our scenario the two coordinates (t and T_i) have fundamentally different interpretations, and we may want to control the shape and degree of smoothness in each direction separately. Nonetheless, when a large number of basis functions are used the estimated surface can adapt quite flexibly to the data, and we have found them to work remarkably well in practice. An alternative basis choice would be the finite element basis (Braess, 2007; Brenner and Scott, 2002) that has been used to estimate the trapezoidal coefficient function of the historical functional linear model (Harezlak et al., 2007; Malfait and Ramsay, 2003). This basis was not chosen due to its increased computational complexity, though we do suggest it as an area for future research.

Basis coefficients are penalized with a second-order derivative penalty in order to ensure that estimates are visually smooth in both the t and T_i directions. We take advantage of the well-known connection between penalized likelihood and mixed models (Reiss and Ogden, 2009; Ruppert et al., 2003), which allows us to estimate the parameters of (2.2) using standard mixed model software, such as the `gam` function of the `mgcv` package in `R` (Wood, 2006). In addition to the software being readily available and well-tested, this allows us to take advantage of the inferential machinery for mixed models to obtain covariance estimates for all parameters. These estimates may then be used to obtain pointwise confidence intervals for $\beta(t, T_i)$ using standard sandwich estimators; see Goldsmith et al. (2011) for details. All model parameters are estimated using restricted maximum likelihood (REML) to simultaneously estimate both the coefficients and the smoothing parameters (Wood, 2011).

2.3.3 Computational Issues

In scalar-on-function regression, it is often common practice to subtract the overall mean function from each raw covariate function, and use the resulting de-meaned functions as predictors in the regression model (Goldsmith et al., 2011; Ramsay and Silverman, 2005). In the standard scalar-on-function regression model (2.1), doing so does not effect the model other than in the interpretation of the intercept, but it can lead to increased numerical stability in the computation. In the case of variable-domain data, the overall mean function is not clearly defined. However, we can estimate the conditional mean of $X_i(t)$ given T_i, which we denote $\mu_{X|T_i}(t)$, by fitting the generalized additive model $X_i(t) = \mu_{X|T_i}(t) + \epsilon_i(t, T_i)$, $\epsilon_i(t, T_i) \sim N(0, \sigma^2 I)$. The bivariate mean function falls on a triangular or trapezoidal surface (the same surface as the associated coefficient function $\beta(t, T_i)$, and may be fit using a thin plate regression spline basis.

Unlike in the standard scalar-on-function regression model (2.1), de-meaning the predictor functions will lead to different estimates of the bivariate coefficient function in the VDFR model (2.2). This is because de-meaning introduces an "offset" into the model that is dependent on T_i:

$$\begin{aligned}\frac{1}{T_i}\int_0^{T_i}\left\{X_i(t)-\mu(t,T_i)\right\}\beta(t,T_i)\,dt &= \frac{1}{T_i}\int_0^{T_i}X_i(t)\beta(t,T_i)\,dt-\frac{1}{T_i}\int_0^{T_i}\mu(t,T_i)\beta(t,T_i)\,dt\\ &= \frac{1}{T_i}\int_0^{T_i}X_i(t)\beta(t,T_i)\,dt-h(T_i)\end{aligned}$$

If one includes the additive term $f(T_i)$ in the model, this term would capture the offset $h(T_i)$, and de-meaning will not have an effect on the estimate of $\beta(t,T_i)$. If one does not include this term, the decision of whether to de-mean can be based on the desired interpretation of the resulting coefficient function, i.e., whether one believes that it is an individual's deviation from the mean predictor function, rather than their predictor function itself, that is most associated with the outcome. Alternatively, the decision may be data-driven, for example by comparing cross-validated prediction errors.

We also note that isotropic smoothers such as the one we employ here were designed to model surfaces for which the arguments of the smoother are measured in the same units, such as points in space. If all of the predictor functions are of similar width, we may be faced with a situation where the coordinates in the t direction span a much wider range than the coordinates in the T_i direction. For these situations, we follow the suggestion of Wood (2003) and scale the coordinates to the unit triangle (i.e., the “upper-left corner” of the unit square).

Example code for all steps of the estimation for this model (as well as the extensions proposed in the next section) is provided in the supplemental material.

2.4 Expanded Class of Variable-Domain Models

In this section we show how we can use simple change of variables and re-parameterization of some of the terms in (2.2) to expand the class of models for variable-domain functional regression. The models in this section are theoretically equivalent to (2.2). However, in practice, each model will give different results due to choice of basis set, smoothness assumptions, and the scale of the numerical approximation of the integral term. We will compare these models more thoroughly in Section 2.5.

2.4.1 Lagged Time

Let $u = t - T_i$ denote the “negative lagged” time, i.e. the time remaining until the end of one’s function, T_i, expressed as a negative number. This is the scale that one would obtain if each function was aligned according to their final measurement rather than their first, and this time was denoted as time $u = 0$. (2.2) becomes

$$\begin{aligned} g(\mu_i) &= \alpha + \boldsymbol{Z}_i\boldsymbol{\gamma} + \frac{1}{T_i}\int_{-T_i}^{0} X_i(u+T_i)\beta(u+T_i, T_i)\,du \\ &= \alpha + \boldsymbol{Z}_i\boldsymbol{\gamma} + \frac{1}{T_i}\int_{-T_i}^{0} X_i^*(u)\beta^*(u, T_i)\,du \end{aligned} \tag{2.3}$$

where $X_i^*(u) = X_i(u+T_i)$ and $\beta^*(u, T_i) = \beta(u+T_i, T_i)$, and the functions $\{X_i^*(u)\}$ fall on the domain $[-T_i, 0]$. The main advantage of this approach is that it assumes smoothness based on the lagged-time as opposed to the original time, which may be more appropriate in certain applications. For example, if $X_i(t)$ is a longitudinally-measured covariate and it is assumed that the most recent measurements will have a stronger effect than the earlier ones, then it makes more sense to impose smoothness based on the lagged time. The coefficient function still falls on a triangular or trapezoidal domain, defined by $\{u, T_i : \min_i -T_i \leq -T_i \leq u \leq 0\}$. Although this domain is the mirror image of that of $\beta(t, T_i)$ in (2.2) (projected over the T_i-axis), the functions $X_i^*(u)$ are translations, rather than reflections, of the original functions $X_i(t)$. The model may be estimated using a thin plate regression spline basis, in much the same way as we proposed to fit (2.2).

2.4.2 Domain-Standardization

In Section 1, we noted that a common approach in the functional regression literature for handling variable-domain data is to transform each function to a common domain. With the change of variable transformation $s = t/T_i$, model (2.2) becomes

$$g(\mu_i) = \alpha + \boldsymbol{Z}_i\boldsymbol{\gamma} + \int_0^1 \tilde{X}_i(s)\tilde{\beta}(s, T_i)\, ds \tag{2.4}$$

where $\tilde{X}_i(s) = X_i(sT_i)$ and $\tilde{\beta}(s, T_i) = \beta(sT_i, T_i)$. The new covariate functions $\{X_i(s)\}$ all fall on the common domain $[0, 1]$, and the new domain variable s has the interpretation of representing the proportion of the way through the function. Thus, $\tilde{X}_i(.5)$ is equal to $X_i(t_0)$ at $t_0 = T_i/2$ (i.e., halfway between 0 and T_i), and $\tilde{X}_i(1)$ is the final recorded value of $X_i(t)$.

The coefficient function $\tilde{\beta}(s, T_i)$ still allows for a weight function $\tilde{\beta}(s, \cdot)$ that changes with T_i. In fact, model (2.4) is a particular instance of a varying coefficient functional regression model (Wu et al., 2010), one for which the functional coefficient varies with T_i. The domain of the coefficient function is the rectangle $\{s, T_i : 0 \leq s \leq 1, 0 \leq T_i \leq \max_i T_i\}$, which allows us to approximate the surface with an anisotropic basis suited for a rectangular surface, such as a tensor-product basis. Since B-splines are the most common basis found in the functional regression literature (Cardot et al., 2003; Cardot and Sarda, 2005; Marx and Eilers, 1999, 2005), we apply a tensor-product B-spline basis to the surface. For comparison, we also fit the model using thin plate regression splines over the same surface.

2.4.3 Parametric Interactions

Domain standardization provides another benefit by allowing us to easily parametrize how T_i affects the weight function $\tilde{\beta}(s,\cdot)$. For example, if we assume $\tilde{\beta}(s,T_i) = \beta_1(s) + \beta_2(s)T_i$, the model becomes a linear interaction model. Letting $\tilde{X}_i(s)T_i = A_i(s)$, (2.4) becomes

$$\begin{aligned} g(\mu_i) &= \alpha + \boldsymbol{Z}_i\boldsymbol{\gamma} + \int_0^1 \tilde{X}_i(s)\left\{\beta_1(s) + \beta_2(s)T_i\right\}\,ds \\ &= \alpha + \boldsymbol{Z}_i\boldsymbol{\gamma} + \int_0^1 \tilde{X}_i(s)\beta_1(s)\,ds + \int_0^1 A_i(s)\beta_2(s)\,ds \end{aligned} \tag{2.5}$$

Thus, restricting $\beta(s,T_i)$ to be linear in T_i reduces the problem to a standard scalar-on-function regression model with two functional predictors, $\tilde{X}_i(s)$ and $A_i(s)$. Similarly, if we assume $\tilde{\beta}(s,T_i) = \beta_1(s) + \beta_2(s)T_i + \beta_3(s)T_i^2$, we obtain a quadratic interaction model. Technically, there is little difference between the linear and quadratic interaction models. Indeed, if $\tilde{X}_i(s)T_i^2 = B_i(s)$ then the model becomes

$$\begin{aligned} g(\mu_i) &= \alpha + \boldsymbol{Z}_i\boldsymbol{\gamma} + \int_0^1 \tilde{X}_i(s)\left\{\beta_1(s) + \beta_2(s)T_i + \beta_3(s)T_i^2\right\}\,ds \\ &= \alpha + \boldsymbol{Z}_i\boldsymbol{\gamma} + \int_0^1 \tilde{X}_i(s)\beta_1(s)\,ds + \int_0^1 A_i(s)\beta_2(s)\,ds + \int_0^1 B_i(s)\beta_3(s)\,ds \end{aligned} \tag{2.6}$$

which is a standard scalar-on-function regression model with three functional predictors. Alternatively, we can make the stricter assumption that $\tilde{\beta}(s,T_i)$ does not change with T_i, that is $\tilde{\beta}(s,T_i) = \beta(s)$. The resulting model, which is not an interaction model at all, is equivalent to the standard functional regression model (2.1) using the domain-standardized predictor functions.

Any of the models presented in this section may be fit using existing functional regression software that accepts multiple functional predictors, such as `pfr` (Goldsmith et al., 2011). To maintain consistency across all models, our implementation does not use existing functional regression software, but instead calls `mgcv::gam` directly, as was done for the non-parametric models discussed previously. As all models are fit using mixed model software, pointwise confidence intervals are available. We use a penalized B-spline basis to approximate the univariate coefficient functions in (2.1), (2.5), and (2.6) above.

2.5 Simulation Studies

2.5.1 Simulation Design

We now investigate the performance of these models via simulations, under a variety of true coefficient functions $\beta(t, T_i)$. For simplicity, we consider the scenario where there are no non-functional covariates Z, and only a single functional predictor $X(t)$. We consider every combination of the following simulation parameters, resulting in $3 \times 2 \times 2 \times 4 \times 2 = 96$ total scenarios:

1. Three choices for the sample size, N: 100, 200, and 500

2. Two distributions for T_i: uniform or right-skewed
3. Four different possibilities for the true coefficient function, $\beta(t, T_i)$, defined below
4. Two different types of outcomes: continuous and binary. We fit gaussian models to the continuous outcomes, and logistic models to the binary outcomes
5. Two choices for measurement error in $X(t)$: none vs. some.

For notational convenience, we will assume that our functions are observed at whole-numbered time points, $\{t_j = 0, 1, \ldots, J = 100\}$. For each value of N, we generate $R = 1000$ datasets of functional covariates according to the following model, which is adapted from Goldsmith et al. (2011):

$$\begin{aligned} W_i(t_j) &= X_i(t_j) + \delta_i(t_j) \\ X_i(t_j) &= u_i + \sum_{k=1}^{10} \left\{ v_{ik1} \sin\left(\frac{2\pi k}{100} t_j\right) + v_{ik2} \cos\left(\frac{2\pi k}{100} t_j\right) \right\} \end{aligned}$$

where $\delta_i(t_j) \sim N(0, \sigma_X^2)$, $u_i \sim N(0, 1)$, and $v_{ik1}, v_{ik2} \sim N(0, 4/k^2)$. In this notation, $\{X_i(\cdot)\}$ are the true underlying functions, whereas $\{W_i(\cdot)\}$ are the observed functions. We consider $\sigma_X^2 \in \{0, 1\}$, corresponding to no measurement error and some measurement error.

The domain width T_i is generated for each function independently, either from a Uniform$(0, 100)$ distribution, or from a NegBin$(1, p = 0.04)$ distribution that is truncated at a maximum T_i of 100. The latter distribution is right-skewed so as to produce more values of T_i that are small, such as those we observe in the ICAP data. Each function $X_i(t_j)$ and $W_i(t_j)$ is then truncated to only allow for $t_j \leq T_i$.

We generate both continuous and binary outcomes for each dataset of functional covariates, based on the model $\eta_i = \frac{1}{T_i}\sum_{t_j=0}^{T_i} \beta_b(t_j, T_i)X_i(t_j)$, where b indexes the particular true coefficient function used. The continuous outcomes are simulated as $Y_i = \eta_i + \epsilon_i$, $\epsilon_i \sim N(0,1)$, whereas the binary outcomes are simulated from a Bernoulli(p_i) distribution, $p_i = \exp(\eta_i)/\left(1+\exp(\eta_i)\right)$. Four possible bivariate coefficient functions $\beta_b(t, T_i)$ are considered:

$$\begin{aligned}
\beta_1(t,T_i) &= 10\frac{t}{T_i} - 5 \\
\beta_2(t,T_i) &= \left(1 - \frac{2T_i}{J}\right) \times \left(5 - 40\left(\frac{t}{T_i} - 0.5\right)^2\right) \\
\beta_3(t,T_i) &= 5 - 10\left(\frac{T_i - t}{J}\right) \\
\beta_4(t,T_i) &= \sin\left(\frac{2\pi T_i}{J}\right) \times \left(5 - 10\left(\frac{T_i - t}{J}\right)\right)
\end{aligned}$$

These coefficient functions all fall within the range $[-5, 5]$, similar to what we will observe in our application, and are plotted as heat maps in Figure 2.2. The coefficient functions are meant to reflect one of two realistic scenarios. The first of these scenarios corresponds to a situation in which the relative position (t/T_i) within the function drives the association between $X_i(t)$ and $g(\mu_i)$. This scenario is reflected in $\beta_1(t,T_i)$ and $\beta_2(t,T_i)$. The remaining two coefficient functions reflect a scenario in which the lag, $T_i - t$, drives the strength of the association.

We fit seven versions of the VDFR model to each simulated data set. The first one, which uses the untransformed predictor functions as described in Section 2.3, will be referred to as the "Untransformed" model. The second uses the lagged predictor functions as described in Section 2.4.1, and is referred to accordingly as the "Lagged" model. The remaining five models use the domain-standardized predictor functions. The first two allow for the interaction with T_i to be non-parametric (Section 2.4.2), either with a thin plate regression spline basis or a tensor-product B-spline basis. We refer to them as "DS (TPRS)" and "DS (TPBS)", respectively. The final three models parametrize the interaction as described in Section 2.4.3. We refer to the models with no interaction, linear interaction, and quadratic interaction as "DS (No Int)", "DS (Lin)", and "DS (Quad)", respectively.

In addition to the seven functional models, we fit 14 "non-functional" models to the simulated data (Table 2.2). The first five of these models were simple linear or logistic regressions of the outcome against a single summary statistic of each subject's predictor function. The remaining nine were more complicated parametric or semi-parametric models involving the within-subject mean $\bar{X}_i$ and the domain width T_i. Variables and parameterizations of each model are listed in Table 2.2. All smooth functions are modeled using a thin-plate regression spline basis.

2.5.2 Evaluation criteria

Performance of models was evaluated in two ways. First, we measure the ability of each model to predict the outcome in each scenario. For the models fit to the continuous outcomes, this is measured through the cross-validated root mean squared error (rMSE), whereas for the binary outcome models, we calculated the cross-validated area under the receiver operating characteristic curve (AUC). All cross-validation is 10-fold. Good predictive accuracy is indicated by low rMSE or high AUC.

For the seven models that produce functional estimates, we also measure the ability of each model to estimate the true coefficient function. This ability is evaluated using the average mean squared error (AMSE) of the estimate over all possible values of t_j and T_i. More precisely,

$$\text{AMSE}^{(r)}\left(\hat{\beta}_b(\cdot,\cdot)\right) = \frac{1}{J(J+1)} \sum_{k=0}^{J} \sum_{j=0}^{k} \left\{\hat{\beta}_b^{(r)}(t_j, k) - \beta_b(t_j, k)\right\}^2,$$

where $\hat{\beta}_b^{(r)}(t_j, k)$ is the estimated coefficient function from the r^{th} simulated dataset evaluated at $t = t_j$, $T_i = k$, and $\beta_b(t_j, k)$ is the value of the true coefficient function at this location. Before this calculation is performed, all estimates (other than the one from the Untransformed model) are converted back to the original (triangular) domain. For the Lagged model, this is a simple translation of the estimates. For the five models fit using the domain-standardized predictor functions over a rectangular grid, we stratify estimates for each T_i into $T_i + 1$ bins, and calculate the mean value in each bin.

2.5.3 Simulation results

The median cross-validated AUC statistics for the case when $N = 200$, T_i is chosen from a skewed distribution, measurement error is present, and the outcome is binary are presented in Table 2.2. This scenario is presented because it is most similar to our application; results from other scenarios appear in the supplemental material. For the presented scenario and in nearly every other scenario, the model with the lowest median cross-validated mean squared error (continuous outcome) or highest median cross-validated AUC (binary outcome) was one of the seven functional models. In many cases, the performance of the non-functional models was extremely poor, resulting in cross-validated AUC statistics around 0.5 or even below, indicating that the given model does not help predict the outcome at all. The only cases where the best model was one of the non-functional models occurred with the smallest sample size, binary outcome, and skewed distribution for T_i, under $\beta_3(t, T_i)$. In these cases there were a number of models that all predicted outcome very well, and the best-performing model happened to be one of the non-functional ones.

The results for the seven functional models under the above scenario are presented more fully in Figures 2.2 and 2.3. The former presents the rAMSE and cross-validated AUC values across the 1000 iterations as box plots, whereas the latter depicts the estimated coefficient function for the estimate with median AMSE among the 1000

	$\beta_1(t,T_i)$	$\beta_2(t,T_i)$	$\beta_3(t,T_i)$	$\beta_4(t,T_i)$
Functional Models:				
Untransformed	0.862	0.744	0.947	0.886
Lagged	0.863	0.746	0.946	0.887
DS (TPRS)	0.902	0.778	0.963	0.904
DS (TPBS)	0.900	0.775	0.933	0.905
DS (No Int)	0.905	0.602	0.948	0.790
DS (Lin)	0.901	0.716	0.964	0.866
DS (Quad)	0.897	0.767	0.959	0.872
Summary Statistic Models:				
Mean	0.434	0.634	0.955	0.854
Median	0.436	0.637	0.949	0.847
Maximum	0.439	0.601	0.842	0.776
Cumulative	0.431	0.646	0.915	0.817
Slope	0.820	0.429	0.438	0.431
Additional Models:				
$\beta_1\bar{X}_i + \beta_2 T_i$	0.459	0.631	0.954	0.850
$\beta_1\bar{X}_i + f_2(T_i)$	0.466	0.628	0.951	0.846
$f_1(\bar{X}_i) + \beta_2 T_i$	0.466	0.621	0.953	0.845
$f_1(\bar{X}_i) + f_2(T_i)$	0.472	0.620	0.949	0.842
$f(\bar{X}_i, T_i)$	0.465	0.720	0.938	0.865
$\beta_1\bar{X}_i + \beta_2 T_i + \beta_3\bar{X}_i T_i$	0.458	0.636	0.958	0.864
$\beta_1\bar{X}_i + f_2(T_i) + f_3(\bar{X}_i T_i)$	0.468	0.647	0.953	0.856
$f_1(\bar{X}_i) + \beta_2 T_i + f_3(\bar{X}_i T_i)$	0.472	0.650	0.955	0.857
$f_1(\bar{X}_i) + f_2(T_i) + f_3(\bar{X}_i T_i)$	0.475	0.649	0.952	0.852

Table 2.2: Median cross-validated AUC for all models applied to the simulated data, for the case when N=200, T_i is skewed, measurement error is present, and the outcome is binary. Models include seven functional models, five simple logistic regressions on the indicated summary statistic, and nine more complicated parametric or semi-parametric functions of the within-subject mean ($\bar{X}_i$) and the domain width (T_i).

iterations, as a heat map. The four nonparametric models seem to perform well regardless of the true coefficient function. For $\beta_1(t, T_i)$ and $\beta_2(t, T_i)$, the domain-standardized models tend to perform better than the Untransformed and Lagged models. The DS (TPRS) and DS (TPBS) models tend to perform similarly for all coefficient functions other than $\beta_3(t, T_i)$, where there appears to be much more variability in the performance of the DS (TPBS) model. The reasons for this are not clear, and this effect does not occur when the outcome is gaussian (supplemental material). The Untransformed and Lagged models tend to perform similarly for all four coefficient functions, including $\beta_3(t, T_i)$ and $\beta_4(t, T_i)$, which were designed to be lag-based.

The parametric functional models are among the best-performing models in the cases when the interaction with T_i is simple enough to be accounted for by the parametric assumption. For example, all three parametric models perform well under $\beta_1(t, T_i)$, which contains no interaction on the domain-standardized scale. However, the DS (No Int) model cannot account for the linear interactions that are present in the other three coefficient functions, resulting in an estimate that is quite homogenous in both the t and T_i directions. The three parametric models are outperformed by the domain-standardized nonparametric models for both $\beta_2(t, T_i)$ and $\beta_4(t, T_i)$, which contain more complicated interactions. The quadratic model especially produces estimates that are quite unstable in the region with high T_i. Recalling that these estimates are from the scenario where T_i is chosen from a right-skewed distribu-

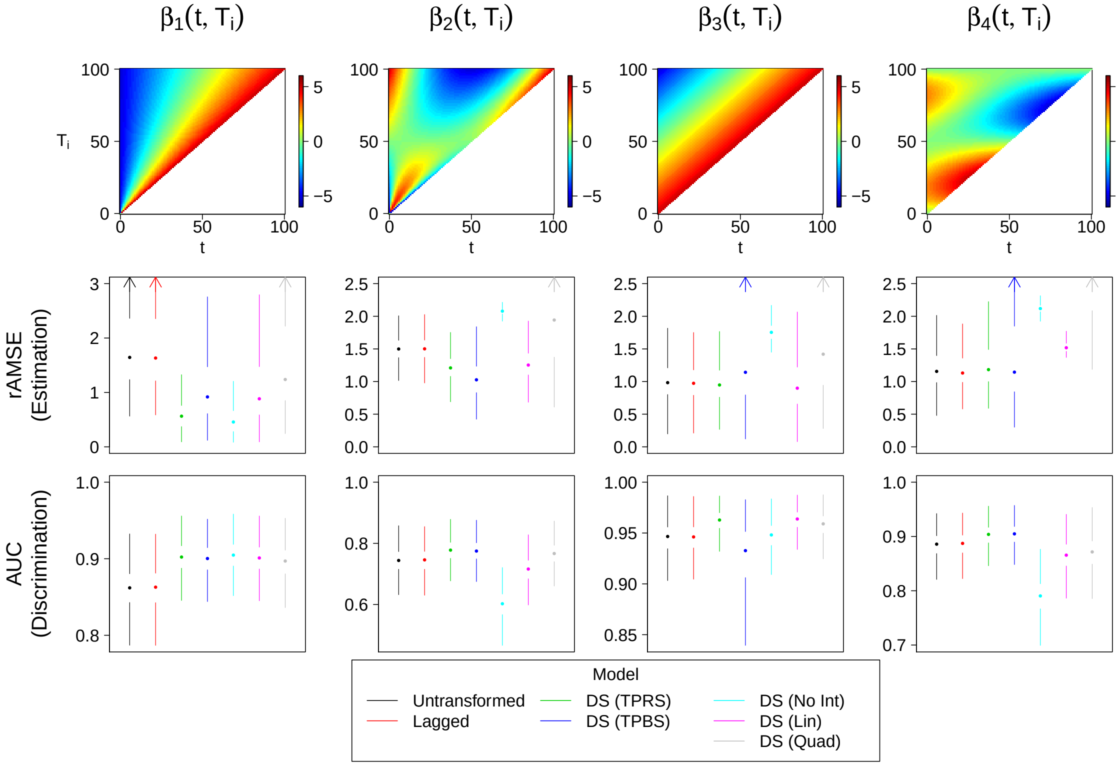


Figure 2.2: Simulation results for the case when $N = 200$, T_i is skewed, measurement error is present, and the outcome is binary. The top row depicts a heat map of the true coefficient functions. The second and third rows depict the root average mean squared error (rAMSE) of $\hat{\beta}(t, T_i)$ and 10-fold cross-validated area under the ROC curve (AUC), respectively, for each of the seven models. Smaller rAMSE and larger AUC indicate better model performance. Results are presented as Tufte box plots, with the median represented by a dot, the interquartile range by the white space around the dot, and the smallest and largest non-outlying points by the endpoints of the lines. Outliers are defined as values not within 1.5 times the interquartile range of the nearest quartile. Arrows indicate lines that extend outside the plotting range.

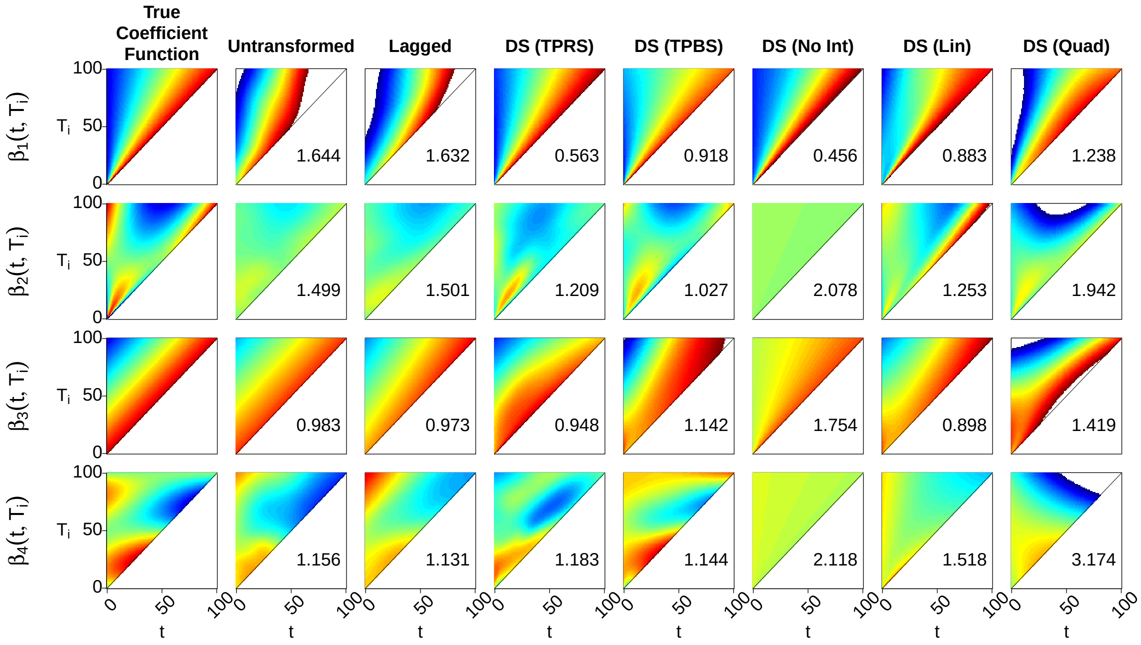


Figure 2.3: Heat maps of the estimate with median AMSE across the 1000 simulated datasets, for each model and coefficient function, in the case where $N = 200$, T_i is skewed, measurement error is included, and the outcome is binary. The range for each plot is from -6 (blue) to 6 (red), with values outside of this range indicated by white space. The numbers in the lower-right corner of each plotting area are the rAMSE statistics for each estimate.

tion, this observation reflects the instability of extrapolating higher-order polynomial functions to regions outside the bulk of the data.

Tables and plots of the rAMSE, rMSE, and AUC under other scenarios (different sample sizes, distribution of T_i, amount of measurement error, and continuous outcomes) are available in the supplemental material. In general, we found the results discussed above to hold true under these scenarios as well. As expected, both estimation and prediction error tend to be lower as the sample size increases, when the distribution of T_i is uniform, and when measurement error is not present. rAMSE values tended to be much lower when the outcome was continuous as opposed to binary. They were also much less variable, resulting in more clear differentiation between models. Overall patterns of comparative model performance were similar to the binary case.

2.6 Application to ICAP Data

2.6.1 Model Specification

For both binary outcomes in the ICAP data, we fit each of the seven VDFR models discussed in the preceding sections. Each model includes SOFA as a functional covariate and controls for age, gender, Charlson comorbidity index (a commonly used index of baseline health, (Charlson et al., 1987)), and the log of the ICU length of stay T_i, as fixed (non-functional) effects.

We investigated whether de-meaning the functional predictors and/or modeling $\log(T_i)$ as a smooth term rather than a linear term improved model performance. Based on cross-validated AUC statistics, we found the optimal performance for the mortality outcome occurred when the SOFA functions were not de-meaned and when $\log(T_i)$ was modeled smoothly. For physical function, highest cross-validated AUC scores occurred without de-meaning when $\log(T_i)$ was included as a linear term. It is these results that we present below. Additionally, since the domain width T_i is highly right-skewed, we investigated whether the interaction in the functional models should occur with a transformation of this variable. In other words, instead of estimating the coefficient function $\beta(t, T_i)$, we estimate $\beta^*(t, w(T_i))$, where $w(\cdot)$ is a monotonic transformation function. The two $w(\cdot)$ functions considered were the log function, and the empirical quantiles of T_i. In addition to evening out the amount of data over the estimated surface on these scales, this approach assumes that the true interaction takes place on the transformed scale. Since the w functions are monotonic, the resulting models are theoretically equivalent, but in practice may be different due to the choice of basis and level of smoothness. However, we found that differences between the three methods were quite small in the ICAP dataset. Since the lengths of stay are approximately log-normally distributed, we present the results using the log-transformed T_i.

In addition to the seven functional models, we fit 13 non-functional models to the data, similar to those fit to the simulated data. All models are adjusted for the same covariates (age, gender, Charlson index, and $\log(T_i)$) as the functional models.

2.6.2 Model Performance

AUC statistics for each model are presented in Table 2.3, both in-sample and under cross-validation, with 95% confidence intervals based on 1000 bootstrapped samples. The in-sample statistic measures the discriminative ability of each model on the existing dataset, whereas the cross-validated statistic estimates discriminative ability for a new sample, and is more relevant for model selection as it is less prone

Table 2.3: AUC statistics for each model, applied to the binary outcomes of in-hospital mortality and physical function in the ICAP dataset. Results are presented as ${}_{L}X_{R}$, where X is the estimate, and (L, R) are the lower and upper bounds, respectively, of a 95% confidence interval based on 1000 bootstrapped samples. Both in-sample and cross-validated AUC statistics are presented. Cross validation is N-fold for the estimates, and 10-fold for the bootstrapped confidence intervals. All models are adjusted for age, gender, Charlson comorbidity index, and the log length of stay. Log length of stay is included as a smooth term for mortality and as a linear term for physical impairment.

	Mortality		**Physical Impairment**	
	In-sample	Cross-validated	In-sample	Cross-validated
Functional Models:				
Untransformed	${}_{0.940}0.948_{0.977}$	${}_{0.913}0.919_{0.955}$	${}_{0.813}0.838_{0.911}$	${}_{0.756}0.784_{0.871}$
Lagged	${}_{0.941}0.947_{0.977}$	${}_{0.911}0.918_{0.955}$	${}_{0.812}0.836_{0.911}$	${}_{0.760}0.790_{0.875}$
DS (TPRS)	${}_{0.942}0.949_{0.983}$	${}_{0.922}0.933_{0.960}$	${}_{0.825}0.847_{0.943}$	${}_{0.766}0.790_{0.888}$
DS (TPBS)	${}_{0.942}0.950_{0.981}$	${}_{0.920}0.936_{0.961}$	${}_{0.811}0.829_{0.945}$	${}_{0.750}0.784_{0.883}$
DS (No Interaction)	${}_{0.934}0.946_{0.971}$	${}_{0.916}0.934_{0.957}$	${}_{0.767}0.826_{0.886}$	${}_{0.727}0.797_{0.860}$
DS (Linear)	${}_{0.937}0.947_{0.973}$	${}_{0.914}0.933_{0.957}$	${}_{0.773}0.831_{0.902}$	${}_{0.725}0.794_{0.862}$
DS (Quadratic)	${}_{0.943}0.950_{0.977}$	${}_{0.919}0.935_{0.959}$	${}_{0.789}0.830_{0.907}$	${}_{0.735}0.788_{0.863}$
Summary Statistic Models:				
Mean	${}_{0.877}0.899_{0.927}$	${}_{0.870}0.893_{0.922}$	${}_{0.755}0.820_{0.870}$	${}_{0.719}0.798_{0.854}$
Median	${}_{0.876}0.898_{0.929}$	${}_{0.868}0.891_{0.923}$	${}_{0.752}0.820_{0.872}$	${}_{0.716}0.798_{0.852}$
Maximum	${}_{0.860}0.885_{0.913}$	${}_{0.853}0.879_{0.908}$	${}_{0.757}0.819_{0.874}$	${}_{0.728}0.799_{0.858}$
Cumulative	${}_{0.805}0.842_{0.876}$	${}_{0.793}0.833_{0.868}$	${}_{0.760}0.818_{0.872}$	${}_{0.731}0.797_{0.854}$
Slope	${}_{0.744}0.795_{0.856}$	${}_{0.728}0.785_{0.847}$	${}_{0.750}0.810_{0.868}$	${}_{0.723}0.790_{0.851}$
Additional Models:				
$\beta_1\bar{X}_i + f_2(T_i)$	${}_{0.880}0.901_{0.930}$	${}_{0.871}0.892_{0.922}$	${}_{0.796}0.821_{0.892}$	${}_{0.747}0.792_{0.862}$
$f_1(\bar{X}_i) + \beta_2 T_i$	${}_{0.877}0.899_{0.927}$	${}_{0.868}0.893_{0.920}$	${}_{0.767}0.820_{0.879}$	${}_{0.713}0.796_{0.852}$
$f_1(\bar{X}_i) + f_2(T_i)$	${}_{0.880}0.901_{0.930}$	${}_{0.869}0.892_{0.920}$	${}_{0.797}0.821_{0.895}$	${}_{0.740}0.790_{0.862}$
$f(\bar{X}_i, T_i)$	${}_{0.878}0.899_{0.933}$	${}_{0.865}0.893_{0.921}$	${}_{0.767}0.820_{0.917}$	${}_{0.727}0.798_{0.865}$
$\beta_1\bar{X}_i + \beta_2 T_i + \beta_3\bar{X}_i T_i$	${}_{0.876}0.900_{0.927}$	${}_{0.869}0.893_{0.922}$	${}_{0.753}0.818_{0.871}$	${}_{0.717}0.795_{0.851}$
$\beta_1\bar{X}_i + f_2(T_i) + f_3(\bar{X}_i T_i)$	${}_{0.882}0.901_{0.933}$	${}_{0.869}0.891_{0.921}$	${}_{0.802}0.820_{0.905}$	${}_{0.745}0.781_{0.858}$
$f_1(\bar{X}_i) + \beta_2 T_i + f_3(\bar{X}_i T_i)$	${}_{0.880}0.901_{0.930}$	${}_{0.869}0.891_{0.922}$	${}_{0.773}0.820_{0.899}$	${}_{0.736}0.791_{0.859}$
$f_1(\bar{X}_i) + f_2(T_i) + f_3(\bar{X}_i T_i)$	${}_{0.884}0.901_{0.934}$	${}_{0.868}0.891_{0.921}$	${}_{0.809}0.820_{0.915}$	${}_{0.745}0.781_{0.858}$

to over-fitting. The cross-validated statistics are based on leave-one-out (i.e., N-fold) cross-validation, but the confidence intervals are based on 10-fold cross-validation to reduce the computation time. This will likely produce slightly wider confidence intervals than if we had used N-fold cross-validation.

Comparing the seven functional models for the mortality outcome, we see that the five models that used the domain-standardized functions all performed quite similarly by both metrics. The Untransformed and Lagged models both performed slightly worse under cross-validation. Each of the functional models resulted in higher AUC statistics than the non-functional ones, both in-sample and cross-validated. While the absolute differences in AUC between the functional and non-functional models may not be very large, we note that a perfectly discriminating model would have an AUC of 1. Thus, the difference between each AUC and 1 offers a measure of the "imperfection" of each model. From this perspective, the best-performing functional model (Domain-standardized B-splines) offers an improvement over the best-performing summary statistic model (mean SOFA) of 40% in terms of cross-validated AUC.

For physical function, AUC statistics were quite a bit lower than those for the mortality outcome, reflecting the weaker association between SOFA patterns and impaired physical function. Additionally, we do not see the same benefit in using a functional modeling approach. Although the functional models result in higher in-sample AUC statistics than the non-functional models, this does not hold under cross-validation. This result indicates that the functional nature of the SOFA curves is not a strong predictor of impaired physical function. The functional model with no interaction performs quite well under cross-validation in both of the two outcome scenarios, indicating that there is not much benefit in terms of discriminative ability to allowing $\beta(t, T_i)$ to change with T_i.

2.6.3 Estimated Coefficient Functions

The estimates for the coefficient functions for mortality and physical function are presented in Figures 2.4 and 2.5, respectively. Rather than presenting the triangular surface $\hat{\beta}(t, T_i)$ estimated by each model as a heat map, we present the univariate weight functions $\hat{\beta}(t, T_0)$ for 10 different values of T_0 spread evenly across the domain of T_i. The top row in these figures displays these estimates, with T_0 indicated by color as well as the support along the t-axis, and the bottom row of plots displays the corresponding pointwise Z-scores, $\hat{\beta}(t, T_0)/SE(\hat{\beta}(t, T_0))$.

For mortality, we see a consistent pattern among all models of a strong, positive spike in the association between death and high SOFA scores at the end of one's ICU stay, regardless of T_i. In most cases, the pointwise associations in these regions is statistically significant, according to a Wald test with $\alpha = 0.05$. This pattern is expected: subjects with higher SOFA scores (i.e., more severe organ failure) right before the end of their ICU stay are likely to have their ICU stay end in death, rather than be discharged alive. Moreover, increasing SOFA scores have been associated with withdrawal of life support, leading to subsequent mortality (Turnbull et al., 2014).

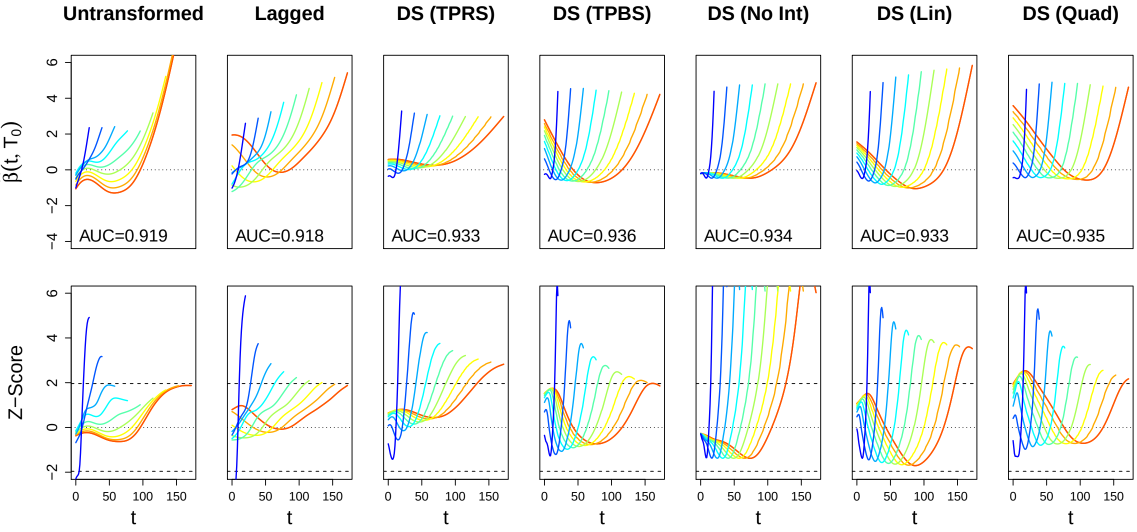


Figure 2.4: Estimated coefficient functions for the association between daily SOFA score and in-hospital mortality in the ICAP dataset. Each column corresponds to one of our six functional models. In the top row of plots, estimates are depicted as $\hat{\beta}(t, T_0)$ for 10 evenly-spaced values of T_0. AUC statistics subject to 10-fold cross-validation are also provided. The bottom row displays the corresponding pointwise Z-scores, $\hat{\beta}(t, T_0)/SE(\hat{\beta}(t, T_0))$, as a function of t. The value of T_0 is indicated both by color and by the support of each function. The zero line is indicated with a horizontal dotted line, and dashed lines correspond to Z-scores of ±1.96.

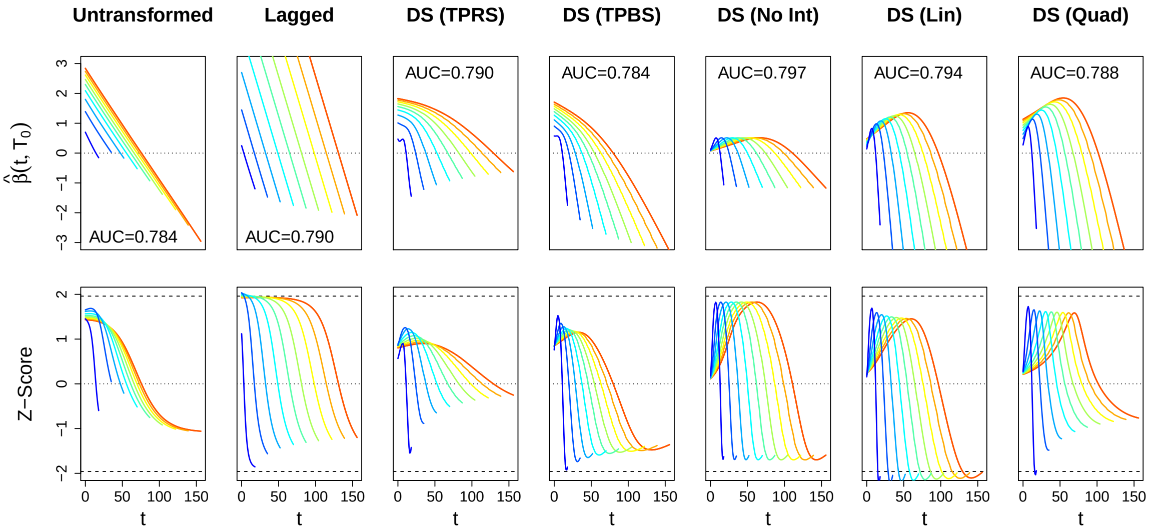


Figure 2.5: Estimated coefficient functions for the association between daily SOFA score and impaired physical function in the ICAP dataset, presented similarly to Figure 2.4.

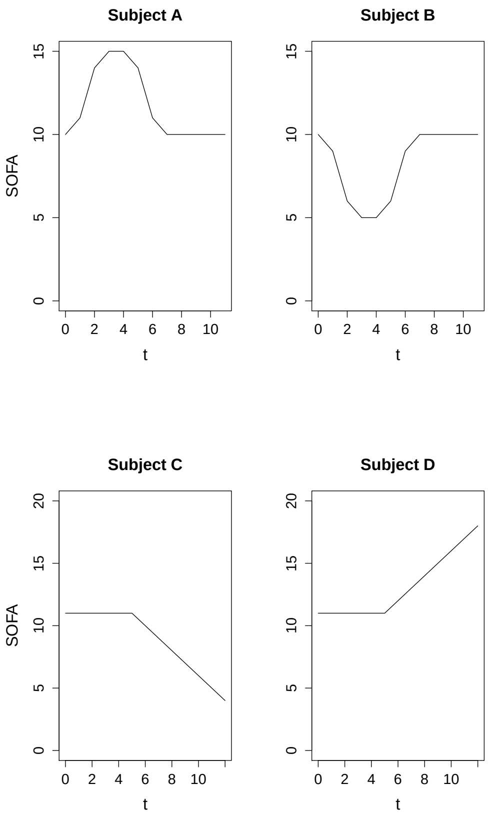
The linear and quadratic models show a tendency for this spike to move later into one's ICU stay when T_i is long. This pattern suggest that the last few days in the ICU are most important for predicting mortality, regardless of T_i. The models that allow for a more flexible interaction with T_i also estimate a positive association between early SOFA scores and mortality for subjects with long lengths of stay, resulting in "U-shaped" weight functions. Although there may be some effect on mortality related to the severity of the event that caused the onset of ALI/ARDS, we note that there are very few subjects that have these high lengths of stay (only seven subjects with $T_i > 75$), and this effect may be spurious. This hypothesis is supported by the fact that the pointwise associations in these regions are not statistically significant.

For physical function, one might be tempted to ignore the coefficient estimates from the functional models, which had lower cross-validated AUC statistics than some of the simpler, parametric models. However, these estimates may still be revealing, as they are able to estimate types of associations that are not possible to be estimated by traditional approaches, and still may identify important trends in the data. For this outcome, we find that the functional association decreases over one's ICU stay, quite linearly in each case. However, the magnitude of these associations is relatively small, and the pointwise 95% confidence intervals cover 0 in every region of all models, except for some very small locations in the Untransformed and DS (TPRS) models. This lack of a strong association reflects our observations from Figure 2.1 and Table 2.3, each of which showed weak functional relationships between SOFA and physical impairment.

For both mortality and physical impairment, there are certain features in the estimates that were somewhat unexpected, and these features have fundamental implications on the interpretation of the coefficient functions. For mortality, we were initially surprised that there were regions of each estimate that lied below the zero line. This means that, for two subjects with the same SOFA scores during the regions where the coefficient function is positive, the model predicts that the one with lower scores (i.e., the healthier patient) in the region with negative coefficient function is more likely to die. Similarly, even though the estimates for physical impairment were not statistically significant (in a pointwise sense), we were surprised the the predominant trend was for the weight functions to decrease over one's ICU stay. According to these models, a subject whose condition gradually deteriorates throughout their hospitalization will be less likely to have impaired physical function upon hospital discharge than a subject who improves.

To further illustrate these points, consider the hypothetical SOFA curves plotted in Figure 2.6. Subjects A and B have the same SOFA scores during the latter portion of their hospitalization, but Subject A experiences a temporary spike in his SOFA scores during the middle of his hospitalization, whereas Subject B experiences a temporary drop during this same time period. Both subjects were assigned the same values for their non-functional covariates (age, gender, Charlson index, and length of stay). According to each of the functional models, the subject who experienced the lower scores is more likely to die than the subject with higher scores. High SOFA scores, in the early and middle portion of one's ICU stay, appear to be associated with higher likelihood of survival. Similarly, in the lower plots both subjects had the same SOFA scores for the first 7 days of their ICU stay, but then Subject C's condition improved over the final 6 days whereas Subject D's health declined. However, the functional models predict that the subject whose health improved is more likely to leave the ICU with impaired physical function than the one whose condition deteriorated. The lone exception is the Lagged model, whose estimate has a shorter period where it is negative compared to the other models.

At first glance these observations may appear counter-intuitive, but each may be explained in the context of the full model. Since subjects similar to Subject A contain peaks in their SOFA scores early in their ICU stay, it means that they survived a



Estimated Probability of Mortality

	Subject A	Subject B
Functional Models:		
Untransformed	$_{0.45}0.72_{0.89}$	$_{0.66}0.85_{0.94}$
Lagged	$_{0.49}0.75_{0.90}$	$_{0.65}0.83_{0.93}$
DS (TPRS)	$_{0.34}0.60_{0.81}$	$_{0.74}0.89_{0.96}$
DS (TPBS)	$_{0.37}0.61_{0.81}$	$_{0.69}0.86_{0.95}$
DS (No Int)	$_{0.43}0.65_{0.82}$	$_{0.70}0.86_{0.94}$
DS (Lin)	$_{0.43}0.64_{0.80}$	$_{0.74}0.88_{0.95}$
DS (Quad)	$_{0.36}0.59_{0.78}$	$_{0.76}0.90_{0.96}$
Summary Statistic Models:		
Mean	$_{0.72}0.81_{0.87}$	$_{0.36}0.45_{0.53}$
Median	$_{0.61}0.70_{0.78}$	$_{0.50}0.59_{0.67}$
Maximum	$_{0.60}0.68_{0.76}$	$_{0.19}0.25_{0.32}$
Cumulative	$_{0.40}0.47_{0.54}$	$_{0.28}0.34_{0.41}$
Slope	$_{0.34}0.40_{0.47}$	$_{0.47}0.55_{0.63}$

Estimated Probability of Impaired PF

	Subject C	Subject D
Functional Models:		
Untransformed	$_{0.49}0.76_{0.91}$	$_{0.21}0.72_{0.96}$
Lagged	$_{0.52}0.76_{0.90}$	$_{0.12}0.57_{0.93}$
DS (TPRS)	$_{0.58}0.82_{0.94}$	$_{0.05}0.42_{0.92}$
DS (TPBS)	$_{0.53}0.72_{0.86}$	$_{0.05}0.32_{0.81}$
DS (No Int)	$_{0.57}0.75_{0.87}$	$_{0.06}0.35_{0.82}$
DS (Lin)	$_{0.56}0.74_{0.87}$	$_{0.05}0.31_{0.80}$
DS (Quad)	$_{0.50}0.71_{0.86}$	$_{0.02}0.24_{0.80}$
Summary Statistic Models:		
Mean	$_{0.50}0.66_{0.79}$	$_{0.50}0.77_{0.92}$
Median	$_{0.50}0.68_{0.82}$	$_{0.50}0.73_{0.88}$
Maximum	$_{0.49}0.61_{0.72}$	$_{0.56}0.79_{0.91}$
Cumulative	$_{0.48}0.62_{0.74}$	$_{0.48}0.72_{0.87}$
Slope	$_{0.48}0.59_{0.70}$	$_{0.16}0.36_{0.61}$

Figure 2.6: Trajectories of four hypothetical patients, along with their predicted probabilities of outcome according to each model. Top row corresponds to the mortality outcome, and bottom corresponds to impaired physical function.

serious episode that caused a temporary peak (worsening of health); if they did not, then their SOFA function would end at this point. Thus, subjects whose SOFA scores peak early in their ICU stay might have greater baseline physiological reserve than those who did not experience this peak, as demonstrated by their ability to survive this severe organ failure. Moreover, for two subjects who have the same SOFA pattern towards the end of their ICU stay, the one whose scores peaked earlier during their hospitalization must have experienced improvement since then. Thus, although Subject A was quite sick in the middle of his ICU stay, the trend in the latter half of hospitalization indicates relative improving health. In contrast, Subject B may have improved early on, but since that point their condition declined. The negative estimated coefficient function early in one's ICU stay captures this effect, whereas the non-functional models do not.

The high predicted probabilities of physical impairment for Subject C relative to D may be explained by recognizing that the subjects eligible for this outcome not only survived their ICU stay, but were also deemed healthy enough to be discharged from the ICU. If a subject has experienced a rapid improvement in their physiological metrics, a physician may be more likely to allow the subject to leave the ICU, even if he is still has some physical limitations as measured by the ADL scale. Conversely, a subject who still has some evidence of poor organ function will likely only be discharged from the ICU if he has demonstrated tremendous improvement in his outward appearance, such as proving to be unrestricted in daily activities. Though it is possible that these effects are only spurious, as these models were outperformed by some of the non-functional ones, we feel these trends are noteworthy and should perhaps be investigated in future studies.

2.7 Discussion

In this paper, we investigated methods to capture the effect of a functional predictor, where the domain of this predictor may vary widely from subject-to-subject. Such a situation is most commonly encountered when the domain variable is time, and each subject (or unit) is measured for a different length of time, as in our application. This investigation motivated our development of the variable-domain functional regression model (VDFR), which estimates a weight function to capture the effect of a functional predictor, but allows this weight function to vary (smoothly) based on the total follow-up time for each subject.

The VDFR models were able to identify features of the association between a longitudinally collected covariate and an outcome that traditional multivariate regression methods are not equipped to handle. In the analysis of ICAP mortality, we saw specifically how the functional models incorporate information related to both the magnitude of one's SOFA score and their trajectory over time to provide better discriminative ability than naive (non-functional) approaches. They also allow us to ask previously unanswerable questions, such as whether or not it is optimal to treat a longitudinal covariate with subject-specific domain as a function, and how the domain width affects the covariate-outcome relationship. Although we were not able to identify any evidence for a strong functional relationship between SOFA score and physical function at hospital discharge, without these methods we would not have known how to answer such a question.

It is important to recognize that the models that we fit are not causal models, and we do not employ them to try to identify a causal relationship between the covariate function and outcome. For example, we identified a pattern in the data that increasing SOFA scores towards the end of one's hospitalization, which indicate a decline in one's overall health, are associated with a higher likelihood that a surviving subject has limitations in their activities of daily living. We do not believe that organ failure causes a subject to have improved physical function; such a claim would run contrary to logic. One must take care in the interpretation of the coefficient functions not only in the VDFR models, but in any functional regression model. The magnitude of the coefficient function at any particular point $(t, T_i) = (t_0, T_0)$ should only be interpreted conditional on the rest of the curve, the domain width T_i, and the patient population under consideration.

Among the various VDFR models, we observed an advantage in domain-standardization as compared to the Untransformed and Lagged models, both in our simulations and when applied to the ICAP data. The key difference between the fit via the domain-standardized models and the Untransformed/Lagged models is the scale on which the smoothness is applied. The Untransformed/Lagged models apply the same degree of smoothness between adjacent time points regardless of the domain width, T_i. The domain-standardized models, on the other hand, implicitly relax the amount of smoothness between adjacent days when T_i is short, as compared to when T_i is long, because these points are stretched further apart on the domain-standardized scale. It is quite likely that one would want to allow for a greater separation in the estimated weights on days 1 and 2 when a subject is only followed for 3 only days, for example, than when he is followed for 30 or 100. A potential solution would be to employ a smoothness criterion that allows the degree of smoothness to vary with T_i, which we do not attempt in this paper.

It may seem unnatural to stretch a function with a domain of only a few days to be the same width as a function with a domain of 100 or more days, however we remind the reader that we avoid any problems by allowing the coefficient function to change with T_i. We are unsure whether or not we would see the same advantage to domain-standardization if there was not such a large amount of variability in T_i, or if the minimum T_i was greater than just a single day, as is the case in the ICAP data. These questions should be explored in future work. Another advantage of domain-standardization is that it easily allowed us to implement three parametric interaction models (no interaction, linear interaction, and quadratic interaction). These models were usually outperformed by their nonparametric counterparts, but they did offer a number of potential benefits, including more robust estimates, tighter confidence intervals, and greater interpretability.

Our proposed methodology is not without limitations. First and foremost is its inability to dynamically predict mortality during one's ICU stay. The ability to predict whether a subject is likely to survive while they are in the ICU would be quite useful for patient prognostication and treatment. We intend to investigate whether this methodology can be extended to this scenario in future work, perhaps by incorporating ideas from joint models for longitudinal and survival data. Additionally, our methods currently fail to account for missing observations, or sparse or unevenly sampled functional covariates. For the SOFA data, we avoid this scenario by imputing SOFA scores to fill the gaps in our functions. This approach ignores the informative missingness of this data, but we were encouraged by the fact that our sensitivity analyses, including a complete-case-only analysis, produced similar results. This is likely in part due to the fact that only 4% of possible patient days are missing. In cases where the missing data mechanism is assumed uninformative, a preferred approach is to approximate each function using a functional principal components expansion, which would impute each function by borrowing strength from similar functions that do not contain gaps. This procedure has not yet been developed for variable-domain functions. We hope to explore these issues in future work.

Chapter 3

Cox Regression with Functional Predictors

3.1 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 (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; Therneau and Grambsch, 1998; Verweij and Van Houwelingen, 1993); and 3) it estimates the amount of smoothing of the functional parameter using a likelihood-based information criterion such as AIC (Therneau et al., 2003; Verweij and Houwelingen, 1994), 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 (Cardot et al., 1999, 2003; Cardot and Sarda, 2005; Ferraty, 2011; Ferraty and Vieu, 2006; Goldsmith et al., 2011; Harezlak and Randolph, 2011; James, 2002; James et al., 2009; Marx and Eilers, 1999; McLean et al., 2013; Müller and Stadtmüller, 2005; Ramsay et al., 2009; Reiss and Ogden, 2007), 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 effect 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, 2014), 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 $\{\boldsymbol{Z}_i\}$ are 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 $g(\mu_i) = \alpha + \boldsymbol{Z}_i\boldsymbol{\gamma} +$

$\int_0^1 X_i(t)\beta(t)\,dt$, where Y_i follows an exponential family distribution with mean μ_i, and $g(\cdot)$ is an appropriate link function (James, 2002; McCullagh and Nelder, 1989; 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 3.3 describes a number of extensions to the model that allow for added flexibility. Section 3.4 assesses the performance of our model in a simulation study. In Section 3.5, we apply our model to our application of interest, and we conclude with a discussion of our findings in Section 3.6. Additional details regarding our software implementation using `R`'s `pcox` package may be found in the supplemental material.

3.2 Cox Model with Functional Covariates

3.2.1 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 $\log h_i(t; \boldsymbol{\gamma}) = \log h_0(t) + \boldsymbol{Z}_i\boldsymbol{\gamma}$, where $h_i(t; \boldsymbol{\gamma})$ 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 $\exp(\boldsymbol{\gamma})$ 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, $X_i(s) \in \mathcal{L}^2[0,1]$, 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

$$\log h_i\left[t; \boldsymbol{\gamma}, \beta(\cdot)\right] = \log h_0(t) + \boldsymbol{Z}_i\boldsymbol{\gamma} + \int_0^1 X_i(s)\beta(s)\, ds \tag{3.1}$$

The functional parameter, $\beta(s)$, is slightly more difficult to interpret than its non-functional counterpart, $\boldsymbol{\gamma}$. 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 $\beta(s)$ as a functional log hazard ratio. More generally, $\beta(s)$ 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 (Ibrahim et al., 2010; Rizopoulos, 2012; Tsiatis and Davidian, 2004).

As proposed above, (3.1) is under-determined unless we make assumptions on the form of the functional coefficient $\beta(s)$. We will take the common approach of approximating $\beta(s)$ using a spline basis (Cardot and Sarda, 2005; Goldsmith et al., 2011; Marx and Eilers, 1999; Ramsay and Silverman, 2005). Let $\boldsymbol{\phi}(s) = \{\phi_1(s), \phi_2(s), \ldots, \phi_{K_b}(s)\}$ be a spline basis over the s-domain, so that $\beta(s) = \sum_{k=1}^{K_b} b_k \phi_k(s)$.

Then (3.1) becomes

$$\begin{aligned}\log h_i(t;\boldsymbol{\gamma},\boldsymbol{b}) &= \log h_0(t)+\boldsymbol{Z}_i\boldsymbol{\gamma}+\int_0^1 X_i(s)\boldsymbol{\phi}(s)\boldsymbol{b}\,ds \\ &= \log h_0(t)+\boldsymbol{Z}_i\boldsymbol{\gamma}+\boldsymbol{c}_i'\boldsymbol{b} \qquad (3.2)\end{aligned}$$

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, 1988; O'sullivan et al., 1986). We have fit the model using other penalized bases, with minimal change to the estimated coefficient function.

3.2.2 Estimation via Penalized Partial Likelihood

For notational convenience, let $\boldsymbol{\theta}=\begin{bmatrix}\boldsymbol{\gamma} & \boldsymbol{b}\end{bmatrix}$ and $\eta_i(\boldsymbol{\theta})=\boldsymbol{Z}_i\boldsymbol{\gamma}+\boldsymbol{c}_i'\boldsymbol{b}$. Then the partial likelihood for this model is $L^{(p)}(\boldsymbol{\theta})=\prod_{i:\delta_i=1}\left[\exp\{\eta_i(\boldsymbol{\theta})\}\Big/\sum_{j:Y_j\geq Y_i}\exp\{\eta_j(\boldsymbol{\theta})\}\right]$. 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

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

where $P(\boldsymbol{b})$ is an appropriate penalty term for the spline coefficients $\boldsymbol{b}$, and λ 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 $P(\boldsymbol{b}) = \frac{1}{2}\lambda\boldsymbol{\theta}'\boldsymbol{D}\boldsymbol{\theta}$, where $\boldsymbol{D}$ is a symmetric, non-negative definite penalty matrix. Also, let $\boldsymbol{W}_i = \begin{bmatrix} \boldsymbol{Z}_i & \boldsymbol{c}_i' \end{bmatrix}$ 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 λ, 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 $\boldsymbol{\theta}$ by maximizing the partial log-likelihood (for a given λ) using a Newton-Raphson procedure. In the results presented in Sections 3.4 and 3.5, 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).

3.2.3 Optimization of the smoothing parameter

An essential step in the algorithm is the optimization of the smoothing parameter, λ. 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 $\hat{\boldsymbol{\theta}}^{\lambda}_{(-i)}$ be the value of $\boldsymbol{\theta}$ that maximizes $\ell^{(p)}_{\lambda,(-i)}(\boldsymbol{\theta})$, the penalized partial log-likelihood when observation i is left out. Then the CVL for a given value of λ is given by $\mathrm{CVL}\{\lambda\} = \sum_{i=1}^{n} \ell^{(p)}_{\lambda,i} \left\{ \hat{\boldsymbol{\theta}}^{\lambda}_{(-i)} \right\}$, where $\ell^{(p)}_{\lambda,i}(\cdot) = \ell^{(p)}_{\lambda}(\cdot) - \ell^{(p)}_{\lambda,(-i)}(\cdot)$ is the contribution of subject i to the penalized partial log-likelihood.

This expression is quite computationally intensive, as it requires calculating $\ell^{(p)}_{\lambda,i}(\hat{\boldsymbol{\theta}}^{\lambda}_{(-i)})$ for each i. In practice, a computationally efficient alternative is the penalized version of the AIC, $\mathrm{AIC}(\lambda) = \ell^{(p)}_{\lambda}(\hat{\boldsymbol{\theta}}^{\lambda}) - e(\lambda)$, where $e(\lambda) = \mathrm{tr}\left[\boldsymbol{H}_{\lambda}(\hat{\boldsymbol{\theta}}^{\lambda})^{-1} \boldsymbol{H}(\hat{\boldsymbol{\theta}}^{\lambda}) \right]$ 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 (AIC_c) of Hurvich et al. (1998), and the

EPIC criterion of Shinohara et al. (2011). The 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.

3.2.4 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 $\boldsymbol{V}_1 = \boldsymbol{H}_\lambda(\hat{\boldsymbol{\theta}}^\lambda)^{-1}\mathcal{I}(\hat{\boldsymbol{\theta}}^\lambda)\boldsymbol{H}_\lambda(\hat{\boldsymbol{\theta}}^\lambda)^{-1}$. On the other hand, Verweij and Houwelingen (1994) use $\boldsymbol{V}_2 = \boldsymbol{H}_\lambda(\hat{\boldsymbol{\theta}}^\lambda)^{-1}$, 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 $\beta(s_0) = \boldsymbol{\phi}(s_0)\boldsymbol{b}$ may be constructed as $\hat{\beta}(s_0) \pm 1.96\sqrt{\boldsymbol{\phi}(s_0)\boldsymbol{V}_{22}\phi(s_0)}$, 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 λ is fixed. When λ is optimized using one of the methods discussed in Section 3.2.3, 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., 2012) 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 3.4.

3.3 Extensions

3.3.1 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 (3.1) to include additional penalized terms. This makes the inclusion of additional functional predictors, smooth effects of scalar covariates, or frailty terms straightforward. Time-varying coefficients for scalar covariates, $X_i\beta(t)$, may also be included by applying a penalized spline basis to the coefficient function, allowing for non-proportional hazards (Grambsch and Therneau, 1994; Zucker and Karr, 1990). 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.

3.3.2 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 (3.2). Let $\boldsymbol{\psi}(t) = \{\psi_1(t), \psi_2(t), \ldots, \psi_{K_0}(t)\}$ 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 $\boldsymbol{b}_0 = \{b_{01}, b_{02}, \ldots, b_{0K_0}\}$. Then the PFL is

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

where $\eta_i(t, \boldsymbol{\theta}) = \boldsymbol{Z}_i\boldsymbol{\gamma} + \boldsymbol{b}_0\boldsymbol{\psi}(t) + \boldsymbol{c}_i'\boldsymbol{b}$, and λ_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.

3.3.3 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., 2012).

3.4 Simulation Study

3.4.1 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 $\beta_3(s) = -2\phi(s|2, 0.3) + 6\phi(s|5, 0.4) + 2\phi(s|7.5, 0.5)$, where $\phi(\cdot|\mu, \sigma)$ is the density of a normal distribution with mean μ and standard variance σ^2. The three coefficient functions appear in the top row of Figure 3.1.

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).

3.4.2 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 3.2.4. 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.

3.4.3 Simulation results

Box plots of the AMSE and cross-validated C-Index appear in Figure 3.1, 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.

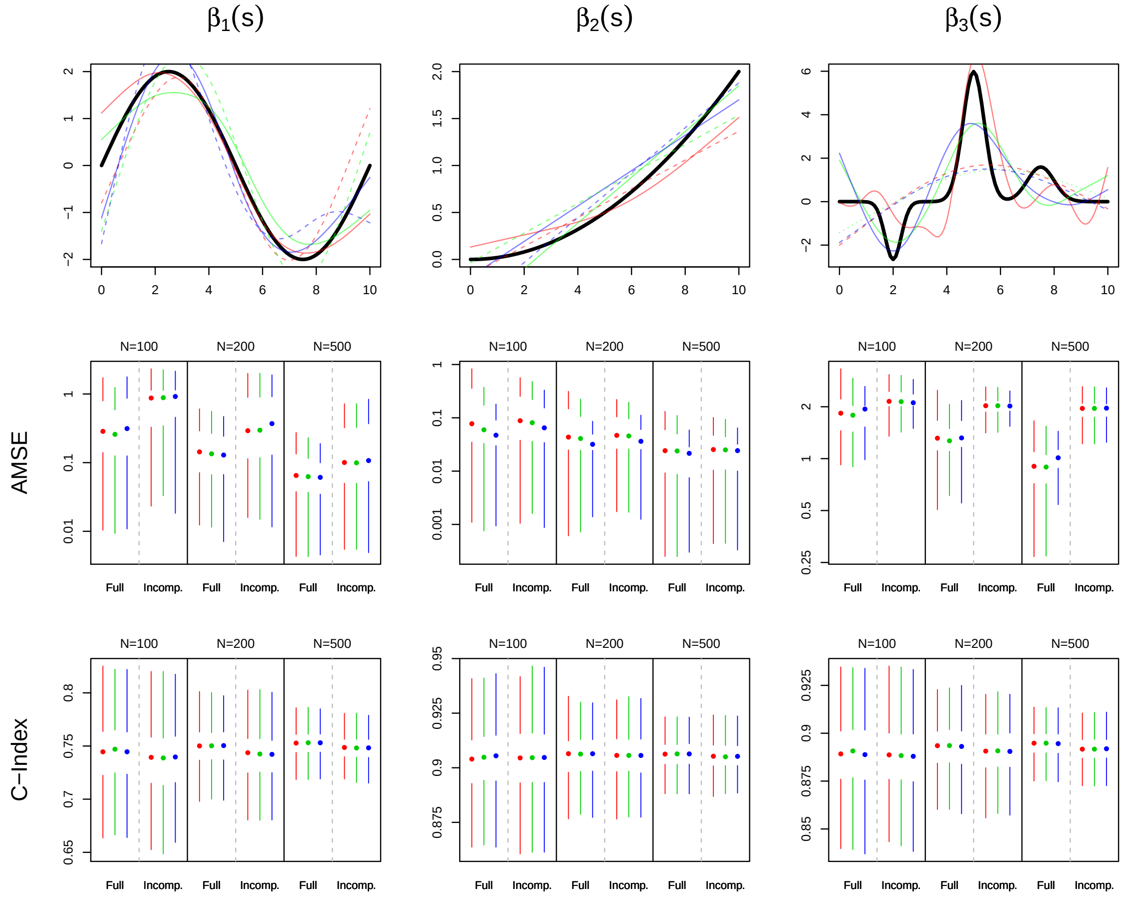


Figure 3.1: Simulation results. In all plots, color indicates the method for optimizing the smoothing parameter: red (AIC), green (AIC$_c$), or blue (EPIC). The top row displays the true coefficient functions (black), as well as the estimates with median AMSE when $N = 200$. The estimates based on the 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.

$\beta_3(s)$, 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

Table 3.1: Mean coverage probability of each of the four pointwise 95% confidence intervals (averaged across s), under each scenario. V1 and V2 are Wald-type confidence intervals based on the model-based estimates of the variance V_1 and V_2, defined in Section 3.2.4. 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

(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 3.1. 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).

3.5 Effect of SOFA Score on Post-ICU Mortality

3.5.1 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 3.2 as a lasagna plot (Swihart et al., 2010).

3.5.2 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

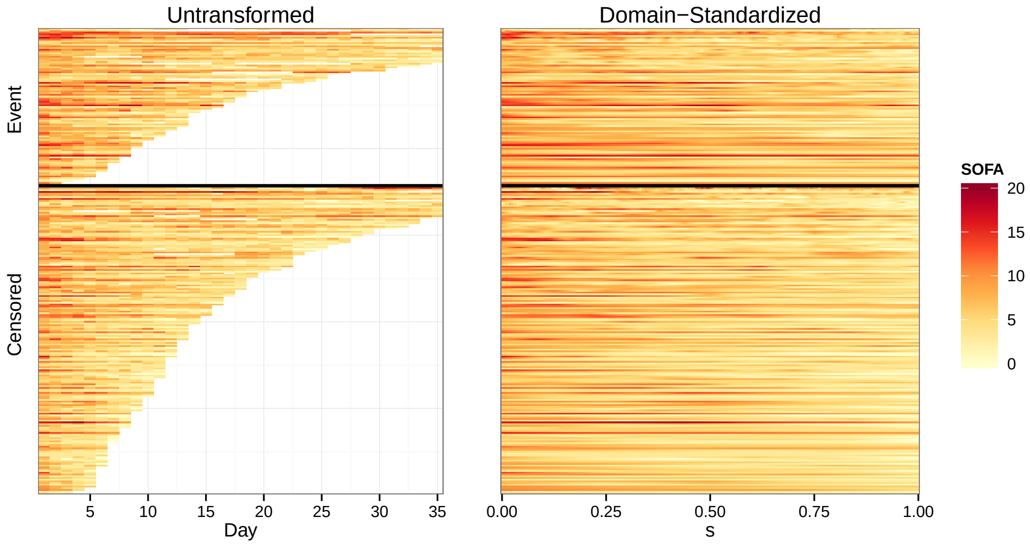


Figure 3.2: SOFA functions, before and after domain-standardization, depicted as lasagna plots. Each row corresponds to a subject, with color indicating the SOFA score at each time point. Subjects are ordered by domain width of the untransformed functions, U_i, within each outcome category (event vs. censored), in both plots.

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 3.2). 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 3.2, 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 3.2, 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 3.3) 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

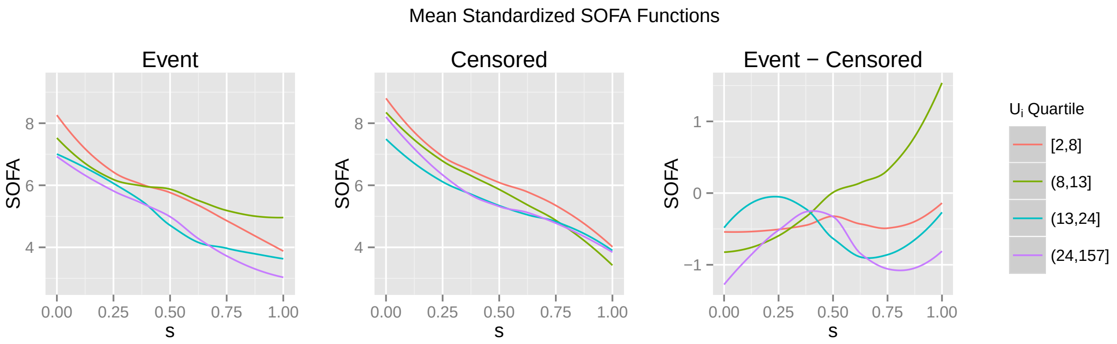


Figure 3.3: Mean domain-standardized SOFA functions, stratified by U_i and event status, as well as the difference in mean functions between those who did and did not experience the event. All curves are smoothed using a lowess smoother.

model using a smooth effect on the log scale. The resulting model is

$$\log h_i(t; \boldsymbol{\gamma}, \beta(\cdot)) = \log h_0(t) + \boldsymbol{Z}_i\boldsymbol{\gamma} + \int_0^1 \tilde{X}_i(s)\beta(s)\, ds + f\{\log(U_i)\} \tag{3.3}$$

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 $f(\cdot)$.

3.5.3 Results

We plot the estimated additive function $\hat{f}\{\log(U_i)\}$ and coefficient function $\hat{\beta}(s)$ based on (3.3) in Figure 3.4, under each of the three optimization criteria. Overall,

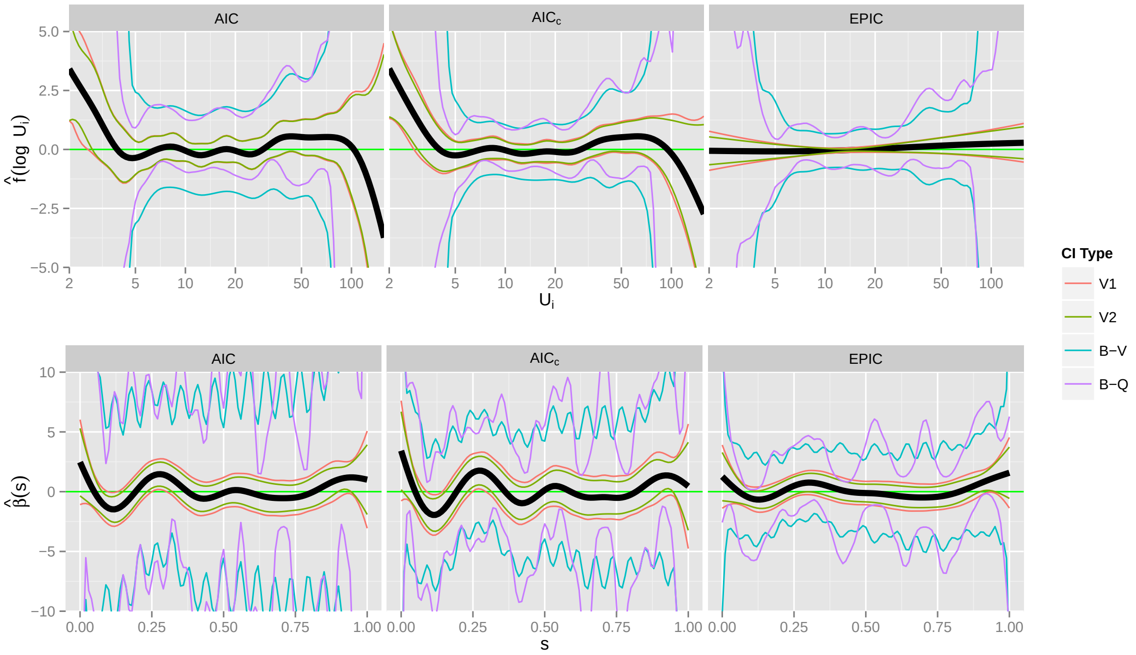


Figure 3.4: Estimated associations between one's log hazard of mortality and their ICU length of stay (top row of figures) and the standardized SOFA functions (bottom row), under each of the three methods for optimizing λ. Pointwise 95% confidence intervals are formed by one of four methods. The first 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.

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 3.2, 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.

3.5.4 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 3.4 reveals some potentially important patterns; in this section we investigate those patterns further. First, we note that the plots for $\hat{f}\{\log(\text{LOS})\}$ 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 3.2). 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

Table 3.2: 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	M1 + M2 + M3 + M4		0.737
Linear Spline	M1 + M2 + M4		0.738
Linear Spline	M1 + M4		0.736
Linear Spline	M1		0.735
Linear Spline	M4		0.739
Linear Spline	S1 + S2 + S3		0.729
Linear Spline	None		0.733
None	None		0.739

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 3.2). The mean SOFA over $s \in [0, 0.05]$ (M1) also shows a potentially important association. Investigating the effects in these regions further, Figure 3.5 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.

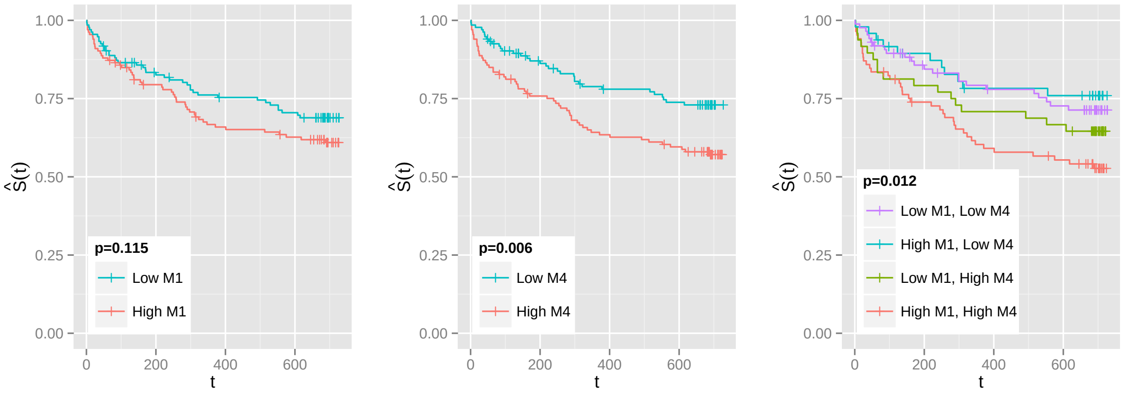


Figure 3.5: Kaplan-Meier estimates of the survival function, stratified by whether the subject has a higher or lower than median value of M1 (left), M4 (middle), and M1 and M4 (right). M1 is the mean SOFA score in the first 5% of one's ICU stay, whereas M4 is in the last 15%. p-values correspond to the log-rank test for the equivalence of the survival functions. "+" indicates censoring.

3.6 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 (3.3) 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.

Chapter 4

The Historical Cox Model

4.1 Introduction

Consider the scenario when one observes a time-to-event outcome and a densely-sampled time-varying predictor. For example, in the ICU we observe if and when people die from the moment they are admitted in the ICU (time to event outcome) and the severity of their health status every day (densely sampled SOFA score). Let (T_i, C_i) denote the event and censoring times, respectively, for subject i, and we observe only $Y_i = \min(T_i, C_i)$, and $\delta_i = I(T_i \leq C_i)$. Suppose that we also observe a time-varying predictor $\{X_i(t_{ij}),\, j = 1, \ldots, J_i,\, t_{iJ_i} = Y_i\}$, measured over a dense grid until time Y_i, as well as fixed, baseline covariates $\boldsymbol{Z}_i = \{Z_{i1}, Z_{i2}, \ldots, Z_{ip}\}$. We assume that T_i and C_i are independent, given the covariates.

One of the primary challenges in including time-varying covariates in a Cox regression model is the need to specify the form of the predictor (Fisher and Lin, 1999). Traditionally, these terms are incorporated into the model by allowing one's hazard function at time t to depend only on the value of the time-varying covariate at that time; i.e., by including the term $\beta X_i(t)$ in the model formula (Cox, 1972). We refer to this type of effect as a concurrent effect of a time-varying covariate. The coefficient β has the interpretation of being the log ratio of the hazard functions at any fixed time t, comparing two subjects whose value of $X_i(t)$ differs by 1 unit, holding all other variables constant. This parameterization is familiar and commonly used, but in some cases it can be overly restrictive. For example, in the intensive care unit (ICU), mortality has been found to be associated not only with the patient's status at a particular time, but with patterns in their status leading up to that time (Gellar et al., 2014; Sakr et al., 2012). This problem was reported in other scenarios as well (Cavender et al., 1992).

More generally, one can allow the hazard function to depend on any function $g(\cdot)$ of the covariate history up to time t, $X_i^{\mathcal{H}(t)}(s) = \{X_i(s) : s \in [0,t]\}$. $g\{X_i^{\mathcal{H}(t)}(s)\} = X_i(t)$ results in the concurrent effect described in the preceding paragraph. Another common technique is to use a time lag, i.e. $g\{X_i^{\mathcal{H}(t)}(s)\} = X_i(t-\delta_0)$ for some fixed δ_0, or a summary statistic of $X_i^{\mathcal{H}(t)}(s)$, such as the mean, maximum, or cumulative exposure. Another option is to compute the slope of a linear regression through some or all of the domain $s \in [0,t]$, and then plug in the slope of the regression into the Cox survival model as a parameter.

Here, we present a method for incorporating time-varying covariates into a Cox model without making assumptions on the form of the predictor. Instead, we apply a flexible method of weighting the entire history $X_i^{\mathcal{H}(t)}(s)$, to express its contribution towards the hazard function $h_i(t)$. This allows all values of the covariate history up to time t to impact the hazard at time t, using functional regression techniques. The key innovation of our approach is the use of a bivariate coefficient function, $\beta(s,t)$, which defines a weight function for $X_i^{\mathcal{H}(t)}(s)$ that varies for each time t. The weights are integrated across the domain $s \in [0,t]$ to provide the total contribution of $X_i^{\mathcal{H}(t)}(s)$ towards $h_i(t)$. A flexible spline basis is applied to the coefficient function, and the spline parameters are estimated by maximizing the penalized partial likelihood. Our method is implemented as part of the `pcox` package in `R`.

The remainder of this article is organized as follows. Section 4.2 introduces the historical Cox model and describes our estimation procedure. Section 4.3 extends the model to two useful contexts: the competing risks context, and the joint modeling context. We perform a detailed simulation exercise in Section 4.4. Section 4.5 applies our model to a study of mortality in the ICU, and we conclude with a discussion in Section 4.6

4.2 The Historical Cox Model

4.2.1 Model

We propose the following model to relate the hazard function $h_i(t)$ to the covariates $\{X_i(t), \boldsymbol{Z}_i\}$:

$$\log h_i(t) = \log h_0(t) + \boldsymbol{\gamma} \boldsymbol{Z}_i + \frac{1}{t} \int_0^t X_i(s)\beta(s,t)\, ds \tag{4.1}$$

We call this model the historical Cox model. Here, $h_0(t)$ is an unspecified baseline hazard function, $\boldsymbol{\gamma}$ are scalar coefficients, and $\beta(s,t)$ is a bivariate functional coefficient. The key feature of this model is the historical functional term $\frac{1}{t}\int_0^t X_i(s)\beta(s,t)\, ds$, which relates $X_i^{\mathcal{H}(t)}(s)$ to $h_i(t)$. The functional coefficient $\beta(s,t)$, which is the primary target of our estimation, describes this relationship. Note that since the variable of integration $s \in [0,t]$ for all t, we are only interested in $\beta(s,t)$ over the restricted domain $0 \leq s \leq t$. This domain covers a triangular surface (Figure 4.1).

The interpretation of $\beta(s,t)$ is most easily understood by fixing t at a particular value, say t_0 (Figure 4.1). With t fixed, $\beta(s,t_0)$ is a univariate weight function with domain $s \in [0,t_0]$, which is applied to $X_i^{\mathcal{H}(t_0)}(s)$. The weighted covariate function $\{X_i(s)\beta(s,t_0) : s \in [0,t_0]\}$ is integrated over its domain and scaled by $1/t_0$, with the result representing the total contribution of $X_i^{\mathcal{H}(t_0)}(s)$ towards $h_i(t_0)$. Regions of $X_i^{\mathcal{H}(t_0)}(s)$ that are positively associated with $h_i(t_0)$ will result in positive estimates of $\beta(s,t_0)$ at the corresponding s locations, and vice versa for negative associations.

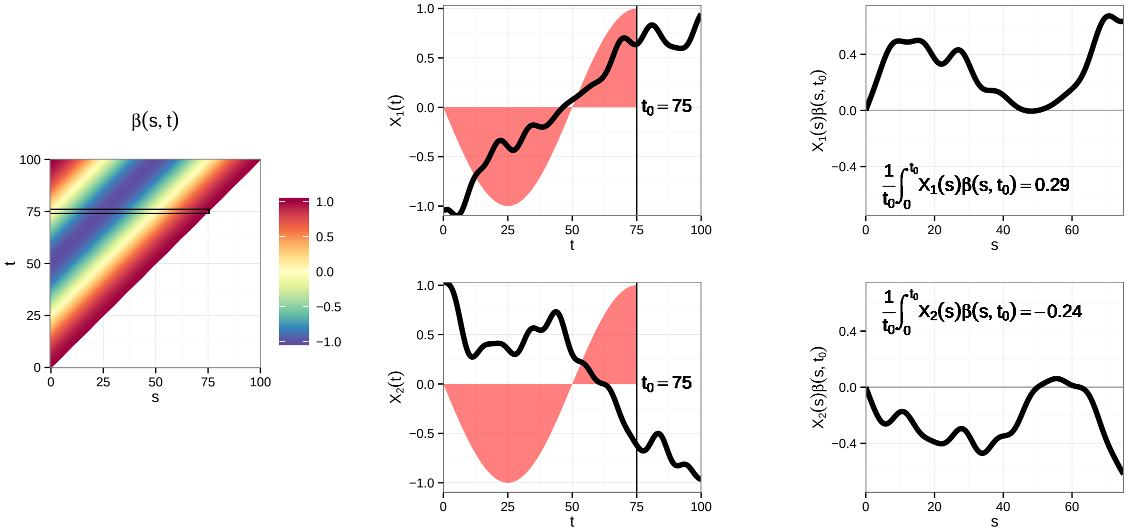


Figure 4.1: Illustration of how $\beta(s,t)$ acts on the covariate functions $X_i(t)$, at $t = t_0 = 75$. Both the coefficient function and covariate functions used in this figure are fabricated. (a) Heat map of a bivariate $\beta(s,t)$. For this illustration, we set $\beta(s,t) = \sin(2\pi s/100 - 2\pi t/100 - 3\pi/2)$. $\beta(s,t_0)$ can be viewed as a single horizontal slice from this heat map. (b) Two examples of potential covariate functions, $X_i(t)$. The weight function $\beta(s,t_0)$ is overlaid in red over each covariate function. (c) The weight function is multiplied by each covariate history $X_i^{\mathcal{H}(t_0)}(s)$, pointwise, to create the weighted covariate functions $X_i(s)\beta(s,t_0)$, $s \in [0,t_0]$. This function is then integrated over its domain and divided by t_0, with the result interpreted as the total contribution of $X_i^{\mathcal{H}(t_0)}(s)$ to $h_i(t_0)$.

The $\beta(s,t_0)$ in Figure 4.1, which uses fabricated data, implies that increasing covariate histories such as $X_1^{\mathcal{H}(t_0)}(s)$ are positively associated with $h_i(t_0)$, and decreasing covariate histories such as $X_2^{\mathcal{H}(t_0)}(s)$ are negatively associated with $h_i(t_0)$.

As t changes, $\beta(s,t)$ changes, smoothly, in both dimensions. This corresponds to our intuition that covariate histories at nearby times will be weighted similarly, while covariate histories at times farther apart may be weighted quite differently.

Note that the fraction $1/t$ in (4.1) is unnecessary, as it could be absorbed into the nonparametric $\beta(s,t)$. However, we include it for two reasons. The first is that its inclusion standardizes the magnitude of $\beta(s,t)$ across different widths of integration, and gives the weight function $\beta(s,t_0)$ an interpretation per unit of time. The second is that it suggests a solution for how to handle the term for $t=0$. Since $X_i^{\mathcal{H}(0)}(s)$ consists of a single observed point, the integration range at $t=0$ is zero. Without the factor $1/t$, the integral would be zero regardless of the value of $X_i(0)$, implying that $X_i(0)$ does not affect the hazard at that time. When $1/t$ is included the term is undefined, but we can replace the expression with its limit as t approaches 0. Since $\lim_{b\to a}\frac{1}{b-a}\int_a^b f(x)\,dx = f(b)$, we have $\lim_{t\to 0}\frac{1}{t}\int_0^t X_i(0)\beta(0,t) = X_i(0)\beta(0,0)$, which is just a concurrent effect for the one available value of the time-varying covariate.

We are not the first to introduce historical terms such as $\frac{1}{t}\int_0^t X_i(s)\beta(s,t)\,ds$ into regression models. Malfait and Ramsay (2003) included a similar term when they introduced the historical functional linear model, which expresses the contribution of the past history of a functional predictor to a functional outcome when both functions are measured concurrently. This model was later expanded on by Harezlak et al. (2007), who introduced various penalization schemes for the coefficients, and Scheipl et al. (2012), who placed the term in a broader mixed model framework. Gellar et al. (2014) used a similar term to express the contribution of covariate functions with subject-specific domains to scalar outcomes. Müller and Zhang (2005) incorporated a historical functional term in a parametric survival model, though their approach did not allow for censored observations. By taking advantage of the Cox regression framework, we allow for censored observations, and also are able to leave the baseline hazard function unspecified. Also, to the best of our knowledge we are the first to make available working code for historical functional regression in general and for its extensions to Cox regression, in particular. Developing reproducible and robust code requires special attention to methodological details and smoothing choices. Below we describe these choices.

4.2.2 Estimation

In order to estimate $\beta(s,t)$ in (4.1), we assume $\beta(s,t) = \sum_{k=1}^{K} b_k \phi_k(s,t)$, where $\{\phi_k(s,t)\}$ are known bivariate basis functions. We favor a thin-plate regression spline (TPRS) basis (Wahba, 1990; Wood, 2003), which adapts well to estimating non-rectangular surfaces. We did not find the finite element basis of Harezlak et al. (2007); Malfait and Ramsay (2003) to be particularly stable, which led to computational instability. We also avoided a tensor product basis because they are designed for rectangular surfaces and not for the triangular surfaces we have to deal with here. One possible criticism of the TPRS basis is that it is isotropic, and smoothness cannot be controlled separately in the s and t directions. Nonetheless, simulations have shown that the basis quite flexible and can easily capture a wide variety of bivariate surfaces (Section 4.4).

We also take the approach of applying a roughness penalty to the TPRS coefficients, which is common in both the smoothing and functional regression literature

(Cardot and Sarda, 2005; Goldsmith et al., 2011; Marx and Eilers, 1999; Ruppert et al., 2003; Wood, 2006). The penalty we use is a second derivative penalty on the scale of the log likelihood, $\lambda \iint \left[\left(\frac{\partial^2 \beta}{\partial s^2} \right) + \left(\frac{\partial^2 \beta}{\partial s \partial t} \right) + \left(\frac{\partial^2 \beta}{\partial t^2} \right) \right] ds\, dt$, which is common for thin plate regression splines. Applying the penalty has two benefits. The first is that it promotes smoothness in $\beta(s,t)$. The second is that it eliminates the need to optimize K, the basis dimension. As long as K is large enough to capture the complexity of $\beta(s,t)$, additional increases in K will likely not affect the model fit considerably (Ruppert, 2002). Here we will assume some familiarity with these concepts and we are simply using well established choices in nonparametric smoothing.

Estimation of functional coefficients in survival models is described in detail in Gellar et al. (2015), which we briefly review here. Let $\boldsymbol{\theta} = [\ \boldsymbol{\gamma} \quad \boldsymbol{b}\ ]$ for spline coefficients $\boldsymbol{b} = (b_1, \ldots, b_K)$ and $\eta_i(t; \boldsymbol{\theta}) = \boldsymbol{Z}_i \boldsymbol{\gamma} + \boldsymbol{c}_i(t)' \boldsymbol{b}$, where $\boldsymbol{c}_i(t)$ is the length K vector with the kth element $\frac{1}{t} \int_0^t X_i(s) \phi_k(s,t)\, ds$. Then the log penalized partial likelihood (PPL) function (Gray, 1992; Therneau and Grambsch, 1998) for this model is

$$\ell_\lambda^{(p)}(\boldsymbol{\theta}) = \sum_{i:\delta_i=1} [\eta_i(\boldsymbol{\theta}) - \log(\sum_{j:Y_j \geq Y_i} \exp\{\eta_j(\boldsymbol{\theta})\})] - \lambda P(\boldsymbol{b}) \tag{4.2}$$

where λ is a smoothing parameter, and $P(\boldsymbol{b})$ is the appropriate penalty. We restrict our discussion to quadratic penalties $P(\boldsymbol{b}) = \boldsymbol{b}' \boldsymbol{D} \boldsymbol{b}$, where $\boldsymbol{D}$ is a positive semi-definite matrix. For a given λ, the PPL may be maximized using the Newton-Raphson algorithm. The smoothing parameter λ can be optimized using a likelihood based information criterion such as AIC, AIC$_c$ (Hurvich et al., 1998), or EPIC (Shinohara et al., 2011). See Gellar et al. (2015) for a comparison of these criteria. In practice, EPIC imposes the highest degree of smoothness on $\beta(s, t)$, and AIC the lowest.

Pointwise confidence intervals may be obtained using either of the two model-based variance estimates suggested by Gray (1992) or Verweij and Houwelingen (1994). Gellar et al. (2015) found through simulations that both estimates tend to underestimate the true standard error, and suggest that confidence intervals with correct coverage may be obtained through a nonparametric bootstrap procedure.

As an alternative to the PPL approach, we could have used the penalized full likelihood as a cost function, by approximating the baseline hazard function with a penalized spline basis (Cai et al., 2002; Kneib and Fahrmeir, 2007; Strasak et al., 2009). As our model already includes penalized splines, this approach would fit neatly into our framework. However, we have found these methods to be difficult to implement, slow to fit, while providing similar results to the PPL approach. Because we are aiming for simplicity and implementability, we will not use the penalized full likelihood here. Additionally, we are unaware of any software that maximizes the full likelihood and is flexible enough to incorporate penalized regression coefficients. We therefore leave the implementation of the PFL estimation to future work.

4.2.3 Software Implementation

The model has been implemented in `R` (R Development Core Team, 2014) language as part of the `pcox` package. `pcox` is a wrapper for two other widely used `R` packages: `mgcv` for setting up the basis and penalization, and `survival` for maximizing the PPL. The model may be fit with just a single line of code:

```
fit <- pcox(Surv(time, event) ~ Z + hf(X), data=dat)
```

In this formula, `X` is an $N \times (\max_i J_i)$ matrix containing the time-varying covariates in what we call "variable-domain functional form", a way of representing a two-dimensional ragged array. In this form, $x_{ij} = X_i(t_{ij})$ is the matrix element at row i and column j of `X`. Cells that correspond to unobserved time points (e.g., they occur after that subject's event or censoring time) should contain a value of `NA`. Here, `Z` is an $N \times p$ matrix containing the baseline covariates, and `dat` is a data frame with N rows and 4 variables: `time`, `event`, `Z`, and `X`. Note that two of these variables are matrices.

The accessory function `hf()` defines the historical functional term from (4.1). The above code assumes that all subjects are observed at the same integer time points $t \in 0, 1, \ldots, t_{J_i}$, but `hf()` contains options to allow for non-integer and also subject-specific time indices. Options also exist to change the basis type and/or penalization, use modified integration limits, or to make the effect non time-varying. There is also an option to make the functional effect of $X_i^{\mathcal{H}(t)}(s)$ nonlinear, analogous to the functional generalized additive model of McLean et al. (2013).

4.3 Extensions

4.3.1 Partial Integration Limits

The historical Cox model allows one's hazard at any time t to depend on the entire history of the time-varying covariate, $X_i^{\mathcal{H}(t)}(s)$. This approach is quite flexible, but sometimes this amount of flexibility is unnecessary. In particular, when the effect of the time-varying covariate history is contained in a small number of observations leading up to time t, estimating the full triangular surface may be unnecessary. To address this issue the integration limits for the integral term in (4.1) can be changed

$$\log h_i(t) = \log h_0(t) + \boldsymbol{\gamma} \boldsymbol{Z}_i + \frac{1}{\zeta(t) - \alpha(t)} \int_{\alpha(t)}^{\zeta(t)} X_i(s)\beta(s,t)\, ds \tag{4.3}$$

for some pre-specified functions $\alpha(t)$, $\zeta(t)$ such that $0 \leq \alpha(t) \leq \zeta(t) \leq t$ for all t. For example, setting $\alpha(t) = \max\{0, t - \alpha_0\}$ and $\zeta(t) = t$ will restrict the effect of $X_i^{\mathcal{H}(t)}(s)$ on $h_i(t)$ to its value over the α_0 time units leading up to time t. For $t > \alpha_0$, this will be $\{X_i(s) : s \in [t - \alpha_0, t]\}$, and for $t \leq \alpha_0$, this will be the entire covariate history $X_i^{\mathcal{H}(t)}(s)$. Note that setting $\alpha_0 = \infty$ results in (4.1).

A further extension of the historical Cox model allows for subject-specific integration limits by replacing $\alpha(t)$ with $\alpha_i(t)$ and $\zeta(t)$ with $\zeta_i(t)$ in (4.3). This may be useful if the time-varying covariate was not observed over the full range $[0, Y_i]$ for some or all subjects, and the start and stop times of measurement are subject-specific. Since our data application does not exhibit this feature, we leave further discussion of this model to future work.

4.3.2 Alternative Basis Parameterizations

The variable-domain functional regression model of Gellar et al. (2014) estimates a triangular surface similar to that of $\beta(s,t)$ in (4.1). The authors introduced an expanded class of models built by re-scaling or re-parameterizing the functional predictors, resulting in improved model fit. Here, we show how an equivalent effect can be achieved by re-parameterizing the basis functions, and leaving the data untouched. The two parameterizations that we focus on are the lagged-time basis and the domain-standardized basis.

Recall that we approximate $\beta(s,t)$ with $\sum_k b_k \phi_k(s,t)$. The lagged-time basis sets $\phi_k(s,t) = \phi_k^*(s^*,t)$, with $s^* \equiv s - t$, and the basis functions $\{\phi_k^*(\cdot,\cdot)\}$ are used to fit the model. Note that s^* is the negative time between any observed $X_i(s)$ and the observation at time t. By using this re-parameterization, the (s,t) coordinates along the right edge (hypotenuse) of the triangular surface in 4.1 are stacked vertically from the perspective of the basis, bringing them closer together. This causes the smoothness across different levels of t to be based on the amount of time until time t, as opposed to the amount of time since time 0. The practical effect is that smoothness is increased along this edge of $\beta(s,t)$, and decreased along the left (vertical) edge.

For the domain-standardized basis, we set $\phi_k(s,t) = \phi_k^*(s^*,t^*)$, with $t^* = t/\max_i(Y_i)$

and $s^* = s/t$ for $t \neq 0$ and $s^* = 0.5$ for $t = 0$. Note that the coordinates (s^*, t^*) will fall on the unit square $[0,1] \times [0,1]$ for the observed data. Using this basis causes the smoothness to be based on the proportion s/t, instead of on s, over different levels of t. The t coordinate is scaled down by $\max_i(Y_i)$ so that $t^* \in [0,1]$, because isotropic bases such as thin-plate regression splines assume the scale of each coordinate of the basis to be the same. Since the coordinates (s^*, t^*) fall on the unit square, we could alternately use a tensor product basis for $\{\phi_k^*(\cdot,\cdot)\}$. The practical effect of this re-parameterization is that a greater amount of smoothness is assumed at high levels of t than at low levels. It may be desirable to allow the resulting weight functions $\beta(s, t_0)$ to be more variable for low t_0 than for high t_0.

4.3.3 Competing Risks

Suppose that instead of a single possible event, each subject experienced one of K different types of events, and let $\Pi_i \in \{1, \ldots, K\}$ indicate the event type experienced by subject i, and that the event times are possibly correlated.

The classical analysis for competing risks data models the cause-specific hazard functions, $h_k(t) = \lim\limits_{\Delta t \to 0^+} Pr(t \leq T < t + \Delta t, \Pi_i = k | T \geq t)/\Delta t$, $k = 1, \ldots, K$, under a proportional hazards assumption (Holt, 1978; Larson, 1984; Prentice et al., 1978). More recently, it has been argued that it is more appropriate to model the subdistribution hazard, $h_k(t) = \lim\limits_{\Delta t \to 0+} Pr\{t \leq T \leq t + \Delta t, \Pi = k | T \geq t \cup (T \leq t \cap \Pi \neq k)\}$, $k = 1, \ldots, K$ (Gray, 1988; Pepe, 1991), as in the model of Fine and Gray (1999).

Although arguments can be made for either model, we will focus on the former as it is the more natural extension of the historical Cox model (4.1), and it is easier to implement using existing software.

Combining this approach with model (4.3), the cause-specific historical Cox model with partial integration limits may be written as

$$\log h_{ik}(t) = \log h_{0k}(t) + \boldsymbol{\gamma}_k \boldsymbol{Z}_i + \frac{1}{\zeta(t) - \alpha(t)} \int_{\alpha(t)}^{\zeta(t)} X_i(s)\beta_k(s,t)\, ds \tag{4.4}$$

, $k = 1, \ldots, K$. The model may be fit as two separate historical Cox models, one for each of the two outcome types (Beyersmann et al., 2013; Lunn and McNeil, 1995). Estimation and inference extend naturally from the approaches presented in Section 4.2.2.

4.4 Simulations

In this section we conduct two detailed simulation studies to investigate the performance of model (4.1) under a variety of conditions and modeling choices. The first simulation looks at model performance when the data-generating model is of the type of model (4.1), and the goal is to accurately estimate the bivariate coefficient function, $\beta(s,t)$. The second scenario investigates how the model performs when the data-generating model includes a simple concurrent effect $\beta X_i(t)$ of the time-varying covariate, or a lagged version $\beta X_i(t-\delta_0)$. For both simulations, we vary the following parameters:

1. Three different sample sizes N: 100, 200, or 500 subjects
2. Three different true coefficient functions $\beta(s,t)$ for Simulation 1, and four different true lags δ_0 (0, 1, 3, or 5 days) for Simulation 2
3. Models based on the full history, and on the partial history of width 3, 5, and 10 days
4. Models based on three different domain transformations: s (untransformed), $s-t$, and s/t.
5. Three different criteria for optimizing the smoothing parameter: AIC, AIC$_c$, or EPIC

For simplicity we consider time-varying covariates observed over an integer-valued time grid, $t_j = j : j \in [1, \ldots, 100]$. For each sample size N we generate $R = 1000$ datasets of predictor functions according to the model $X_i(t_j) = u_i + \sum_{k=1}^{10} \{v_{ik1}\sin(2\pi k t_j/100) + v_{ik2}$ where $u_i \sim N(0,\,1)$, $v_{ik1}, v_{ik2} \sim N(0,\,4/k^2)$. This model is adapted from Goldsmith et al. (2011) and was also used in Gellar et al. (2014).

Survival times are generated according to the model $\log h_i(t) = \log h_0(t) + \eta_i(t)$, where $\eta_i(t)$ differs for each of the two simulation exercises. This is done through an adaptation of the permutation algorithm of Sylvestre and Abrahamowicz (2008). The adaptation we implemented allows for both time-dependent covariates and time-dependent effects, and are included as part of the `pcox` package in `R`. We assume uniform censoring over the time interval $[1, 100]$.

4.4.1 Simulation 1: Estimation of $\beta(s, t)$

In this simulation, $\eta_i(t) = \int_0^t X_i(s)\beta_b(s,t)\,ds$ for three different true coefficient function shapes, $\beta_b(s,t)$. The coefficient functions (displayed as heat maps in the left column of Figure 4.2) are $\beta_1(s,t) = M\left(1 - \frac{t-s}{50})\right)$, $\beta_2(s,t) = M\,(2s/t - 1)$, and $\beta_3(s,t) = M\sin(2\pi t/100)(1 - \frac{t-s}{50})$. Three different values of the effect size (magnitude) M (1, 2, and 5) are investigated. For this exercise, the goal is the ability to estimate $\beta_b(s,t)$, so our results focus on the model fit over the full history. This is measured by the Average Mean Squared Error, $\text{AMSE}\left(\hat{\beta}_b(\cdot,\cdot)\right) = \frac{1}{J(J+1)}\sum_{k=0}^{J}\sum_{j=0}^{k}\left\{\hat{\beta}_b(t_j,t_k) - \beta_b(t_j,t_k)\right\}^2$, i.e., the average of the squared difference between the estimate and the true value of the coefficient function, averaged over its entire surface. We also consider model fit, as measured by the 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 $\eta_i(t)$.

Results for this simulation for moderate effect size ($M = 2$) and EPIC optimization appear in Figure 4.2. As expected, both estimation (as evidenced by decreasing AMSE) and model fit (as evidenced by increasing C-Index) improve as the sample size is increased. We also see that model performance may be substantially affected by transforming the domain over which the basis is applied. Applying the s/t transformation improves the estimation of $\beta_2(s,t)$, but results in much poorer estimation of $\beta_1(s,t)$. From looking at the functional form or the heat map of the two functions,

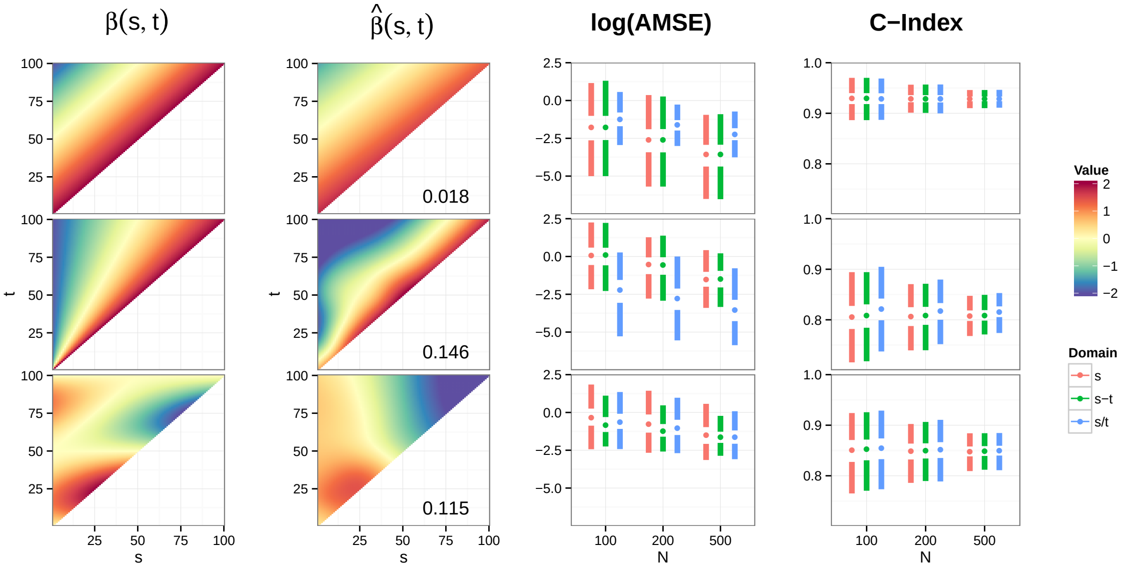


Figure 4.2: Results for Simulation 1, for EPIC estimation under effect size $M = 2$. Each row of plots corresponds to a different true coefficient function, which is depicted in the first column as a heat map. The second column displays the median-performing estimate (as measured by AMSE) over 1000 datasets for the scenario where $N = 200$ and no domain transformation is used; this is done to provide context to the AMSE values. The third and fourth columns present Tufte box plots of the distributions of the AMSE and C-Index, respectively, stratified by sample size and domain transformation. For these plots, the median of each distribution is indicated by the center dot, the interquartile range by the white space around the dot, and the "whiskers" by the colored bars. Lower values of AMSE and higher values of C-Index are indicative of better model performance.

it is clear why this would be. The value of $\beta_2(s,t)$ is based on the the fraction s/t, so a procedure that explicitly take this into account in the smoothness assumption performs better. $\beta_1(s,t)$, on the other hand, is based on $s-t$, so the s/t transformation is counterproductive. We saw very little difference between the untransformed (s) and $s-t$-transformed estimates, though the $s-t$ estimates tend to perform slightly better. This is due to the relatively minor changes to the smoothness assumption that is induced by applying this transformation.

Increasing the effect size M of the coefficient function expected increase in C-Index and decrease in AMSE (Supplemental Material). It is notable, though, that the C-Index was affected to a much greater degree than the AMSE, indicating that the model was able to adequately estimate $\beta(s,t)$ even when less information was available to associate the covariate history with the outcome. We generally saw very little difference in model performance based on the optimization method, with EPIC tending to perform slightly better than both AIC and AIC$_c$.

4.4.2 Simulation 2: Performance Under Concurrent and Lagged Data-Generating Models

For our second simulation we consider a model where data is generated according to a model with $\eta_i(t) = \beta X_i(t-\delta_0)$, for four different values of δ_0 (0, 1, 3, and 5 days). For $t < \delta_0$, we assume $\eta_i(t) = 0$, i.e. no information is supplied by the time-varying covariate. For this simulation, we set $\beta \equiv 1$ for all scenarios. Since these simulations are not based on a true coefficient function, AMSE is not an appropriate metric of model performance, so we instead focus on the C-Index for each fit.

For each simulated dataset, we fit the full historical Cox model (4.1), as well as three versions of the model with partial integration limits (4.3). For the partial models we set $\zeta(t) = t$ and $\alpha(t) = \max\{0, t - \alpha_0\}$ for $\alpha_0 \in \{3, 5, 10\}$; thus the model considers the covariate history over the 3, 5, or 10 days leading up to time t. As in

Section 4.4.1, the model is fit for three different domain transformations and three different criteria to optimize the smoothing parameter. We also fit the model with two more traditional representations of the time-varying covariate: the concurrent effect model, i.e. $\beta X_i(t)$, and the lagged effect model that "knows" the true lag δ_0, i.e. $\beta X_i(t - \delta_0)$.

Figure 4.3 presents heat maps of the historical estimate from the median-performing fit (as measured by C-Index) for each combination of α_0 (defining the integration limits) and δ_0 (true lag), as well as the full distribution of C-Index statistics for each model over the 1000 iterations. Note that the heat maps are plotted with the lagged domain $s - t$ along the x-axis, for interpretability and to avoid excessive white space in the plots. For some of the combinations of (α_0, δ_0), we observe a stronger signal around $s - t = \delta_0$, which indicates that the model is highlighting the true lag in its estimate. However, this signal tends to be spread out over neighboring $s - t$ values, likely due to the correlation in the functional predictors $X_i(t)$. Other estimates do not show any type of noticeable pattern based on $s - t$, and instead show a tendency towards increasing magnitude at larger values of t. Since this effect is not present in the data-generating model, it may be related to the fact that less information is available for larger values of t, because most subjects have either experienced the event or been censored before these times.

It is notable that all the historical models tend to perform well in terms of the C-Index. The "Lagged" model should perform optimally, because it corresponds most

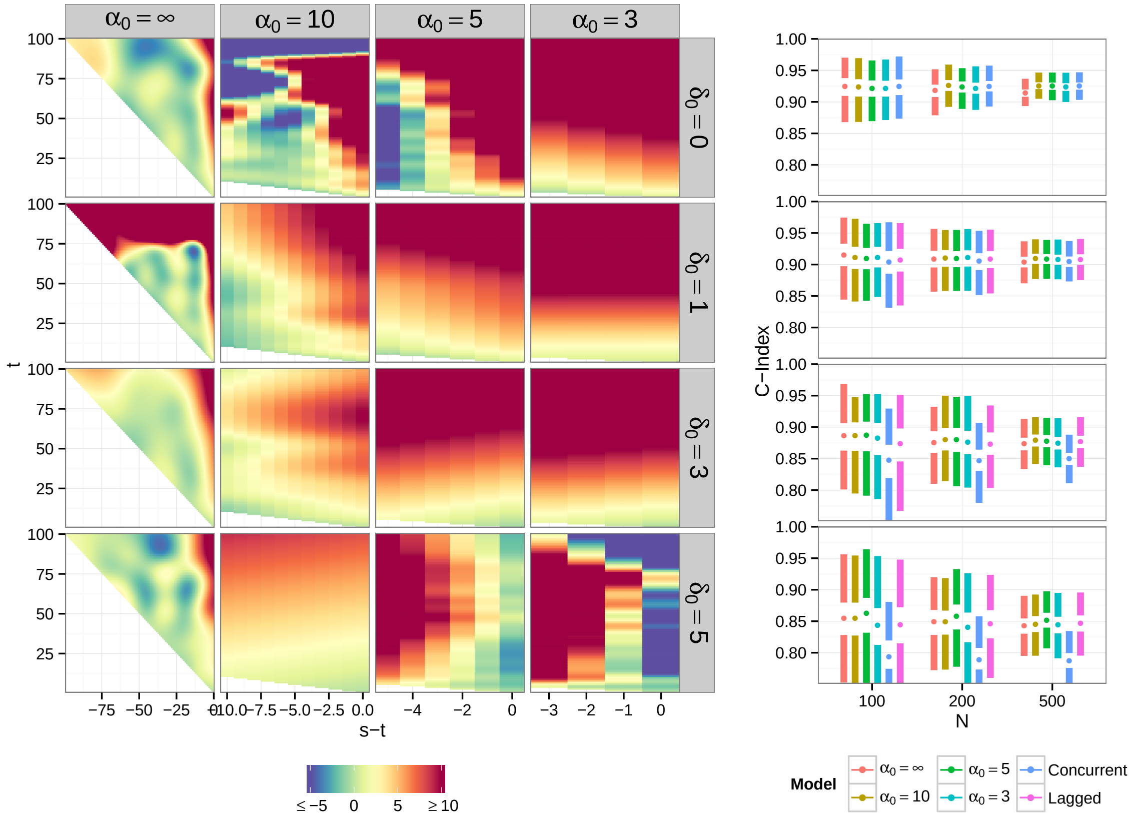


Figure 4.3: Results for Simulation 2, for AIC estimation with the $s - t$ domain transformation. δ_0 indicates the true lag of the model ($\delta_0 = 0$ indicates a concurrent effect), and α_0 indicates the maximum integration width ($\delta_0 = \infty$ corresponds to the full historical Cox model). The left panel of the figure depicts a heat map of the median-performing estimate (as measured by C-Index) for each combination of α_0 and δ_0, for the scenario when $N = 200$. Note that the heat maps are transformed to the lagged domain $(s - t)$ on the x-axis, to avoid excessive white space in the plots. The right panel of the figure displays the distribution of C-Index across the 1000 iterations of the simulation, for each model, true lag, and sample size. See Figure 4.2 for a description of Tufte box plots. The "Lagged" model assumes the true value δ_0 is known.

closely to the data-generating model, but there is not a large drop off in performance using any of the historical models. In some cases the historical models are the top performers, but this is likely due to over-fitting. As δ_0 increases, there is a substantial decrease in performance in the "Concurrent" model, but the historical models continue to perform well.

4.5 ICU Mortality among ARDS Patients

4.5.1 Data Description

We apply our methods to study the relationship between organ failure and mortality for patients hospitalized with acute respiratory distress syndrome (ARDS). ARDS, sometimes referred to as Acute Lung Injury (ALI), is a severe lung condition characterized by inflammation of the lung tissue. It can be initiated by a wide variety of causes, including trauma, infection, or sepsis. A patient with ARDS is typically treated with mechanical ventilation in the intensive care unit (ICU), while physicians attempt to treat the underlying cause of the inflammation. Studies of ARDS report mortality rates as high as 40 and 60 percent (Ware and Matthay, 2000).

Our data data comes from the Improving Care of ALI Patients (ICAP) study, (Needham et al., 2006), a multi-site prospective cohort study conducted in Baltimore, MD. Of the 520 subjects enrolled in ICAP, 283 (54%) survived their hospitalization. Data regarding patient treatment and status is recorded daily while the patient remains in the ICU. Among these daily measurements is the Sequential Organ Failure Assessment score, a composite measure of overall organ function. The SOFA score is based on a set of physiological criteria, with larger values indicating poorer organ

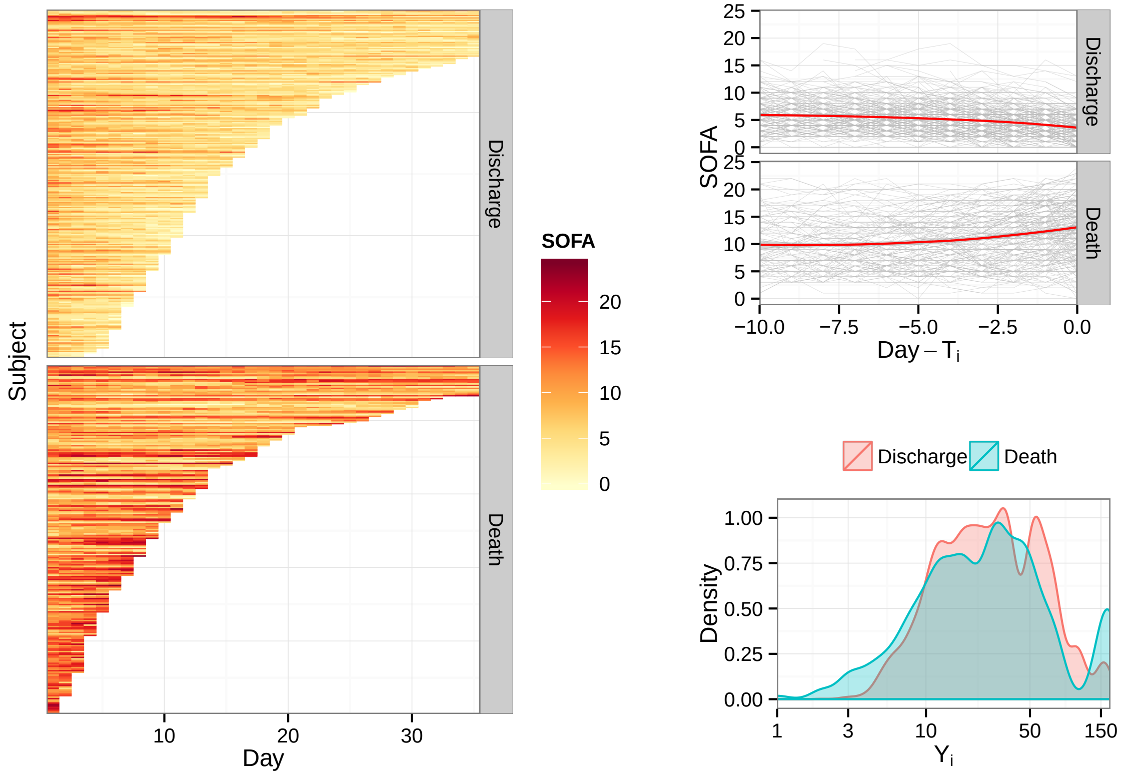


Figure 4.4: SOFA scores in the ICAP dataset, stratified by event type. Left: lasagna plot of the first 35 days of SOFA scores, with each row corresponds to a subject and the SOFA score is indicated by color. Top right: spaghetti plot of each subject's last 10 (or fewer if $Y_i < 10$) SOFA scores, with a lowess smooth overlaid. Bottom right: density plot of the distribution of ICU length of stay, Y_i, on the log scale.

function. When physiological measurements are recorded multiple times during a day, the worst 24-hour score is used. Our goal is to understand the relationship between SOFA and mortality, and to evaluate its use as a predictive biomarker.

SOFA scores are depicted in the left panel of Figure 4.4 as a lasagna plot (Swihart et al., 2010). We see that SOFA scores tend to be higher (more red) for those who die in the hospital, especially towards the end of their hospital stay. We also see a trend over the last 10 days of one's hospitalization of increasing scores for those who died, and decreasing scores for those who survived (Figure 4.4, bottom right). This is not surprising, but it highlights the fact that longitudinal trends, as opposed to just the last SOFA value, may contribute to one's mortality risk at any point in time.

4.5.2 Model Specification

Subjects in the ICAP study leave the risk set in one of two ways: death or hospital discharge. One modeling option would be to consider those who are discharged alive from the hospital to be censored in a historical Cox model. However, Cox regression models assume independent censoring, meaning that one's instantaneous risk for death at time t, given the covariates up to that time, is unaffected by whether or not that subject has been discharged from the hospital. Although this assumption is untestable, we feel it is unlikely to hold for this dataset. The alternative approach that we take is to model hospital discharge as a competing event (Beyersmann et al., 2013; Jackson et al., 2014)

We fit four versions of our comepting risks model (4.4) to the data, each with different time-specific integration limits $[\alpha(t), \zeta(t)]$. All models set $\zeta(t) = t$ and $\alpha(t) = \max\{0, t - \alpha_0\}$, but they differ in the choice of α_0: 5, 10, 25, or ∞. Recall that $\alpha_0 = \infty$ results in integration over the domain of the full covariate history, $[0, t]$. All models are adjusted for age, gender, and baseline health status as measured by the Charlson co-morbidity index (Charlson et al., 1987).

4.5.3 Results: Historical Models

We assess the performance of each model by calculating the C-Index. Uncertainty in these estimates is expressed by constructing a 95% confidence interval based on the 2.5^{th} and 97.5^{th} percentiles of 1000 bootstrap samples. We also compute the proportion of bootstrap samples for which the C-Index of the concurrent model is greater than or equal to that of each historical model. This statistic is a p-value for the null hypothesis that the concurrent model out-performs the corresponding historical model, with performance measured by the C-Index.

C-Index results are presented in Table 4.1 for models fit with the EPIC criterion (results for AIC and AIC_c are available in the supplemental material). Concordances for all models are higher for death than for hospital discharge, which may reflect SOFA having a stronger association with death than with hospital discharge. This makes scientific sense, as increasing SOFA scores (and thus declining health) will often precede death, but a patient with declining SOFA scores might not necessarily be discharged immediately if the physician does not deem it appropriate. The historical models tend to outperform the concurrent model for death, but for hospital discharge the concurrent model resulted in a higher C-Index than most of the historical models. One exception is for the model with $\alpha_0 = 5$ and s/t transformation, which resulted in a higher concordance probability for hospital discharge than the concurrent model (0.783 vs. 0.776, p=0.037).

We plot the estimated coefficient functions $\beta(s, t)$ for the s/t transformation with

Table 4.1: C-Index for each historical model applied to the SOFA data, using the EPIC optimization criterion. C-Index values are expressed as ${}_{L}X_{U}$, where X is the C-Index, and (L, U) are the lower and upper bound of the 95% confidence interval based on 1000 bootstrap samples. p-values correspond to the null hypothesis that the C-Index for each indicated model is less than or equal to the C-Index for the concurrent model, which were ${}_{0.781}0.812_{0.844}$ for Death and ${}_{0.743}0.776_{0.807}$ for hospital discharge. α_0 indicates the range of integration for models fit with partial limits.

		Death		Discharge	
Model	Domain	C-Index	p-value	C-Index	p-value
	s	${}_{0.790}0.816_{0.851}$	0.082	${}_{0.716}0.752_{0.793}$	0.936
Full History	$s-t$	${}_{0.792}0.822_{0.851}$	0.044	${}_{0.721}0.752_{0.797}$	0.888
	s/t	${}_{0.789}0.812_{0.851}$	0.170	${}_{0.750}0.773_{0.816}$	0.216
	s	${}_{0.789}0.821_{0.852}$	0.125	${}_{0.719}0.749_{0.793}$	0.944
$\alpha_0 = 25$	$s-t$	${}_{0.789}0.822_{0.851}$	0.131	${}_{0.721}0.750_{0.798}$	0.851
	s/t	${}_{0.799}0.815_{0.855}$	0.036	${}_{0.752}0.775_{0.815}$	0.146
	s	${}_{0.793}0.820_{0.850}$	0.050	${}_{0.724}0.753_{0.793}$	0.969
$\alpha_0 = 10$	$s-t$	${}_{0.793}0.820_{0.851}$	0.041	${}_{0.734}0.756_{0.802}$	0.802
	s/t	${}_{0.800}0.819_{0.856}$	0.013	${}_{0.749}0.774_{0.812}$	0.268
	s	${}_{0.789}0.817_{0.848}$	0.108	${}_{0.735}0.765_{0.799}$	0.836
$\alpha_0 = 5$	$s-t$	${}_{0.790}0.818_{0.848}$	0.094	${}_{0.735}0.765_{0.798}$	0.863
	s/t	${}_{0.802}0.821_{0.857}$	0.019	${}_{0.755}0.783_{0.818}$	0.037

EPIC optimization as heat maps in Figure 4.5. Estimation is unstable at high values of t; this problem occurs because length of stay is right-skewed and approximately log-normal, see the bottom-right plot of Figure 4.4. Accordingly, our analysis focuses on $t \leq 50$, as this region contains 92% of all observed days.

All four models depict the same overall pattern: a strong positive association along the right ($s = t$) edge of the estimate for death, and a strong negative association along this edge for hospital discharge. This is unsurprising, and reflects our observations based on Figure 4.4 that subjects who die have higher SOFA scores immediately before their death. In the estimates based on the full historical model, the association is

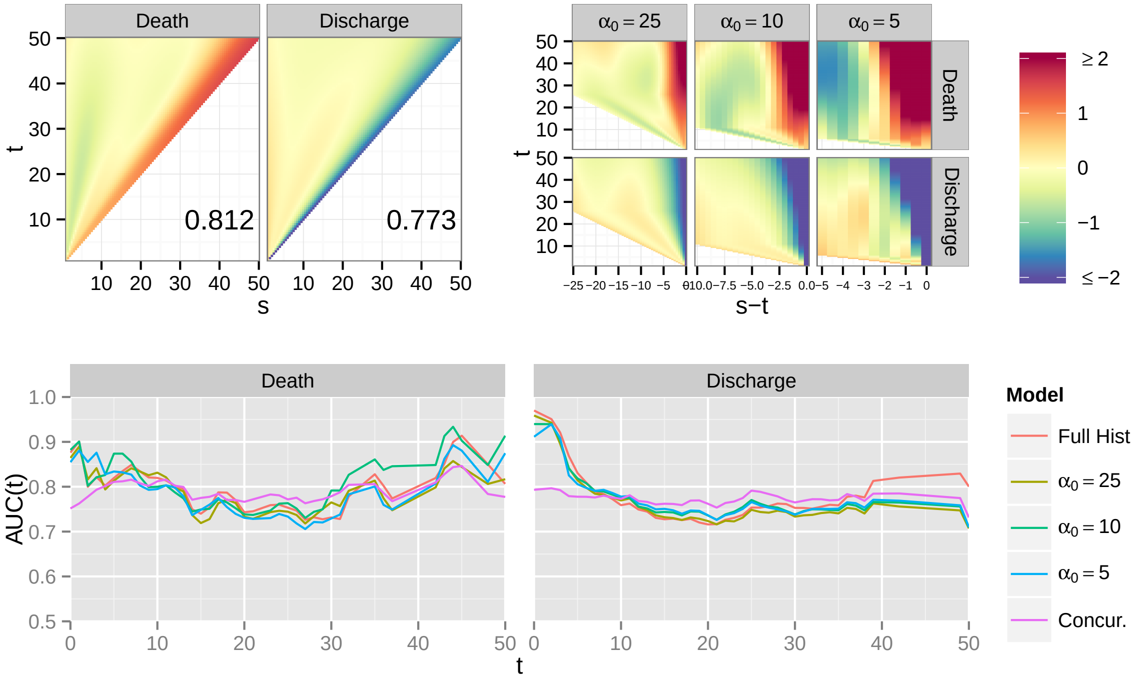


Figure 4.5: Estimation results for the historical Cox models fit to the SOFA data, using the s/t transformation and EPIC optimization criterion, for $t \leq 50$. Top left: Estimates of $\beta(s,t)$ for the full historical Cox model ($\alpha_0 = \infty$), with corresponding C-Index values listed. Top right: Estimates for the historical Cox models with partial integration limits, with the lagged scale $s - t$ used for the x-axis to prevent excessive white space. C-Indices for these fits (and others) appear in Table 4.1. Bottom: Time-varying AUC curves for the four historical models, as well as the concurrent time-varying covariate model.

very close to zero outside this range. This observation suggests that we may be able to obtain a better fit if we focus our estimation along this edge by using a model with partial integration limits, a proposal that is supported by the C-Index estimates in Table 4.1

Somewhat less expected is the trend that occurs in the partial integration estimates around $s - t \in [-15, -5]$. Around this range (with slight differences among the three models), the association with both death and hospital discharge flips directions. In other words, high SOFA scores 5-15 days prior to time t are positively associated with hospital discharge, and negatively associated with death. This seemingly counterintuitive observation was noted in Gellar et al. (2014) with the same data (with mortality was treated as a binary as opposed to time-to-event outcome), and reflects the ability of the model to capture information regarding the trajectory of one's SOFA scores. For two subjects with the same scores in the five days leading up to time t, the subject with worse scores before then is at lower risk of death at time t, because that subject experienced a relative decline in their SOFA scores over time, or at least a less steep incline in those scores. Note that this does not imply that high SOFA score are in any way reflective of better health, as the association of expected direction in the $s-t \in [0,5]$ range is of greater magnitude than the previous period.

In order to further investigate the differences in performance of the various models, we calculate the time-varying AUC curve for each model, $AUC(t) = P(\eta_j(t) > \eta_k(t) | T_j = t, T_k > t)$. We calculate this statistic using the `risksetROC` package in `R` (Heagerty and Saha-Chaudhuri, 2012), which follows the procedure of Heagerty and Zheng (2005), using the "incident/dynamic" definition of time-varying sensitivity and specificity. From these plots (Figure 4.5, bottom) we see that the historical models are most useful for predicting outcome at low or high values of t, but the concurrent model performs best at moderate values of t. This effect is especially pronounced in the models for hospital discharge at low values of t, where $AUC(t)$ for the concurrent model is around 0.8, but it is above 0.9 for all historical models. One possible explanation for this observation is that in the first few days of one's hospital stay, the trajectory of their SOFA scores, not just the magnitude, is strongly associated with whether or not the subject will be discharged from the hospital. Another potential reason for this result is that the true association between SOFA and discharge is concurrent but time-varying, i.e. $\beta(t)X_i(t)$, and since the historical model allows for time-varying effects it captures this association.

4.6 Discussion

In this paper, we introduce a new method for accounting for time-varying covariates in a Cox regression model that allows one's entire (or partial) history of the covariate to impact the hazard at any time. This is accomplished by including the covariate in the model as a historical functional term, and estimating a time-varying functional coefficient that changes smoothly both across one's history at any time t, and with t itself. The resulting coefficient function is interpretable, for a fixed t, as a weight function applied to one's covariate history at that time.

When we applied various versions of our historical Cox model to the SOFA data from the ICAP study, we observed a slight improvement in the concordance probability associated with mortality over the more traditional concurrent effect model. For the competing risk of hospital discharge, we only observed a significant improvement in concordance for the model with $\alpha_0 = 5$ and the s/t transformation. In fact, we observed the s/t transformation to generally improve model performance for both outcomes. This improvement occurs because the transformation effectively relaxes the smoothness for low values of t, when few observations are available, and increases it at higher values of t when more information is available.

It is unclear how often in practice it is necessary to consider such a historical effect. We developed this model for the purpose of applying it to the SOFA data because previous work (Gellar et al., 2014) suggested that one's SOFA trajectory over the last 5-10 days of their hospitalization, and not just their final SOFA score, is useful in distinguishing survivors from non-survivors. Our analysis shows only small improvements over more traditional approaches by using the historical model. Nonetheless, without these methods we would have no way of determining whether the current value of a time-varying covariate captures its entire relationship with one's mortality risk, or if previous values are also informative.

It is important to note that we do not intend for the parameter estimates from these models to be interpreted as causal effects. In particular, a causal model would have to take account for the potential of time-varying confounders (Daniel et al., 2013). For example, Turnbull et al. (2014) showed that limitations or withdrawal of life support is associated with higher SOFA scores, and SOFA scores are in turn likely to affect subsequent decisions to limit or withdraw life support. Our method does not control for this type of confounding, and in order to estimate causal effects in this situation specialized methods are required, e.g. Hernán et al. (2000); Robins (1986); Robins et al. (1992). It would be interesting to see if these approaches could be improved by incorporating the techniques presented in this paper.

As specified in this paper, the historical Cox model cannot be directly used for dynamic prediction of mortality: the ability to predict a subject's future survival curve while they are still in the hospital, based on all information collected up to that point in time. This is because the hazard function in (4.1) is conditional on the most current values of the time-varying covariate, and dynamic prediction looks at future mortality conditional on the current information. However, the model may be combined with existing approaches for dynamic prediction, such as landmarking models (van Houwelingen and Putter, 2012; Van Houwelingen, 2007) or joint modeling of the longitudinal and survival processes (Rizopoulos, 2011). These approaches should be explored in future work.

Chapter 5

Final Remarks

In this thesis I have presented three novel methods for the inclusion of longitudinal predictors in regression models. Chapter 2 introduced variable-domain functional regression, a flexible approach for relating an outcome to a longitudinal predictor in the common scenario where each subject is followed for a different length of time. Chapter 3 presents a blueprint for the inclusion of a baseline functional predictor in a Cox regression model, which to our knowledge had not been attempted prior to this work. Finally, Chapter 4 defines the historical Cox model, which combines approaches from the previous two chapters to present a novel method for modeling time-varying covariates in a Cox model.

The approaches presented in this thesis inspire an entire new area of methodological research that involves functional regression approaches to longitudinal and survival data. For example, one could use the same historical approaches presented in the historical Cox model of Chapter 4 in the context of time-series data, where the outcome is an observed stochastic process recorded over time. Such an approach would have applications, for example, in growth models, which measure child or fetal development regularly over time.

Another important extension of these techniques would be to allow for dynamic prediction of an outcome. The goal of dynamic prediction is to predict future outcomes based on all currently available information, and update those predictions as more information becomes available (van Houwelingen and Putter, 2012). Current approaches to this problem include landmarking (Van Houwelingen, 2007) and joint modeling of longitudinal and survival data (Garre et al., 2008; Proust-Lima and Taylor, 2009; Rizopoulos, 2011; Yu et al., 2008). These models usually assume a concurrent effect of the longitudinal variable, but one could build on their work by treating the history of that variable as functional or historical. These principles could also be applied to landmarking approaches when the outcome is a time series, as opposed to mortality. I am unaware of any existing attempts to perform a landmark analysis with time series data.

Finally, another possible direction of future research would be to extend nonparameteric approaches such as functional principal component analysis (fPCA) or functional penalized least squares (fPLS) to variable-domain functional data. Such approaches would be useful either as an exploratory tool or as a method for presmoothing variable-domain data. The principal component or scores could also be used directly in a regression model as an extension of functional principal component regression or functional penalized least squares (Reiss and Ogden, 2007).

I believe the approaches presented in this thesis will become increasingly important and useful as biomedical studies become more complex in the years ahead, and I am pleased to present this contribution towards the field.

Bibliography

Belitz, C., A. Brezger, T. Kneib, S. Lang, and N. Umlauf (2013). BayesX: Software for Bayesian Inference in Structured Additive Regression Models.

Bender, R., T. Augustin, and M. Blettner (2005, June). Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine 24*(11), 1713–23.

Beyersmann, J., A. Allignol, and M. Schumacher (2013). *Competing Risks and Multistate Models with R.* New York, NY: Springer.

Braess, D. (2007). *Finite elements - Theory, Fast Solvers, Applications in Solid Mechanics.* Cambridge University Press.

Brenner, S. C. and L. R. Scott (2002). *The mathematical theory of finite element methods; Texts in applied mathematics ; 15.*, Volume 2nd. Springer.

Brown, E. R. and J. G. Ibrahim (2003, June). A Bayesian Semiparametric Joint Hierarchical Model for Longitudinal and Survival Data. *Biometrics 59*(2), 221–8.

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

Burnham, K. and D. Anderson (2002). *Model selection and multimodel inference: a practical information-theoretic approach* (2nd ed.). New York: Springer Verlag.

Cai, T. and R. a. Betensky (2003, September). Hazard regression for interval-censored data with penalized spline. *Biometrics 59*(3), 570–9.

Cai, T., R. J. Hyndman, and M. P. Wand (2002, December). Mixed Model-Based Hazard Estimation. *Journal of Computational and Graphical Statistics 11*(4), 784–798.

Cardot, H., F. Ferraty, and P. Sarda (1999, October). Functional linear model. *Statistics & Probability Letters 45*(1), 11–22.

Cardot, H., F. Ferraty, and P. Sarda (2003). Spline Estimators for the Functional Linear Model. *Statistica Sinica 13*, 571–591.

Cardot, H. and P. Sarda (2005, January). Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis 92*(1), 24–41.

Cavender, J. B., W. J. Rogers, L. D. Fisher, B. J. Gersh, C. J. Coggin, and W. O. Myers (1992). Effects of smoking on survival and morbidity in patients randomized to medical or surgical therapy in the Coronary Artery Surgery Study (CASS):

10-year follow-up. CASS Investigators. *Journal of the American College of Cardiology 20*(2), 287–294.

Charlson, M. E., P. Pompei, K. L. Ales, and C. R. MacKenzie (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal Of Chronic Diseases 40*(5), 373–383.

Chiou, J. and H. Müller (2009). Modeling Hazard Rates as Functional Data for the Analysis of Cohort Lifetables and Mortality Forecasting. *Journal of the American Statistical Association 104*, 572–585.

Cox, D. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34*(2), 187–220.

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

Crainiceanu, C. M., A.-M. Staicu, S. Ray, and N. Punjabi (2012, November). Bootstrap-based inference on the difference in the means of two correlated functional processes. *Statistics in medicine 31*(26), 3223–40.

Daniel, R. M., S. N. Cousens, B. L. De Stavola, M. G. Kenward, and J. a. C. Sterne (2013). Methods for dealing with time-dependent confounding. *Statistics in Medicine 32*(9), 1584–1618.

Dinglas, V. D., J. E. Gellar, E. Colantuoni, V. A. Stan, P. A. Mendez-tellez, P. J. Pronovost, and D. M. Needham (2011). Does intensive care unit severity of illness influence recall of baseline physical function? *Journal of Critical Care 26*(6), 1–13.

Eilers, P. H. C. and B. D. Marx (1996, May). Flexible smoothing with B -splines and penalties. *Statistical Science 11*(2), 89–121.

Ferraty, F. (Ed.) (2011). *Recent Advances in Functional Data Analysis and Related Topics*. Berlin: Springer Verlag.

Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. New York: Springer Verlag.

Fine, J. and R. Gray (1999). A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association 94*(446), 496–509.

Fisher, L. D. and D. Y. Lin (1999, January). Time-dependent covariates in the Cox proportional-hazards regression model. *Annual review of public health 20*(6), 145–57.

Garre, F., A. Zwinderman, R. B. Geskus, and Y. W. Sijpkens (2008). A joint latent class changepoint model to improve the prediction of time to graft failure. *Journal of the Royal Statistical Society: Series A 171*(1), 299–308.

Gellar, J. E., E. Colantuoni, D. M. Needham, and C. M. Crainiceanu (2014). Variable-Domain Functional Regression for Modeling ICU Data. *Journal of the American Statistical Association 109*(508), 1425–1439.

Gellar, J. E., E. Colantuoni, D. M. Needham, and C. M. Crainiceanu (2015). Cox Regression Models with Functional Covariates for Survival Data (in-press). *Statistical Modeling 15*(3).

Goldsmith, J., J. Bobb, C. M. Crainiceanu, B. Caffo, and D. Reich (2011). Penalized Functional Regression. *Journal of Computational and Graphical Statistics 20*(4), 830–851.

Goldsmith, J., C. Crainiceanu, B. Caffo, and D. Reich (2011). Penalized Functional Regression Analysis of White-Matter Tract Profiles in Multiple Sclerosis. *NeuroImage 57*(2), 431–439.

Grambsch, P. and T. Therneau (1994). Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. *Biometrika 81*(3), 515–526.

Gray, R. (1988). A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk. *The Annals of statistics 16*(3), 1141–1154.

Gray, R. (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.

Hanson, T. E., A. J. Branscum, and W. O. Johnson (2011, January). Predictive comparison of joint longitudinal-survival modeling: a case study illustrating competing approaches. *Lifetime data analysis 17*(1), 3–28.

Harezlak, J., B. A. Coull, N. M. Laird, S. R. Magari, and D. C. Christiani (2007, June). Penalized solutions to functional regression problems. *Computational statistics & data analysis 51*(10), 4911–4925.

Harezlak, J. and T. W. Randolph (2011). Structured Penalties for Generalized Functional Linear Models (GFLM). In F. Ferraty (Ed.), *Recent Advances in Functional Data Analysis and Related Topics*, Contributions to Statistics, Chapter 25, pp. 161–167. Heidelberg: Springer Verlag.

Harrell, F. E., K. L. Lee, and D. B. Mark (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine 15*, 361–387.

Hastie, T. and R. Tibshirani (1993). Varying-Coefficient Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 55*(4), 757–796.

Heagerty, P. J. and P. Saha-Chaudhuri (2012). risksetROC: Riskset ROC curve estimation from censored survival data.

Heagerty, P. J. and Y. Zheng (2005, March). Survival model predictive accuracy and ROC curves. *Biometrics 61*(1), 92–105.

Hernán, M. a., B. Brumback, and J. M. Robins (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology Cambridge Mass 11*(5), 550–60.

Holt, J. (1978). Competing Risk Analyses with Special Reference to Matched Pair Experiments. *Biometrika 65*(1), 159–165.

Hurvich, C. M., J. S. Simonoff, and C. L. Tsai (1998). Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion. *Journal of the Royal Statistical Society. Series B (Methodological) 60*(2), 271–293.

Ibrahim, J., M. Chen, and D. Sinha (2001). *Bayesian survival analysis.* New York, NY: Springer Verlag.

Ibrahim, J. G., H. Chu, and L. M. Chen (2010, June). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology 28*(16), 2796–801.

Jackson, D., I. R. White, S. Seaman, H. Evans, K. Baisley, and J. Carpenter (2014). Relaxing the independent censoring assumption in the Cox proportional hazards model using multiple imputation. *Statistics in Medicine 33*(27), 4681–94.

James, G. M. (2002, August). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(3), 411–432.

James, G. M., J. Wang, and J. Zhu (2009, October). Functional linear regression thats interpretable. *The Annals of Statistics 37*(5A), 2083–2108.

Katz, S., A. B. Ford, R. W. Moskowitz, B. A. Jackson, and M. W. Jaffe (1963). Studies of Illness in the Aged. The Index of ADL: A Standardized Measure of Biological and Psychosocial Function. *JAMA 21*(185), 914–9.

Kneib, T. and L. Fahrmeir (2007, March). A Mixed Model Approach for Geoadditive Hazard Regression. *Scandinavian Journal of Statistics 34*(1), 207–228.

Larson, M. G. (1984). Covariate analysis of competing-risks data with log-linear models. *Biometrics 40*(2), 459–469.

Lunn, M. and D. McNeil (1995). Applying Cox Regression to Competing Risks. *Biometrics 51*(2), 524–532.

Malfait, N. and J. O. Ramsay (2003, June). The historical functional linear model. *Canadian Journal of Statistics 31*(2), 115–128.

Marx, B. and P. Eilers (1999). Generalized Linear Regression on Sampled Signals and Curves: A P-Spline Approach. *Technometrics 41*(1), 1–13.

Marx, B. and P. Eilers (2005). Multidimensional penalized signal regression. *Technometrics 47*(1), 13–22.

McCullagh, P. and J. A. Nelder (1989). *Generalized linear models*, Volume 37. Boca Raton, FL: Chapman & Hall/CRC.

McLean, M. W., G. Hooker, A.-M. Staicu, F. Scheipl, and D. Ruppert (2013, September). Functional Generalized Additive Models. *Journal of Computational and Graphical Statistics 23*(1), 249–269.

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

Müller, H. and Y. Zhang (2005). Time-Varying Functional Regression for Predicting Remaining Lifetime Distributions from Longitudinal Trajectories. *Biometrics 61*(4), 1064–1075.

Müller, H.-G. and U. Stadtmüller (2005, April). Generalized functional linear models. *The Annals of Statistics 33*(2), 774–805.

Needham, D. M., C. R. Dennison, D. W. Dowdy, P. a. Mendez-Tellez, N. Ciesla, S. V. Desai, J. Sevransky, C. Shanholtz, D. Scharfstein, M. S. Herridge, and P. J. Pronovost (2006, February). Study protocol: The Improving Care of Acute Lung Injury Patients (ICAP) study. *Critical care (London, England) 10*(1), R9.

O'Sullivan, F. (1988). Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal on Scientific and Statistical Computing 9*(3), 531–542.

O'sullivan, F., B. S. Yandell, and W. J. J. Raynor (1986). Automatic Smoothing of Regression Functions in Generalized Linear Models. *Journal of the American Statistical Association 81*(393), 96–103.

Pepe, M. (1991). Inference for Events with Dependent Risks in Multiple Endpoint Studies. *Journal of the American Statistical Association 86*(415), 770–778.

Prentice, R. L., J. D. Kalbfleisch, a. V. Peterson, N. Flournoy, V. T. Farewell, and N. E. Breslow (1978, December). The Analysis of Failure Times in the Risks Presence of Competing. *Biometrics 34*(4), 541–54.

Proust-Lima, C. and J. M. G. Taylor (2009, July). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. *Biostatistics (Oxford, England) 10*(3), 535–49.

R Development Core Team (2014). R: A Language and Environment for Statistical Computing.

Ramsay, J. O., G. Hooker, and S. Graves (2009). *Functional data analysis with R and MATLAB*. New York: Springer Verlag.

Ramsay, J. O. and B. Silverman (2005). *Functional data analysis*. New York: Springer.

Reiss, P. and R. T. Ogden (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(2), 505–523.

Reiss, P. T. and R. T. Ogden (2007, September). Functional Principal Component Regression and Functional Partial Least Squares. *Journal of the American Statistical Association 102*(479), 984–996.

Rice, J. (2004). Functional and longitudinal data analysis: Perspectives on smoothing. *Statistica Sinica 14*(3), 613–629.

Rizopoulos, D. (2011, September). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics 67*(3), 819–29.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R.* Boca Raton, FL: Chapman & Hall/CRC.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure periodApplication to control of the healthy worker survivor effect.

Robins, J. M., D. Blevins, G. Ritter, and M. Wulfsohn (1992). G-estimation of the effect of prophylaxis therapy for Pneumocystis carinii pneumonia on the survival of AIDS patients. Technical Report 4.

Ruppert, D. (2002, December). Selecting the Number of Knots for Penalized Splines. *Journal of Computational and Graphical Statistics 11*(4), 735–757.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric regression*, Volume 12. New York: Cambridge University Press.

Sakr, Y., S. M. Lobo, R. P. Moreno, H. Gerlach, V. M. Ranieri, A. Michalopoulos, and J.-L. Vincent (2012, November). Patterns and early evolution of organ failure in the intensive care unit and their relation to outcome. *Critical care (London, England) 16*(6), R222.

Scheipl, F., A. Staicu, and S. Greven (2012). Functional Additive Mixed Models. *arXiv preprint arXiv:1207.5947*, 1–26.

Schmee, J. and G. Hahn (1979). A Simple Method for Regression Analysis with Censored Data. *Technometrics 21*(4), 417–432.

Shinohara, R., C. Crainiceanu, B. Caffo, and D. Reich (2011). Longitudinal Analysis of Spatiotemporal Processes : A Case Study of Dynamic Contrast- Enhanced Magnetic Resonance Imaging in Multiple Sclerosis. Technical Report September 2011.

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

Strasak, A., S. Lang, T. Kneib, and L. Brant (2009). 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 19*(1), 15–24.

Swihart, B. J., B. Caffo, B. D. James, M. Strand, B. S. Schwartz, and N. M. Punjabi (2010, September). Lasagna plots: a saucy alternative to spaghetti plots. *Epidemiology (Cambridge, Mass.) 21*(5), 621–5.

Sylvestre, M. and M. Abrahamowicz (2008). Comparison of algorithms to generate event times conditional on time-dependent covariates. *Statistics in Medicine 27*(14), 2618–2634.

Therneau, T. M. (2012). coxme: Mixed Effects Cox Models.

Therneau, T. M. (2014). A Package for Survival Analysis in S.

Therneau, T. M. and P. M. Grambsch (1998). Penalized Cox models and Frailty. *Technical report, Division of Biostatistics. Mayo Clinic; Rochester, MN*, 1–58.

Therneau, T. M., P. M. Grambsch, and V. S. Pankratz (2003, March). Penalized Survival Models and Frailty. *Journal of Computational and Graphical Statistics 12*(1), 156–175.

Tsiatis, A. and M. Davidian (2004). Joint Modeling of Longitudinal and Time-to-Event Data: An Overview. *Statistica Sinica 14*(2004), 809–834.

Tsiatis, A., V. DeGruttola, and M. Wulfsohn (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and

CD4 counts in patients with AIDS. *Journal of the American Statistical Association 90*(429), 27–37.

Turnbull, A. E., A. P. Ruhl, B. M. Lau, P. A. Mendez-Tellez, C. B. Shanholtz, and D. M. Needham (2014). Timing of Limitations in Life Support in Acute Lung Injury Patients: A Multisite Study. *Critical Care Medicine 42*(2), 296–302.

van Houwelingen, H. and H. Putter (2012). *Dynamic Prediction in Clinical Survival Analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Van Houwelingen, H. C. (2007, March). Dynamic Prediction by Landmarking in Event History Analysis. *Scandinavian Journal of Statistics 34*(1), 70–85.

Verweij, P. and H. V. Houwelingen (1994). Penalized Likelihood in Cox Regression. *Statistics in Medicine 13*, 2427–2436.

Verweij, P. J. and H. C. Van Houwelingen (1993, December). Cross-validation in survival analysis. *Statistics in medicine 12*(24), 2305–14.

Wahba, G. (1990). *Spline Models for Observational Data*. Montpelier, Vermont: Captial City Press.

Wang, Y. and J. M. G. Taylor (2001, September). Jointly Modeling Longitudinal and Event Time Data With Application to Acquired Immunodeficiency Syndrome. *Journal of the American Statistical Association 96*(455), 895–905.

Ware, L. and M. Matthay (2000). The Acute Respiratory Distress Syndrome. *New England Journal of Medicine 342*(18), 1334–49.

Wood, S. (2006). *Generalized additive models: an introduction with R*, Volume 66. Chapman & Hall/CRC.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (. . . 73*(1), 3–36.

Wood, S. N. (2003, February). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(1), 95–114.

Wu, Y., J. Fan, and H.-G. Müller (2010, August). Varying-coefficient functional linear regression. *Bernoulli 16*(3), 730–758.

Yao, F., H.-G. Müller, A. J. Clifford, S. R. Dueker, J. Follett, Y. Lin, B. a. Buchholz, and J. S. Vogel (2003, September). Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics 59*(3), 676–85.

Yu, M., N. J. Law, J. M. Taylor, and H. M. Sandler (2004). Joint Longitudinal-Survival-Cure Models and their Appliation to Prostate Cancer. *Statistica Sinica 14*(1), 835–862.

Yu, M., J. M. G. Taylor, and H. M. Sandler (2008, March). Individual Prediction in

Prostate Cancer Studies Using a Joint Longitudinal SurvivalCure Model. *Journal of the American Statistical Association 103*(481), 178–187.

Zambon, M. and J.-L. Vincent (2008, May). Mortality rates for patients with acute lung injury/ARDS have decreased over time. *Chest 133*(5), 1120–7.

Zucker, D. and A. Karr (1990). Nonparametric Survival Analysis with Time-Dependent Covariate Effects: A Penalized Partial Likelihood Approach. *The Annals of Statistics 18*(1), 329–353.

Curriculum Vitae

Jonathan Gellar was born on March 17, 1981, and grew up in Highland Park, IL. In 1999 he received a full-tuition academic scholarship to attend the University of Southern California, where he received dual B.S. degrees in Biomedical/Mechanical Engineering and Computer Science in 2004. He moved to Baltimore in 2008 to attend the Johns Hopkins Bloomberg School of Public Health, where he received his M.P.H. in 2009. After completion of the M.P.H. he enrolled in the Biostatistics Sc.M. program at Johns Hopkins, and transferred into the Ph.D. program the following year. His research at Johns Hopkins has focused on developing novel statistical methods for modeling longitudinal biomarkers using functional regression approaches, with a focus on data collected in the intensive care unit. In May 2015 Jonathan will begin working as a Statistician at Mathematica Policy Research in Washington, DC.

CURRICULUM VITAE

Jonathan Gellar
Johns Hopkins Bloomberg School of Public Health
Department of Biostatistics
615 N. Wolfe St., E3031 Baltimore, MD 21205

Phone: (213) 864-6677
E-mail: jgellar1@jhu.edu
Website: www.jonathangellar.com

Education

2015 (expected)	**PhD, Biostatistics**, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD Advisor: Ciprian Crainiceanu Dissertation: Functional Regression Methods for Densely-Sampled Biomarkers in the ICU
2009	**MPH**, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD Concentration: Epidemiologic & Biostatistical Methods in Public Health Advisor: Fernando Pineda Capstone Project: Approximating z-score by Support Vector Regression for Discovery of ncRNA in Genomic Sequence Data
2004	**BS, Computer Science**, University of Southern California, Los Angeles, CA
2004	**BS, Biomedical/Mechanical Engineering**, University of Southern California, Los Angeles, CA

Professional Experience

2014	**Research Intern**, NIH (NINDS), Bethesda, MD Statistical researcher in the Translational Neuroradiology Unit and the AFNI group Develop algorithms for probabilistic brain segmentation based on multi-contrast MRI images
2009-2014	**Research Assistant**, Johns Hopkins Biostatistics Center, Baltimore, MD Statistical consultant for the CLEAR III and MISTIE clinical trials Designed, tested, and implemented the covariate-adaptive randomization algorithm
2008-2014	**Research Assistant**, Johns Hopkins Hospital, Baltimore, MD Pulmonary/Critical Care Medicine - OASIS team Statistical consultant and researcher for the ICAP and ALTOS studies
2004-2007	**Hospital Lab Technician**, UCLA Medical Center, Los Angeles, CA UCLA Blood Donor Center Conducted donor interviews, drew blood, and assisted in the coordination of blood drives
2003-2004	**Lead Intern**, UCLA Medical Center, Los Angeles, CA Help Optimize the Patient Experience (HOPE) project, Patient Relations department Led a group of 15 students in a research project designed to improve hospital communication
2002-2003	**Research Assistant**, University of Southern California, Los Angeles, CA SCOWR wireless robot project, USC Robotics Embedded Systems Laboratory Programmed wireless communication algorithms in TinyOS and C++ computer languages

Publications

Published/Accepted

2015	**Jonathan E. Gellar**, Elizabeth Colantuoni, Dale M. Needham, and Ciprian M. Crainiceanu. Cox Regression Models with Functional Covariates for Survival Data (in-press). *Statistical Modeling*, 2015.
2014	**Jonathan E. Gellar**, Elizabeth Colantuoni, Dale M. Needham, and Ciprian M. Crainiceanu. Variable-Domain Functional Regression for Modeling ICU Data. *Journal of the American Statistical Association*, 109(508):1425-1439, 2014.
2014	Martin B. Brodsky, **Jonathan E. Gellar**, Victor D. Dinglas, Elizabeth Colantuoni, Pedro A. Mendez-Tellez, Carl Shanholts, Jeffrey B. Palmer, and Dale M. Needham. Duration of oral endotracheal intubation is associated with dysphagia symptoms in acute lung injury patients. *Journal of Critical Care*, 29(4):574-9, 2014.
2013	**Jonathan E. Gellar** and Ciprian M. Crainiceanu. Cox Regression Models with Functional Covariates. In *Proceedings of the 28th International Workshop on Statistical Modelling* (Muggeo VMR, Capursi V, Boscaino G, Lovison G, editors), Vol. 1., pages 157-164, 2013.
2013	O J Bienvenu, **Jonathan E. Gellar**, B M Althouse, E Colantuoni, T Sricharoenchai, P A Mendez-Tellez, C Shanholtz, C R Dennison, P J Pronovost, and D M Needham. Post- traumatic stress disorder symptoms after acute lung injury: a 2-year prospective longitudinal study. *Psychological Medicine*, 43(12):2657-71, December 2013.
2011	Victor D Dinglas, **Jonathan E. Gellar**, Elizabeth Colantuoni, Vanessa A Stan, Pedro A Mendez-tellez, Peter J Pronovost, and Dale M Needham. Does intensive care unit severity of illness influence recall of baseline physical function? *Journal of Critical Care*, 26(6):1-13, 2011.

Submitted

2013	Gayane Yenokyan, **Jonathan E. Gellar**, Michael A. Rosenblum, Richard E. Thompson, and Daniel F. Hanley. Achieving balance on key prognostic covariates a trial of a new treatment for hemorrhagic stroke. [Submitted].

In Progress

2014	**Jonathan E. Gellar**, Fabian Scheipl, Mei-Cheng Wang, Dale M. Needham, and Ciprian M. Crainiceanu. The Historical Cox Model..

Honors and Awards

2014	ENAR Student Travel Award (one of 20 awarded annually): *Variable-Domain Functional Regression for Modeling ICU Data*
2013	Best Student Paper, 28th International Workshop on Statistical Modeling: *Cox Regression Models with Functional Covariates*
2009	Delta Omega Honor Society in Public Health
1999-2004	Full-tuition Trustee Scholarship, University of Southern California
1999-2004	Merit Research Program, University of Southern California
1999-2004	W.V.T. Rusch Engineering Honors Program, University of Southern California
2002	Tau Beta Pi, the national engineering honor society
2001	National Society of Collegiate Scholars
2001	Golden Key International Honour Society
2000	Gamma Sigma Alpha, the national Greek honor society
1999	Los Goyescos Chapter of the National Spanish Honor Society
1999	National Merit Scholar

Skills

- Statistical software: R, STATA, and SAS
- Document Markup Languages: LaTeXand R Markdown
- Programming languages: C/C++, Perl, Java, and MATLAB
- Operating systems: Mac, Unix, and Windows
- Relational databases: MySQL and Microsoft Access
- Languages: English (fluent), Spanish (proficient)

Software

- `pcox` package for `R`: lead author
  - Performs various types of penalized Cox regression

- `refund` package for `R`: contributing author
  - `pfr()`: penalized functional regression
  - `lf()`, `af()`, `peer()`, `fpc()`: implementations of linear functional terms, additive functional terms, partially empirical eigenvectors for regression, and functional principal components regression
  - `lf.vd()`: variable-domain functional regression

Teaching Experience

Teaching Assistant

Spring 2015	Analysis of Longitudinal Data and Multilevel Statistical Models (**lead TA**)
Summer 2014	Analysis of Longitudinal Data
Spring 2014	Statistical Methods in Public Health III (**lead TA**)
Summer 2013	Data Analysis Workshop I-II
Spring 2013	Analysis of Longitudinal Data and Multilevel Statistical Models
Fall 2012	Advanced Methods in Biostatistics V-VI
2011-2012	Methods in Biostatistics I-IV (**lead TA**)
Summer 2011	Statistical Reasoning in Public Health I-II
2010-2011	Methods in Biostatistics I-IV
Summer 2010	Analysis of Longitudinal Data
Fall 2010	Statistical Methods in Public Health IV

Mentorship

Research mentor for the following junior researchers:

ScM	James Pringle
ScM	Andrew Leroux

Activities and Leadership Roles

2009-14	Johns Hopkins Biostatistics Computing Club and Journal Club
2010-11	Johns Hopkins Ice Hockey team
2010-11	Johns Hopkins Biostatistics Information Technology (BIT) student representative
2008-09	Johns Hopkins Epidemiology Student Organization
2000-02	USC Ambassador to the university president
1999-2001	USC Ice Hockey team

Professional Memberships

- American Statistical Association
- International Biometric Society

Presentations

Oral Presentations

2015	The Historical Cox Model. *Miami, FL. ENAR*, 2015.
2014	Functional Regression Methods for Densely-Sampled Biomarkers in the ICU. *Ludwig Maximilian University (Munich, Germany), Department of Statistics* (**Invited**).
2014	Functional Regression Methods for Densely-Sampled Biomarkers in the ICU. *North Carolina State University, Department of Statistics* (**Invited**).
2014	The Historical Cox Model. *Boston, MA. JSM*, 2014.
2014	Variable-Domain Functional Regression. *Baltimore, MD. ENAR*, 2014.
2013	Domain-Interaction Functional Regression Models for Functions with Varying Domains. *Montreal, Canada. JSM*, 2013.
2013	Cox Regression Models with Functional Covariates for Survival Data. *Palermo, Italy. 28th International Workshop on Statistical Modeling.*

Poster Presentations

2013	Domain-Interaction Functional Regression Models for Functions with Varying Domains. *Montreal, Canada. JSM*, 2013.
2012	Terminal Event-Dependent Functional Regression with Application to In-Hospital Mortality. *San Diego, CA. JSM*, 2012.