From the Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Stockholm, Sweden

Generalized Survival Models as a Tool for Medical Research

Xingrong Liu

Karolinska Institutet

Stockholm 2017

All previously published papers were reproduced with permission from the publisher.
Cover image made by Leo and Mei with plus plus mini basic.
Published by Karolinska Institutet.
Printed by E-Print AB 2017.

ISBN 978-91-7676-801-3

Generalized Survival Models as a Tool for Medical Research

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Xingrong Liu

Time and location: kl 09:00, November 23 2017 in the lecture hall Atrium, Nobels väg 12B, Karolinska Institutet, Solna

Principal Supervisor:
Associate Professor Mark Clements
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Co-supervisors:
Associate Professor Arvid Sjölander
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Associate Professor Fredrik Wiklund
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Professor Yudi Pawitan
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Opponent:
Professor Virginie Rondeau
French National Institute of Health and Medical Research
Bordeaux Population Health

Examination Board:
Associate Professor Nicola Orsini
Karolinska Institutet
Department of Public Health Sciences

Professor Thomas Scheike
University of Copenhagen
Department of Public Health
Section of Biostatistics

Professor Fan Yang Wallentin
Uppsala University
Department of Statistics

*To **Mei** and **Leo***

Abstract

In medical research, many studies with the time-to-event outcomes investigate the effect of an exposure (or treatment) on patients' survival. For the analysis of time-to-event or survival data, model-based approaches have been commonly applied. In this thesis, a class of regression models on the survival scale, termed *generalized survival models* (GSMs previously described in Appendix A of [1]), and full likelihood-based estimation methods were presented along with four papers. The overall aim was to provide a rich and coherent framework for modelling either independent or correlated survival data.

Our main contributions to GSMs and related estimation approaches were as follows: **First**, we refined the mathematical and statistical backgrounds of the model components, including the link function, log-time, and smooth univariate functions. **Second**, we broadened the class to include generalized additive functional forms for representing covariate effects, such as non-linear forms, time-dependent effects, joint time-dependent and non-linear effects for age, and multivariate regression splines. **Third**, we introduced the thin plate regression splines [2], which can use knot free bases, as an alternative regression tool to knot-based regression splines into GSMs. **Fourth**, under a penalized likelihood framework, we integrated the process of parametric estimation and model selection for the number of spline basis functions. These refinements, extensions, and related assessments were undertaken in the first three papers. These newly proposed features of GSMs and estimation methods were implemented and integrated into the *rstpm2* package in R.

This thesis consists of four research papers for modeling either independent or correlated survival data, together with either overall or net survival to be the measure of interest. In **Paper I**, the outcomes under study were independent time-to-death due to any cause (or time-to-any recurrence of disease). Parametric and penalized GSMs were introduced with extensions, simulation studies and applications. In **Paper II**, the outcome of interest was correlated time-to-some specific event due to any cause, such as time-to-event data collected from patients in the same clinics. It is reasonable to consider that the subjects within a cluster may share some unmeasured environmental or genetic risk factors, which are commonly modeled by a random effect b (or frailty U) and assumed to be independent of given baseline covariates. In this paper, GSMs with novel extensions were proposed to analyze correlated time-to-event data. In **Paper III**, we extended GSMs with novel features for relative survival analysis; the outcome of interest was time-to-death due to the disease under study. In **Paper IV**, we analyzed time-to-repeated event within the same subject using the proposed methods in Paper II and described the time-dependent cumulative risks of subsequent outcomes for men in different states since study entry.

In summary, these proposed methods performed well in extensive simulation studies under the investigated setting, with good point estimates and coverage probabilities. Through the analysis of example data sets, similar results can also be observed using the proposed methods and other well-established approaches, under proportional hazards or proportional odds models settings. Moreover, novel features were also illustrated in both simulations and applications. Generally, the combination of GSMs and full-likelihood based estimation methods can provide alternative tools for the analysis of survival data in medical research.

List of Scientific Papers

(included in this thesis)

I. Xing-Rong Liu, Yudi Pawitan, Mark Clements. Parametric and penalized generalized survival models. *Statistical Methods in Medical Research*. 2016. http://journals.sagepub.com/doi/full/10.1177/0962280216664760. [Epub ahead of print]

II. Xing-Rong Liu, Yudi Pawitan, Mark Clements. Generalized survival models for correlated time-to-event data. *Statistics in Medicine*. 2017;1-20 https://doi.org/10.1002/sim.7451. [Epub ahead of print]

III. Xing-Rong Liu, Yudi Pawitan, Arvid Sjölander, Mark Clements. Use of knot-free splines in one and two dimensions for relative survival. 2017. [*Manuscript*]

IV. Xing-Rong Liu, Jan Adolfsson, Andreas Karlsson, Thorgerdur Palsdottir, Fredrik Wiklund, Henrik Grönberg, Mark Clements. Patterns of prostate-specific antigen (PSA) testing and subsequent outcomes. 2017. [*Manuscript*]

Related Papers

(not included in this thesis)

I. Xing-Rong Liu, Martin Eklund, Jan Adolfsson, Tobias Nordström, Markus Aly, Henrik Grönberg, Mark Clements. Prostate cancer risks following a negative prostate biopsy: population-based cohort study. 2017. [*Submitted*]

II. Hannah Bower, Michael J Crowther, Mark J Rutherford, Therese M-L Andersson, Mark Clements, Xing-Rong Liu, Paul W Dickman, and Paul C Lambert. Capturing simple and complex time-dependent effects using flexible parametric survival models: A simulation study. 2017. [*Submitted*]

Contents

List of Abbreviations

AFT	Accelerated failure time
AH	Additive hazard
AIC	Akaike Information Criterion
ANOVA	Analysis of variance
CRAN	Comprehensive R Archive Network
EDF	Effective degrees of freedom
$\mathrm{g}(\cdot)$	Link function
$\mathrm{G}(\cdot)$	Inverse link function
GSM	Generalized survival model
$h(\cdot)$	Hazard function
$H(\cdot)$	Cumulative hazard
$I(\cdot)$	Indicator function
$L(\cdot)$	Likelihood function
$L^M(\cdot)$	Marginal likelihood function
LASSO	Least absolute shrinkage and selection operator
LCV	Likelihood based cross-validation
$\log(\cdot)$	Base-e log (or the natural logarithm)
MLE	Maximum likelihood estimation
PH	Proportional hazard
PO	Proportional odd
PC	Principal component
PSA	Prostate-specific antigen
P-spline	Penalized B-spline
$\mathrm{s}(\cdot)$	Spline representation (or function)
$S(\cdot)$	Survival function
SE	Standard error
STHLM3	Stockholm 3
$\mathbf{X}(t, \mathbf{z})$	Design matrix
$\mathbf{X}_D(\cdot)$	First derivative of the design matrix with respect to time

$\eta(\cdot)$ Linear predictor function

λ Smoothing parameter

1 Introduction

In epidemiological and clinical studies, many outcomes are expressed as the time to a specified endpoint, such as the time from diagnosis to recurrence of a tumor, or time from diagnosis to death. The analysis of time-to-event (or survival) data needs specific statistical methods to handle censoring and truncation, such as non-parametric methods [3–6] and regression model-based approaches [7–13], combined with likelihood-based estimation methods [14–17].

A conversation with Sir David Cox was reported by Nancy Reid in 1994 [18]:

Reid: *"So if you had a set of censored survival data today, you might rather fit a parametric model, even though there was a feeling among the medical statisticians that that wasn't quite right."*

Cox: *"That is right, but since then various people have shown that the answers are very insensitive to the parametric formulation of the underlying distribution. And if you want to do things like predict the outcome for a particular patient, it is much more convenient to do that parametrically."*

The conversation provides insight into the advantages of parametric survival models, which can estimate baseline functions of survival times (e.g. baseline hazard and survival functions) and related regression coefficients in one statistical procedure. By contrast, the classical procedure usually requires two steps: first, based on Cox regression, to obtain regression coefficients, and then the Nelson-Aalen estimator of cumulative baseline hazards for predicting the survival. Alternatively, to increase flexibility of functional forms for baseline hazard functions or covariate effects, spline-based regression approaches [19–32] have become of interest in survival analysis for either independent or correlated time-to-event data.

Based on maximum likelihood or penalized likelihood estimation methods, a fitted parametric or regression spline-based survival regression model can provide all model components, including regression coefficients and a baseline function. In this context, several interesting features can be introduced to regression spline-based survival models, such as **(1)** generalized additive functional forms [2] for covariate effects, e.g. non-linear forms for continuous variables [33], joint time-dependent and non-linear effects for age [34, 35]; and **(2)** low rank smoothers [36] for representing smooth regression components, which do not depend on spline knots (e.g. the thin plate regression splines [37]), as an alternative regression tool to knot-based regression splines combined with full likelihood-based estimation methods.

In this thesis, within the framework of generalized survival models (GSMs) [1], we have refined the mathematical and statistical backgrounds of the link function, smooth univariate functions, and transformation of time; we have extended some novel features for GSMs and full likelihood-based estimation methods; We have introduced a penalized likelihood framework to integrate the process of parameter estimation and model selection for the number of spline basis functions in one statistical procedure. Under overall and relative survival settings, GSMs can be an alternative tool for the analysis of either independent or correlated survival data in medical research.

2 Background

In observational cohort studies, an initial event could have occurred before study entry (e.g. the diagnosis of a disease) for some subjects, and some individuals could be censored due to emmigration or at end of study period. The analysis of this type of censored (or truncated) survival data is generally termed *survival analysis*.

2.1 Time-to-event outcomes

In survival analysis, observations of time-to-event outcome involve two quantities: (1) the observed survival time T_o that usually measures the duration from study entry until some endpoints (e.g. death and end of study); and (2) the event indicator Δ that denotes whether a specific event (e.g. death due to any cause) occurs. This means that there should always be paired data (t_o, δ) denoting observed survival times. Of more interest, the time to the occurrence of a specific event (T with $\delta = 1$) is considered as a continuous non-negative random variable and has a continuous probability distribution. However, for some subjects, the specific event does not occur or cannot be observed during the study period. In this case, the type of observed survival times is right censored and we define C to be the censoring time with $\delta = 0$. Left-truncated survival data can be encountered in a delayed entry study [38, 39] and time-to-event data can also be interval-censored [15]. In this thesis, right-censored survival data are mainly illustrated.

The situation in which both random variables (T and C) are potentially related to each other is known as *informative censoring*. However the desirable situation is un-informative censoring. In practice, in addition to observed time-to-event data (t_{oi}, δ_{oi}), there are often several variables ($\mathbf{z_i}$) recorded for the patient i at diagnosis of the disease under investigation, such as demographic variables (e.g. sex and age) and clinical characteristics (e.g. stage at diagnosis and primary tumor site) in observational cohort studies. These variables may be termed risk factors, prognostics factors, exposures or treatments in different medical studies. For simplicity, these variables are often termed *covariates* in this thesis.

In general, if not specifically stated in this thesis, we assume that the common assumption of un-informative censoring holds after adjustment for measured con-founders. In this context, T and C are considered to be conditionally independent, given that

$$T|Z = z \perp\!\!\!\perp C|Z = z. \tag{2.1}$$

Note: the term *baseline* is commonly used in this thesis for: (1) *baseline covariates*, which means that the values of covariates are fixed (or time-constant) and measured at study entry; and (2) *baseline functions*, which denotes the functions of survival times for the specified reference group, with the values of zero for each categorized variable and empirical averages for continuous variables.

2.1.1 Independent or correlated time-to-event data

For time-to-event data, one of our aims was to infer the underlying probability distribution and hazard functions of t given $\mathbf{z}$ from a random sample of observations $\{(t_{o1},\delta_1,z_1),(t_{o2},\delta_2,z_2),\ldots,(t_{on},\delta_n,z_n)\}$, where n is the sample size. Before processing, we need to pay attention to the data types, and whether the data are independent. The common assumption is that these samples are independent and identically distributed given values of Z.

However, in some situations [40–42], clustered time-to-event data or repeated survival data within the same individual may be encountered. In these situations, $\{(T_{oij},\Delta_{oij})|Z = z_{ij}; i = 1,2,\ldots,I\}$ for any subject j in the i^{th} cluster may be correlated and within-cluster dependence needs to be accounted for.

Therefore, it is necessary to choose appropriate statistical methods for the analysis of different types of survival data. For example, the Cox regression and the proportional odds model can be applied to independent survival data [7, 43]; some extended proportional hazards or proportional odds model with random effects can be used for correlated survival data [44, 45].

2.1.2 Overall or net (relative) survival

According to the research question of interest, either overall or net survival can be used as measures of cancer survival [46]. Both these settings were investigated in this thesis (i.e. overall survival and relative survival settings).

Let T be the random variable for all-cause mortality, the corresponding overall survival function $S(t|\mathbf{z})$ [8] is defined as the probability that the specific event does not occur within the time interval $(0,t]$, given $\mathbf{z}$ and $t \in (0,\infty)$, and is the complement of the cumulative distribution function $F(t|\mathbf{z}) = P(T \leq t|\mathbf{z})$ in the forms of

$$S(t|\mathbf{z}) = P(T > t|\mathbf{z}) = 1 - P(T \leq t|\mathbf{z}) \tag{2.2}$$

with $S(0|\mathbf{z}) = 1$. Collett [11] used "the greater than or equal to" sign "$\geq$" for defining the survival function. The hazard function $h(t|\mathbf{z})$ [8] is then defined as

$$h(t|\mathbf{z}) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t; \mathbf{z})}{\Delta t} = -\frac{d\log\{S(t|\mathbf{z})\}}{dt}, \tag{2.3}$$

which is interpreted as an instantaneous rate of event occurring at time t for subjects surviving at that time, with further relationships [8] between hazard, cumulative hazard $H(t|\mathbf{z}) = \int_0^t h(u|\mathbf{z})\mathrm{d}u$, and survival functions of continuous times t

$$S(t|\mathbf{z}) = \exp(-\int_0^t h(u|\mathbf{z})\mathrm{d}u) = \exp(-H(t|\mathbf{z})).$$

In a long-term follow-up cohort study, the mortality for each patient may be due to either the disease in question or all other causes, especially in elderly patients [47]. In the relative survival setting, Pohar-Perme and colleagues [6] defined T_{Ei} as the time random variable of death due to cancer with the survival function: $S_{Ei}(t|\mathbf{z_i}) = P(T_{Ei} > t|\mathbf{z_i})$, and T_{Pi} as the time random variable of death due to other causes with the survival function:

$S_{Pi}(t|\mathbf{x_i}) = P(T_{Pi} > t|\mathbf{x_i})$. Thus $T_i = min\{T_{Ei}, T_{Pi}\}$ is the time random variable of all-cause mortality and the overall (all-cause) survival [48] function for each patient can be given as

$$\begin{aligned} P(T_i > t|\mathbf{z_i}) &= P(min\{T_{Ei}, T_{Pi}\} > t|\mathbf{z_i}) \\ &= P(T_{Ei} > t, T_{Pi} > t|\mathbf{z_i}) \\ &= P(T_{Ei} > t|\mathbf{z_i}) \cdot P(T_{Pi} > t|\mathbf{z_i}) \quad (2.4) \\ &= S_{Ei}(t|\mathbf{z_i}) \cdot S_{Pi}(a_i + t, y_i + t|\mathbf{x_i}) \quad (2.5) \\ &= S_i(t|\mathbf{z_i}), \quad (2.6) \end{aligned}$$

based on two main assumptions: (1) the formula 2.4 holds only if T_{Ei} and T_{Pi} are conditionally independent, given a set of baseline variables $\mathbf{z_i}$; and (2) the formula 2.5 holds only if the mortality rate due to other causes can be represented by the matched population life tables stratified by a set of variables $\mathbf{x_i}$, with a_i and y_i being the age and period at diagnosis for the subject i. This verifies previously proposed assumptions [6, 49] and demonstrates that the net survival for an individual i is equivalent to the individual-level relative survival ratio:

$$S_{Ei}(t|\mathbf{z_i}) = \frac{S_i(t|\mathbf{z_i})}{S_{Pi}(a_i + t, y_i + t|\mathbf{x_i})}. \quad (2.7)$$

Furthermore $h_i(t|\mathbf{z_i})$, $h_{Pi}(t|\mathbf{x_i})$ and $h_{Ei}(t|\mathbf{z_i})$ [6] are the hazard functions for each patient, corresponding to the survival functions $S_i(t|\mathbf{z_i})$, $S_{Pi}(t|\mathbf{x_i})$ and $S_{Ei}(t|\mathbf{z_i})$, respectively. Based on the formula 2.7, the excess mortality due to cancer under study is the difference between the observed (all-cause) hazards and the mortality due to other causes [50, 51], which is expressed as

$$h_{Ei}(t|\mathbf{z_i}) = h_i(t|\mathbf{z_i}) - h_{Pi}(a_i + t, y_i + t|\mathbf{x_i}), \quad (2.8)$$

and the empirically marginal survival (net survival) [6] is

$$S_E(t) = \frac{1}{n}\sum_{i=1}^{n} S_{Ei}(t|\mathbf{z_i}) = \frac{1}{n}\sum_{i=1}^{n} \exp\{-\int_0^t h_{Ei}(u|\mathbf{z_i})\mathrm{d}u\}. \quad (2.9)$$

Similarly, based on proportional hazards models for independent survival data, adjusted all-cause survival curves have been applied in medical research [52] .

2.2 Statistical models for survival data

2.2.1 Parametric survival models and their connections

The log-transformation of T is used to associate with prognostic factors in the broad family of parametric *accelerated failure time* (AFT) models [11, 53] and is usually described as

$$\log(T) = \mu + \beta^T z + \sigma\varepsilon. \quad (2.10)$$

where μ denotes the intercept, β and σ are unknown model parameters, and ε is the random error. It is clear that survival times could be different, even if the given baseline variables are the same for some subjects.

In general, parametric AFT models provide a basic framework to connect other survival models introduced by Collett [11]. According to the definition of survival probability function,

$$\begin{aligned} S(t|\mathbf{z}) = P(T > t|\mathbf{z}) &= P(\log(T) > \log(t)|\mathbf{z}) \\ &= P(\mu + \beta^T\mathbf{z} + \sigma\varepsilon > \log(t)|\mathbf{z}) \\ &= P(\varepsilon > \frac{\log(t) - \mu - \beta^T\mathbf{z}}{\sigma}). \end{aligned}$$

Assuming a particular standard probability distribution for the random error ε, the corresponding survival probability distribution of log-transformed survival times is specified. For example, suppose ε has the standard logistic distribution, then

$$S(t|\mathbf{z}) = \left\{1 + \exp\left(\frac{\log(t) - \mu - \beta^T\mathbf{z}}{\sigma}\right)\right\}^{-1},$$

which can be reformulated as [11]

$$\begin{aligned} \log\left\{\frac{1 - S(t|\mathbf{z})}{S(t|\mathbf{z})}\right\} &= \log\left\{\frac{1 - S_0(t)}{S_0(t)}\right\} + \beta_{po}^T\mathbf{z} \\ &= b_0 + b_1\log(t) + \beta_{po}^T\mathbf{z}, \end{aligned} \tag{2.11}$$

where $S_0(t)$ is the unknown basis survival function, and b and β_{po} are re-arranged parameters. The models have monotonic and unimodal hazards, which have been applied to lung cancer survival data [54] and can be identified as the *log-logistic proportional odds (PO) model* for survival data [43].

Similarly, the Weibull AFT model is related to the *Weibull proportional hazards (PH) model* [11] in the form of

$$h(t|\mathbf{z}) = \lambda\gamma t^{\gamma-1}\exp(\beta_{ph}^T\mathbf{z}),$$

which can be expressed in survival function as

$$\begin{aligned} \log\{-\log(S(t|\mathbf{z}))\} &= \log\{-\log(S_0(t))\} + \beta_{ph}^T\mathbf{z} \\ &= \log\lambda + \gamma\log(t) + \beta_{ph}^T\mathbf{z}. \end{aligned} \tag{2.12}$$

In particular, in relative survival analysis, the excess hazards may monotonically decrease along with the follow-up time, and it could be reasonable to explore Weibull or log-logistic parametric distributions for baseline hazard functions (see the related analysis and results in Paper III).

In general, there are relationships between the parameters in either the Weibull PH model 2.12 or the log-logistic PO model 2.11 and the corresponding parametric AFTs. R code for the related comparison is provided in the Appendix A.1.

2.2.2 Spline-based regression models and semi-parametric approaches

To relax parametric forms for transformed baseline functions and covariate effects, spline-based smooth modeling techniques, such as B-splines [55], restricted cubic splines [53], P-splines [56], smoothing splines [57] and a class of penalized regression splines [2], can be applied to represent smooth regression components. Following Wood [37], regression splines (see Table 2.1) can be classified as: (1) pure regression splines, which may be dependent on knots or not, usually with the number of spline basis functions as the dimension of unknown parameters; and (2) penalized regression splines, implemented in the *mgcv* package in R, including penalized B-splines (P-splines), with effective degrees of freedom (or the *equivalent number of parameters* [36]) to be automatically selected in the estimation procedure; where we need to specify the maximum dimension of spline basis functions for each smooth regression components.

Table 2.1: Four types of regression splines

Regression splines	**knot-based bases**	**knot-free bases**
Pure	1. Require knots locations[c]	Require the number of bases
(MLE[a])	2. Require the number of knots	
	(e.g. B-spline)	(e.g. thin plate spline basis)
Penalized	1. Require knots locations[c]	Require the maximum dimension
(MPLE[b])	2. Require the maximum dimension	
	(e.g. P-spline)	(e.g. thin plate spline basis)

a MLE: Maximum likelihood estimation.
b MPLE: Maximum penalized likelihood estimation.
c Interior knots can be located at either the quantiles of x or equally-spaced points.

In general, B-splines and restricted cubic splines are predominantly used to estimate smooth functions, commonly with the quantiles of continuous variables as predefined knots in both Stata and R [30, 53]. By contrast, P-splines and penalized truncated power basis functions [36] have been compared previously [58]. However, within the framework of penalized knot-based regression splines, Eilers and Marx [58] pointed out that equally spaced knots may be preferred, especially for constructing simple difference penalties. Compared to these knot-based splines, there is one type of knot-free regression splines based on radial basis functions (so-called *penalized thin plate regression splines* [37]), which are mainly applied in generalized additive models implemented by Wood [2].

Truncated power basis functions

The truncated power function (or truncated piece-wise polynomial) of a continuous variable x with exponent 3 can be defined as [36, 59]

$$x_+^3 = \begin{cases} x^3, & x > 0 \\ 0, & x \leq 0. \end{cases}$$

Based on truncated power basis functions, a cubic spline function $s(x)$ [20] can be represented as

$$s(x) = \sum_{j=0}^{3} \beta_{0j} x^j + \sum_{i=1}^{K} \beta_{i3} (x - t_i)_+^3, \tag{2.13}$$

with the position of K knots being $(t_1 < t_2 < \cdots < t_K)$. To avoid having a poorly behaved in the tails, Stone and Koo [60] proposed to construct the cubic function to be linear in the tails, which means that the function $s(x)$ is subject to $\beta_{02} = \beta_0 = 0$ and $\sum_{i=1}^{K} \beta_{i3} = \sum_{i=1}^{K} \beta_{i3} t_i = 0$ (see also [16, 20]).

For example, this type of restricted cubic splines (RCS) has been mainly applied to the family of *Royston-Parmar models* or *flexible parametric survival models* [16, 30], implemented in Stata with the name *stpm2*.

B-spline basis functions

B-splines can be derived by the Cox-de Boor recursive formula [55],

$$B_{i,0}(x) = \begin{cases} 1 & \text{if } \ t_i \le x < t_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

$$B_{i,k}(x) = \frac{x - t_i}{t_{i+k} - t_i} B_{i,k-1}(x) + \frac{t_{i+k+1} - x}{t_{i+k+1} - t_{i+1}} B_{i+1,k-1}(x).$$

where $t_1 < t_2 < \cdots < t_k$ are the ordered k knots. Eilers and Marx [58] provided an example and the R code to demonstrate the relationship between B-splines and truncated power functions using equally spaced knots, see [61] for further details.

Based on B-splines, another type of restricted cubic splines (also called *natural splines* [36]) can be imposed on the same linear constraints in the tails as $s(x)$. For example, in the R package *rstpm2* [62], natural cubic splines are adopted as the default regression splines in the *stpm2* function.

Radial basis functions

Another set of related spline basis functions are the *radial basis functions* described in [36]. One advantage of this type of spline basis functions is that it can construct either univariate or multivariate splines. Based on the radial basis functions, both knot-based and knot-free bases can be constructed for modelling smooth functions [2].

Within knot-based approaches, given knots $x_1 < x_2 < \cdots < x_K$, the univariate radial basis functions can be expressed as

$$B_i(x) = |x - x_i|^3 = 2(x - x_i)_+^3 - (x - x_i)^3.$$

From this it is clear that radial basis functions are also related to truncated basis functions. Based on the above cubic radial basis functions, knot-based thin plate

regression splines [37] can be constructed as

$$s(x) = \sum_{j=0}^{3} \beta_{0j} x^j + \sum_{i=1}^{K} \beta_{i3} |x - x_i|^3$$

and imposed on the same linear constraints in the tails as the above restricted cubic splines.

Within knot-free approaches, Wood [37] introduced to use the truncated eigendecomposition method for producing knot free bases [2]. The thin plate regression splines, which do not use "knots" [2], can be a type of knot-free regression splines. For example, in the R package *rstpm2* [62], the penalized thin plate regression splines are treated as the default regression splines for the penalized estimation approaches with the function *pstpm2*.

2.2.2.1 Spline-based hazard models

Cox regression [7] is the most popular survival regression model applied to the analysis of time-to-event data

$$h(t|\mathbf{z}) = h_0(t)\exp(\beta_z^T \mathbf{z}),$$

with the baseline hazard function $h_0(t)$ unspecified and a partial likelihood estimation framework for estimating regression coefficients β_z.

In the past three decades, spline-based smoothing techniques have been commonly applied to the Cox proportional hazards model and extended models on hazard (or log-hazard) scale, e.g. for representing non-linear effects of continuous covariates [21], for modeling continuous-by-continuous interaction terms [22], for time-dependent effects [63, 64], and for joint time-dependent and non-linear effects for age [34, 35]. More generally, the geoadditive hazard regression (combined P-splines) using a mixed model approach [65, 66], smoothing spline ANOVA models for survival data [31, 57], and the related regression models on hazard scale by Wood [67] have been developed for the analysis of survival data.

For example, within the extended framework of Cox regression, four smoothing techniques including restricted cubic splines, P-splines, natural splines and fractional polynomials [68, 69] were compared for modelling non-linear effects of continuous covariates under several simulated settings; see [70] for details.

2.2.2.2 Spline-based generalized survival models

Alternatively, a class of spline-based regression models on the survival scale was originally introduced by Younes and Lachin [71] using the term *link-based models*,

$$g_c\{S(t|\mathbf{z})\} = g\{S_0(t)\} + \beta_z^T \mathbf{z}, \tag{2.14}$$

where B-splines were applied to represent the baseline hazard function and link functions can be generalized to the parametric family $g_c(x) = \log\frac{x^{-c}-1}{c}$.

Royston and Parmar [16] developed the class of spline-based survival models using restricted cubic splines (based on truncated power functions) to represent transformed baseline functions $g\{S_0(t)\}$ [16]. Within this framework, Royston, Lambert, and

colleagues [1, 16, 30] further extended Model 2.14 with time-dependent effects to produce a family of *flexible parametric survival models* in Stata, such as for clustered survival data [72], competing risk [73], and population-based cancer survival data [74–76]. The term *"generalized survival models"* was first introduced by Royston and Sauerbrei in the Appendix A of [1].

2.2.2.3 Semi-parametric approaches

More specifically, under the similar model framework with 2.14, Dabrowska and Doksum introduced the class of semiparametric *"generalized odds-rate models"* [77, 78]

$$g_\rho\{S(t|\mathbf{z})\} = \alpha(t) + \beta_z^T \mathbf{z} \tag{2.15}$$

to include both proportional hazards and odds model as special cases, where $\alpha(t)$ is an increasing function of t and a family of link functions, with $g_\rho(\cdot)$ the same as $g_c(\cdot)$. The class of semiparametric models have been developed for handling interval censored data [79].

Additionally, a class of *semiparametric transformation models* with censored data [80] were developed to include both proportional hazards and proportional odds models with random effects [45, 81–83]. The class of models are closely connected to the class of spline-based survival models [84] and the class of semiparametric survival models [77].

In addition to the class of regression models 2.14 on the survival scale, there is also a class of models for the cumulative incidence functions within the competing risk setting. Fine and Gray [85] introduced the concept of the subdistribution hazard (termed *Fine and Gray competing risk regression model*), which can be connected to the cause-specific cumulative incidence functions. Furthermore, the class of absolute risk regression models on the cumulative incidence functions have been proposed, such as using penalized estimation methods [86], with high-dimensional covariates [87], for interval-censoring data [88].

2.2.3 Multistate models

For subjects who experience multiple events (or states) since study entry, *multi-state models* with non-parametric estimation methods can be applied. In Paper IV, we applied multistate models to estimate time-dependent risks of subsequent outcomes (or different states) from different initial states. For illustration, assume that a stochastic process $\{X(t) : t \geq 0\}$ is a continuous-time Markov process, with a finite state space $D_s = \{0, 1, \ldots, p\}$. Transition probabilities can be defined as [89]

$$P_{ij}(t, t+\mathrm{d}t) = \mathrm{Prob}(X(t+\mathrm{d}t) = j | X(t) = i) \tag{2.16}$$

and transition hazards can be defined as

$$h_{ij}(t) = \lim_{\Delta t \to 0} \frac{P_{ij}(t, t+\mathrm{d}t)}{\Delta t}$$

Furthermore, the transition probabilities $P(0,t)$ from an initial state i to a subsequent state j over a time period $(0,t)$ is a more interesting measure, and can be the solution to

the Kolmogorov forward differential equations [89]

$$\frac{\partial P_{ij}(0,t)}{\partial t} = P_{ij}(0,t)h_{ij}(t)$$

based on the Chapman-Kolmogorov equations $P_{ij}(0,t) = P_{ik}(0,u)\,P_{kj}(u,t)$, where $0 < u < t$ and $k \in D_s$. One solution to the above equation in the matrix form is the Aalen-Johansen estimators [89]

$$\widehat{\mathbf{P}}(0,t) = \prod_{0<t_k\leq t} (\mathbf{I} + \Delta\widehat{\mathbf{H}}(t_k) \tag{2.17}$$

where the cumulative translation hazard function $H_{ij}(t)$ can be estimated by the Nelson-Aalen estimator $\widehat{H}_{ij}(t) = \int_0^t \frac{\mathrm{d}N_{ij}(u)}{Y_i(u)}$ and let $\widehat{H}_{ii}(t) = -\sum_j \widehat{H}_{ij}(t)$ [89], where $N(\cdot)$ is a corresponding counting process and $Y(\cdot)$ is the at risk indicator (see [40]).

2.3 Full likelihood-based estimation methods

A partial likelihood function [7] only involves the functional forms of covariate effects, canceling out the information on the baseline functions of time. By contrast, a full likelihood function often includes specified functional forms for both baseline hazard functions (or other transformed basis functions of time) and covariate effects (e.g. time-dependent and non-linear effects).

2.3.1 Maximum (full) likelihood estimation

For right-censored survival data, under the assumption of non-informative censoring 2.1, a full log-likelihood of β can be formulated as follows [8]

$$\log L(\beta) = \sum_{i=1}^{n} \{\delta_i \log h(t_{oi}|\mathbf{z}_i;\beta) + \log(S(t_{oi}|\mathbf{z}_i;\beta))\}, \tag{2.18}$$

which involves hazard functions only for subjects with the occurrence of an event, and survival functions for all individuals.

The negative log-likelihood function $-\log L(\beta)$ can be set to an objective function for estimating optimal model parameters β, which is then converted to a mathematical optimization problem of finding the optimal β in a feasible parametric space. For further details on maximum likelihood estimation, see [90, 91]. Similarly, for correlated data, the related marginal likelihood function [90] can be derived from integrating out random effects for each cluster.

2.3.2 Maximum penalized log-likelihood estimation

For uses of penalized regression splines, such as P-splines by Eilers and Marx [92] and all penalized regression splines implemented by Wood [2], the corresponding maximum penalized likelihood estimation methods can be applied [22, 32, 40]. In this type of estimation method, the maximum degrees of freedom for each smooth regression component is prespecified, and then an effective degrees of freedom [22, 32, 93] for each

smooth regression component can be automatically estimated through this statistical estimation procedure.

The penalized log-likelihood function can be constructed to be the logarithm of the full likelihood function 2.18 minus a penalty term for each smooth function. For example, based on a spline-based proportional hazards model with only one smooth function of t, the corresponding penalized log-likelihood can be formulated as

$$\log L_p(\beta|\lambda) = \log L(\beta) - \frac{1}{2}\lambda P(\beta) \tag{2.19}$$

where $P(\beta)$ is the roughness penalty function and λ is the smoothing parameter to control the smoothness of a fitted function.

Let $\mathcal{H}_{pl}$ and $\mathcal{H}_l$ be the Hessian matrices of the corresponding penalized log-likelihood function 2.19 and log-likelihood function 2.18 evaluated with the same estimates, respectively. *pl* denotes the related Hessian matrix is derived from the penalized log-likelihood function, and *l* denotes the Hessian matrix is derived from the log-likelihood function. Based on both Hessian matrices, the corresponding effective degrees of freedom for each smooth regression component can be extracted from the total degrees of freedom [22, 40] for a proposed model, which is given as

$$EDF_{\lambda}(\text{model}) = \text{Trace}(\mathcal{H}_{pl}^{-1}\mathcal{H}_l).$$

Similarly, for correlated data, the related penalized marginal likelihood estimation procedures for spline-based flexible parametric models [94] can be applied to estimate model parameters and the variance of random effects.

2.4 Survival models as dynamic regression models

2.4.1 Modeling components

As demonstrated in the above parametric models 2.11 and 2.12, a proposed survival model can generally be decomposed as two additive terms: (1) the *background component* to be a function of time for the reference subgroup; and (2) the *comparing component* to be linearly combined relative ratios or absolute differences between comparison subgroups and the reference subgroup. Martinussen and Scheike [95] proposed the concept of *dynamic regression models* for time-to-event data. Younes and Lachin [71] earlier proposed to take into account both effects of time t and covariates $\mathbf{z}$, where the effect of time is a transformed baseline survival function.

The idea can be illustrated through the well-defined multiplicative and additive hazards modeling (e.g. Cox regression and Aalen's additive hazards models). The Cox regression [7] can be presented in an additive form, with the condition of $h_0(t) > 0$,

$$\log h(t|\mathbf{z}) = \log h_0(t) + \beta_{ph}^T \mathbf{z},$$

in which the term on the right-hand side was termed the *linear predictor* by Andersen and Skovgaard [96]. It is clear that the multiplicative effects of covariates on the baseline hazard function could be converted to the additive effects of covariates on the

log-transformed baseline hazard function with some conditions. The Aalen's additive hazards model [97] can be given as

$$h(t|\mathbf{z}) = h_0(t) + \beta_{ah}(t)^T \mathbf{z}.$$

The main differences between the two classical models are: (1) the link functions for baseline hazards, such as the log and identity links; and (2) the functional forms of covariate effects, such as either time-constant or time-dependent effects, which can be interpreted as log-hazard ratio and hazard difference, respectively.

2.4.2 Specifications

2.4.2.1 Modeling baseline hazard function of *t*

In practice, multivariate regression models are commonly applied to involve multiple prognostic factors or investigate the effect of a treatment with adjustment for multiple potential confounders (or balance the difference between compared groups at baseline). In this context, the baseline hazard function is conditional on multiple baseline variables being zero (for categorical variables) or the empirical average (for continuous variables). It may be not clear whether standard parametric models (e.g. the Weibull or log-logistic parametric baseline hazards) are appropriate for the specific reference group.

Alternatively, spline-based representations and non-parametric methods can be applied. For larger survival data (or at the beginning of a follow-up study), it is desirable that there is little difference between spline-based methods and non-parametric approaches. However, for a cohort study with a small or moderate sample size, both approaches could be influenced by the fact that there are commonly limited data at the end of follow-up. For non-parametric methods, it is usual to suppose that there are constant effects over some intervals without the occurrence of an event; in this context, spline-based regression models also only provide a data-driven estimation, since the continuous assumption of survival times will be valid during that period. It is noteworthy that the related estimates of baseline functions of t and time-dependent effects of covariates should be interpreted carefully at the end of follow-up.

Scheike proposed to set the maximum time (or a threshold point) as an argument in the R function *Gprop.odds* for fitting a generalized semiparametric proportional odds model [95, 98]. The idea could be very useful to gain stable estimates of regression coefficients and especially the baseline hazard function and time-dependent effects.

2.4.2.2 How to model covariate effects

With empirical data (a random sample from a population), ideally the final selected survival regression model should use appropriate functional forms for all given baseline covariates. However, when proposing a statistical model for empirical survival data, several statistical issues are naturally raised. For example, the question of how to specify the effects of categorical variables, which may be time-constant or time-dependent [64, 99]. The question of how to model functional forms for continuous covariates in observational studies, investigating log-linear, step functional or nonlinear relationships [33, 100–104]; joint time-dependent and non-linear effects for age have

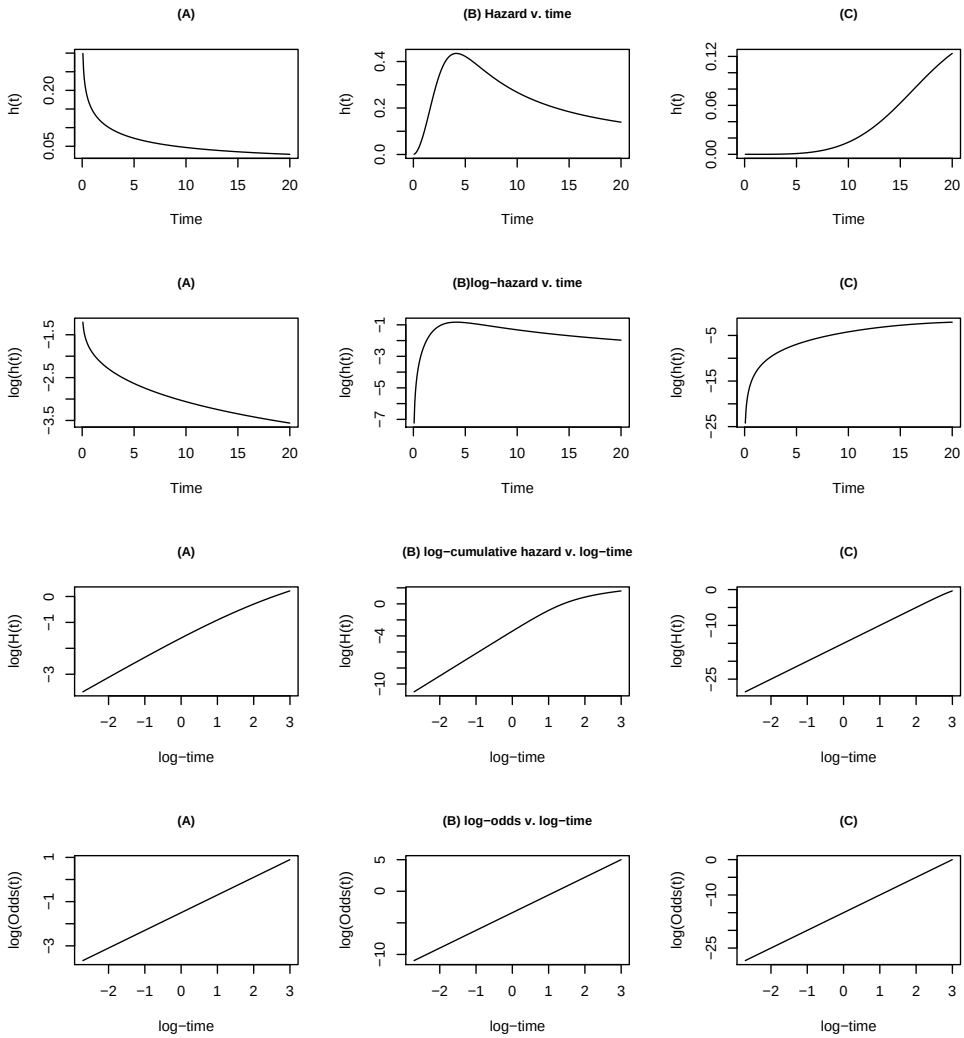
been considered by Abrahamowicz and colleagues [34] in the all-cause survival setting, and by Remontet and colleagues [35] in the relative survival setting. All these issues were noted in the international collaborative *Strengthening analytical thinking for observational studies* (STRATOS) project, which was initiated and led by Willi Sauerbrei [33].

In data analysis, it would be desirable to reduce the impact of misspecification of functional forms for covariate effects. For example, the proportional assumption may not be valid due to multiple reasons: (1) supposing a full model is a proportional hazards model, Therneau and Grambsch [105] provided an example to demonstrate that the hazards ratio varies over time due to omitting an important independent variable, see [106] for the case of omitting confounders; and (2) due to a changing biological effect of an exposure [76, 107].

In this context, the different choice of link functions will have the result that covariate effects with different interpretation, such as in measures of either relative ratios or absolute differences. Furthermore, the transformations [96, 108] by different links may make the transformed baseline functions more linear over the follow-up period (or transformed times). For illustration, three types of baseline hazard functions: (A) with a decreasing shape; (B) with a unimodal shape; and (C) with an increasing shape are presented in the upper panel of Figure 2.1. Based on these three baseline functions, the corresponding baseline cumulative hazard functions and baseline odds can be derived against original time (or log-transformed time).

In general, when using model-based methods in survival analysis, three issues need to be considered: (1) which link function should be used, which will result in regression coefficients in measures of either different relative ratios (e.g. log-hazards ratio, log-odds ratio) or absolute differences (e.g. hazard differences [109]); (2) whether or how to specify baseline hazard functions, which determines the choice of parametric or spline-based flexible parametric models; and (3) how to model covariate effects, especially for continuous covariates [33].

Figure 2.1: Three types of baseline hazard functions presented in the top panel: (A) with a decreasing shape, (B) with a unimodal shape, and (C) with an increasing shape. Based on these three baseline hazard functions, the corresponding log-hazard v. time given in the second row, log-cumulative hazard v. log-time in the third row, and log-odds v. log-time presented in the bottom panel, respectively.



3 Aims

The overall aim of the four studies was to provide a rich and coherent framework for modeling independent and correlated time-to-event data for medical research. More specifically, the aim of each study was:

I. To refine, extend GSMs and apply proposed approaches to independent survival data, with either time-to-death due to any cause or time-to-any occurrence of a disease as the outcome of interest.

II. To extend GSMs for correlated time-to-event data, with applications to simulated and real data sets.

III. To extend GSMs for population-based cancer survival data analysis, with a comparison of knot-based and knot-free regression splines; parametric and penalized estimation procedures under a penalized likelihood framework.

IV. To investigate patterns of prostate-specific antigen (PSA) testing and subsequent outcomes, partly using the proposed methods in Paper II.

These newly refined model components and extended GSMs were to be implemented and integrated into the *rstpm2* package in R [62].

4 Refinement, extension and assessment

The class of GSMs [1] were originally in the form of

$$g\{S(t|\mathbf{z})\} = g\{S_0(t)\} + \beta_z^T \mathbf{z} \tag{4.1}$$

where g is a specified link function, β_z denotes unknown regression coefficients, $\mathbf{z}$ are a vector of baseline covariates, and $S_0(t)$ is the baseline survival function defined for the reference group.

4.1 Understanding components of GSMs

4.1.1 View of the link functions from statistical perspectives

The choice of a specific link function g in GSMs 4.1 can be related to the specification of a cumulative distribution for random error in the following semi-parametric transformation models [77, 80]

$$T_f(Y) = \mu + \beta_z^T \mathbf{z} + \sigma\varepsilon \tag{4.2}$$

where $T_f(\cdot)$ denotes a monotonic increasing transformation function, Y is the time random variable T (or $\log(T)$), μ is an unknown location parameter, σ is an unknown scale parameter, ε is a standard random error with $E(\varepsilon)$=0, $Var(\varepsilon)$=1, and β is the column vector of regression coefficients.

There are two unknown functions: the transformation function $T_f(\cdot)$ and the cumulative probability distribution of ε. In this context, at least one assumption must be made for estimation. In general, there are several options for estimating model parameters:

(1) Set two parametric forms for both functions: $T_f(\cdot)$=$\log(\cdot)$ and multiple specific parametric distributions for ε, such as normal and logistic distributions. In this setting, it can be identified as the class of classical parametric AFT models 2.10 [11];

(2) Only set $T_f(\cdot)$ to be the log-transformation function, but represent a cumulative distribution function of ε by a flexible smooth functions, such as a truncated series expansion [110] or splines [111];

(3) Only set a specific cumulative distribution for ε, but leave the transformation $T_f(\cdot)$ as a flexible function, such as a well-defined monotonic I-spline or common regression splines with a monotonic condition.

For instance, based on a monotonic increasing spline function $s(\cdot)$, we can make a theoretical connection between 4.1 and 4.2. By definition, survival function of t (or $\log(t)$) given baseline covariates $\mathbf{Z}$ can be expressed as

$$\begin{aligned} S(y|\mathbf{z}) = P(Y > y|\mathbf{z}) &= P(s(Y) > s(y)|\mathbf{z}) \\ &= P(\mu + \beta_z^T \mathbf{z} + \sigma\varepsilon > s(y)|\mathbf{z}) \\ &= 1 - P(\mu + \beta_z^T \mathbf{z} + \sigma\varepsilon \leq s(y)|\mathbf{z}) \end{aligned} \tag{4.3}$$

$$= 1 - F_\varepsilon\left(\frac{s(y) - \mu - \beta_z^T \mathbf{z}}{\sigma}\right)$$

where $y = t$ or $\log(t)$, $F_\varepsilon(\cdot)$ is a standard cumulative probability function of ε. One commonly used distribution of ε is the standard Gumbel (minimum) distribution with $F_\varepsilon(\cdot) = 1 - \exp(-\exp(\cdot))$. The baseline survival function is then in the form of $S_0(t) \overset{\text{Gum.}}{=} \exp\{-\exp(s_\varepsilon^*(y))\}$, where the flexible spline representation $s_\varepsilon^*(y)$ absorbs two naive parameters μ and σ. That is, the corresponding transformation function is the log$-$log link, and then the transformed survival function is given as

$$\log\{-\log(S(t|\mathbf{z}))\} = \log\{-\log(S_0(t))\} + \beta_{ph}^T \mathbf{z},$$

where σ is absorbed into $\beta_z^T \mathbf{z}$ by the definition of $\beta_{ph}^T \mathbf{z} \equiv -\frac{\beta_z^T \mathbf{z}}{\sigma}$. The above model can be treated as the classical proportional hazards model on the survival scale with the condition $h_0(t) > 0$.

Similarly, the standard logistic distribution for ε corresponds to the proportional odds model and the corresponding link function is $-\text{logit}(\cdot)$,

$$\log\left\{\frac{1 - S(t|\mathbf{z})}{S(t|\mathbf{z})}\right\} = \log\left\{\frac{1 - S_0(t)}{S_0(t)}\right\} + \beta_{po}^T \mathbf{z}.$$

More specifically, the standard exponential distribution for ε can be identified for an additive hazard model on the survival scale with extended time-dependent coefficients,

$$-\log\{S(t|\mathbf{z})\} = -\log\{S_0(t)\} + \beta_{ah}(t)^T \mathbf{z} = H_0(t) + \beta_{ah}(t)^T \mathbf{z},$$

the corresponding link function is $-\log(\cdot)$ with the conditions of both baseline cumulative function $H_0(t)$ and time-dependent effects $\beta_{ah}(t)$ passing through $(0,0)$.

4.1.2 Build GSMs using a linear predictor function

Under the model framework 4.1, functional forms of effects can be expressed in an additive manner (see 2.4.1 and 4.1.1) on a transformed baseline function (e.g. log-cumulative baseline hazard function and log-baseline odds function), so it is possible to adopt a linear predictor function [96] for GSMs. For example, univariate smooth regression components can be expressed as a linear combination of spline basis functions.

In general, the transformed survival function can be expressed in a linear predictor in the matrix form of

$$g\{S(t|\mathbf{z})\} = \eta(t, \mathbf{z}; \beta) = \mathbf{X}(t, \mathbf{z})\beta \tag{4.4}$$

where β is a column vector of parameters and $\mathbf{X}(t, \mathbf{z})$ is a design matrix (usually including the first column of ones for the constant term) with rows for each subject and with columns to include observations of corresponding predictors (or transformed predictors by spline basis functions), with the sum-to-zero constraint for each additive smoother excluding time-dependent effects.

Suppose $G\{\cdot\}$ to be the inverse link functions of $g\{\cdot\}$. Survival and hazard functions then can be rewritten as

$$\left.\begin{aligned} S(t|\mathbf{z};\beta) &= G\{\eta(t,\mathbf{z};\beta)\} \\ h(t|\mathbf{z};\beta) &= -\frac{G'\{\eta(t,\mathbf{z};\beta)\}}{G\{\eta(t,\mathbf{z};\beta)\}}\frac{d\eta(t,\mathbf{z};\beta)}{dt} \end{aligned}\right\} \tag{4.5}$$

with

$$\frac{d\eta(t,\mathbf{z};\beta)}{dt} = \lim_{\Delta t\to 0}\frac{\mathbf{X}(t+\Delta t,\mathbf{z})-\mathbf{X}(t-\Delta t,\mathbf{z})}{2\Delta t}\beta = X_D(t,\mathbf{z})\beta, \tag{4.6}$$

where $X_D(t,\mathbf{z})$ denotes the first derivative of the design matrix $\mathbf{X}(t,\mathbf{z})$ with respect to t, calculating by the finite difference method.

4.1.3 Select $\log(t)$ or t to match a specific link

Generalized survival models 4.4 are defined on $t \in (0,\infty)$, with unknown parameters β and the condition of $S(0|\mathbf{z}) = 1$. This implies that

$$\lim_{t\to 0} g\{S(t|\mathbf{z})\} = \lim_{t\to 0}\eta(t,\mathbf{z}) = -\infty,$$

when the link function is specified as either the log$-$log link or the $-$logit link. The above formula holds, for example, if $\eta(t,\mathbf{z})$ is a polynomial function of $\log(t)$ with corresponding bounded parameters. For example, a GSM with the log$-$log link can involve a transformed survival baseline function in the form of $\log-\log\{S_0(t)\} = s(\log(t))$ or a time-dependent effect as $\beta(\log(t))$.

However if g is specified as the $-$log link, that is

$$\lim_{t\to 0}\log\{S(t|\mathbf{z})\} = \lim_{t\to 0}\eta(t,\mathbf{z}) = 0,$$

which is valid when $\eta(t,\mathbf{z})$ is to be a polynomial function of time t (or $\sqrt{t}$) with bounded parameters, e.g. $H_0(t) = s(t)$ or $s(\sqrt{t})$.

4.1.4 Connecting smooth functions to regression splines

Various flexible smoothing techniques [33] have been used to explore non-linear relationship between covariates and outcomes of interest, with linear relationship as a special case. These potentially non-linear relationship are unknown, but can be assumed as smooth functions that are related to regression splines (e.g. truncated power basis functions or radial basis functions).

For example, consider the ordered sequence, $a < x_1 < x_2 < \cdots < x_N < b$, as a random sample of the continuous variable x with N observations. Suppose a smooth function of x (i.e. $f(x)$) is differentiable four times with respect to x over its domain (a,b), with a value $c \in (a,b)$. Mathematically, the explicit representation of $f(x)$ with a integral form for reminder function [112] can be expanded at c by Taylor's theory in the forms of

$$f(x) = f(c) + \sum_{k=1}^{3}\frac{f^{(k)}(c)}{k!}(x-c)^k + \int_c^x \frac{f^{(4)}(u)}{k!}(x-u)^3\mathrm{d}u$$

$$= f(c) + \sum_{k=1}^{3} \frac{f^{(k)}(c)}{k!}(x-c)^k + \int_c^b \frac{f^{(4)}(u)}{k!}(x-u)_+^3 \, \mathrm{d}u, \tag{4.7}$$

where (k) denotes $f(x)$ is k times differential. By the Riemann integral method, one can approximate the last integral form in the formula 4.7 that depends on the location of c. In order to use of all the information on the given observations, suppose $c < x_1$ (for specifying spline knots over the whole range of measured values **x**). In this context, $f(x)$ can be approximated and rearranged to be in the form of

$$f(x) \approx \sum_{j=0}^{3} \beta_{0j} x^j + \sum_{i=1}^{N} \beta_{1i} (x - x_i)_+^3,$$

with $N+4$ unknown parameters in this equation. Extra conditions are required for estimation with only N values.

As suggested by Stone and Koo [60], and further derived by Durrleman and Simon [20], restricted cubic splines can be imposed to be linear in the tails ($x < x_1$ and $x > x_N$). These constraints imply that $\beta_{02} = \beta_{03} = 0$, $\sum_{i=1}^{N} \beta_{1i} = 0$ and $\sum_{i=1}^{N} \beta_{1i} x_i = 0$. The smooth function is then be approximated by

$$f(x) \approx \beta_{00} + \beta_{01} x + \sum_{i=1}^{N} \beta_{1i} (x - x_i)_+^3 \quad \text{subject to} \sum_{i=1}^{N} \beta_{1i} = \sum_{i=1}^{N} \beta_{1i} x_i = 0, \tag{4.8}$$

in which there are $N+2$ parameters with two conditions.

Furthermore, based on the relationship between truncated basis functions and radial basis functions

$$(x - x_i)_+^3 = \frac{|x - x_i|^3 + (x - x_i)^3}{2}$$

and those boundary constraints, $f(x)$ can be re-expressed in the notation of radial basis functions, with new rearranged coefficients (say β) in the form of

$$f(x) \approx \beta_1 + \beta_2 x + \sum_{i=1}^{N} \beta_{i+2} |x - x_i|^3 \quad \text{subject to} \sum_{i=1}^{N} \beta_{i+2} = \sum_{i=1}^{N} \beta_{i+2} x_i = 0. \tag{4.9}$$

It is clear to see that: (a) the formula 4.8 is a general version of the formula (4) provided by Durrleman and Simon in [20], but with N available observations; and (b) the formula 4.9 is the full univariate thin plate spline described in [113], which is based on cubic radial basis functions. Both of them are restricted cubic splines.

In practice, it is preferable to adopt an approximated function $s(x)$ that lies in a finite dimensional linear space. For example, $s(x) = \beta_1 + \beta_2 x$, with the remainder part (e.g. $\sum_{i=1}^{N} \beta_{i+2} |x - x_i|^3$ in 4.9) to be zero. Generally, based on few basis functions, regression splines can be used for estimating smooth regression components. For example,

$$s(x) = \begin{cases} \sum_{i=1}^{k} \beta_{1i} B_{1i}(x) & \text{without linear form} \\ \\ \beta_1 + \beta_2 x + \sum_{i=1}^{k-2} \beta_{i+2} B_i(x) & \text{involving constant and linear forms,} \end{cases}$$

where β and β_1 are spline coefficients estimated through likelihood-based approaches, and both $B_1(x)$ and $B(x)$ denote basis functions. Based on all observations of x, a regression spline including constant and linear forms can be expressed by either knot-based or knot-free bases in the matrix form as

$$s(\mathbf{x}) = \mathbf{B}(\mathbf{x})\beta = \begin{pmatrix} 1 & x_1 & B_3(x_1) & \dots & B_k(x_1) \\ 1 & x_2 & B_3(x_2) & \dots & B_k(x_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & B_3(x_N) & \dots & B_k(x_N) \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{pmatrix} \tag{4.10}$$

This implies that an estimated smooth function $s(x)$ can be represented by a sum of weighted basis functions, with coefficients β to be estimated by likelihood-based estimation methods, which can be illustrated in the following graph [114].

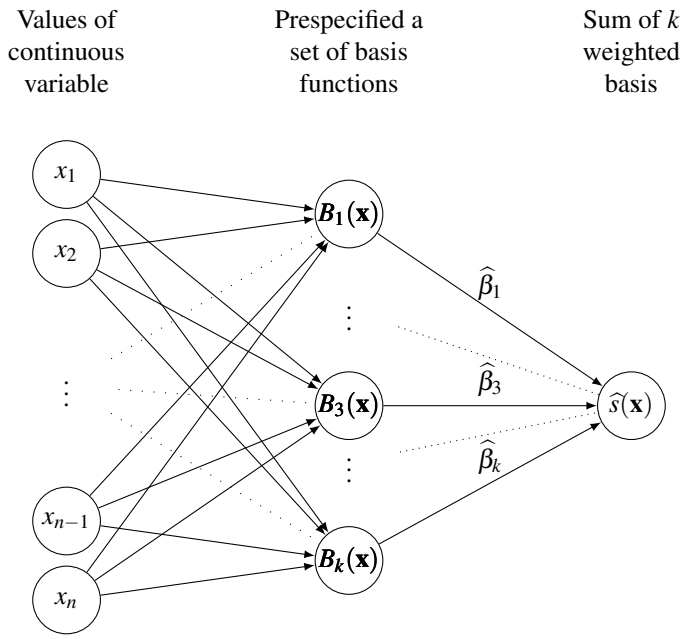


Figure 4.1: A estimated smooth function as a sum of k weighted basis functions

4.2 Extensions of GSMs

4.2.1 Inclusion of knot-based and knot-free regression splines

Generally, based on either knot-based or knot-free basis functions, regression splines can be used for representing smooth regression components in a proposed GSM.

For example, using knot-based approaches for pure regression splines, one needs to pre-specify the number of knots or spline basis functions (e.g. $\{3,4,5,6,7\}$) for each smooth regression component, then separately construct knot-based pure regression splines, e.g. the use of restricted cubic splines [20, 30] to approximate 4.8. Finally, the maximum likelihood estimation method can be applied to estimate corresponding model parameters.

By contrast, for penalized regression splines with knot-free bases [2], one only needs to set the maximum degrees of freedom for each smooth regression component, and then produce low-rank penalized regression splines by the following truncated eigen decomposition method [36]. One feature of this type of spline basis functions is that they do not depend on any spline knots, e.g. the thin plate regression splines implemented for generalized additive models by Wood [2].

For illustration, a univariate thin-plate spline function [113] of time t can be identify to be 4.9 in the form of

$$s(t) = b_0 + b_1 t + \frac{1}{12}\sum_{i=1}^{n} \delta_i |t - t_i|^3 \quad \text{subject to} \sum_i \delta_i t_i = \sum_i \delta_i = 0$$

where t_i indicates observed time for each subject i (n is the total number of subjects), with $t_1 < t_2 < \cdots < t_n$. b_0, b_1 and δ_i are spline coefficients that specify $s(t)$. Since the cubic spline $s(t)$ is represented by many basis functions, it is preferable to use fewer basis functions.

Based on all observations of t, a low rank smoother [36] can be achieved by the truncated eigndecomposition method [37]. Suppose $\mathbf{s}= (s(t_1), s(t_2), \ldots, s(t_n))^{\mathsf{T}}$ that can be expressed in the form of

$$\begin{aligned}
\mathbf{s} &= E\,\delta + T^{\mathsf{T}}\mathbf{b} \\
&= U\,D\,U^T\,\delta + T^{\mathsf{T}}\mathbf{b} && (4.11)\\
&\approx U_k D_k\,\delta_k + T^{\mathsf{T}}\mathbf{b} && (4.12)\\
&= U_k D_k Q_{k-2}\,\delta_{k-2} + T^{\mathsf{T}}\mathbf{b}, && (4.13)
\end{aligned}$$

under three conditions: (1) the equivalent 4.11 bases on the eigendecomposition of the symmetric matrix E with $E_{ij} = |t_i - t_j|^3/12$, i, j=1,2,...,n [113]; (2) the approximation 4.12 holds, with eigendecomposition approximation of E by its principal k components of eigenvectors and eigenvalues $(U_k D_k U_k^T)$, where $\delta = U_k \delta_k$ with U_k to be the first k orthogonal eigenvectors in the eignvectors matrix U; and (3) the formula 4.13 absorbs the linear constraints of $T\delta = 0$, where Q_{k-2} is the last k-2 columns of the orthogonal factor Q with $U_k^T\,T^T$=QR and T is the $2 \times n$ matrix,

$$T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_n \end{pmatrix}.$$

See [2, 37] for further details and multivariate splines.

Similarly to principal component analysis, the principal eigenvector of the basis matrix E with the linear constraints in the tails are converted to the k-2 new basis functions PC=$U_k D_k Q_{k-2}$, which do not depend on any knots. Combined the constant and linear terms, the cubic spline $s(t)$ evaluated at all observations can be calculated as

$$\mathbf{s} \approx (PC_1, PC_2, \ldots, PC_{k-2})\boldsymbol{\delta}_{\mathbf{k-2}} + \mathbf{T}^{\top}\mathbf{b}.$$

Comparing to knot-based regression splines, this type of knot-free regression splines can: (1) avoid the knot placement issue; (2) model not only univariate functions but also multivariate splines; and (3) be a linear combination of $k-1$ known basis functions that are nested with the set of k basis functions. To improve the numerical stability of the final basis functions, Wood proposed to reset the mean square size of each column of the corresponding basis matrix to be unity [115]. For identifiability, the sum-to-zero constraint for all additive functional forms is commonly imposed in the process of estimation.

Note that the thin plate regression spline basis can be a type of global polynomial functions with local features [2]. For example, fractional polynomials [116, 117] and the knot removal approach proposed in [61] can be (or produce) global polynomials. In the appendix A.3, an implementation in R for a low-rank smoother has been made available.

4.2.2 Introduction of generalized additive functional forms

Within the framework of GSMs, most potential functional forms for categorical variables can be fitted, such as time-constant effects, time-dependent effects, categorical-by-categorical and continuous-by-categorical interaction terms.

In this thesis, we have introduced mature regression spline-based smoothing techniques, e.g. penalized regression splines developed for generalized additive models by Simon Wood [2], for the effects of continuous covariates. For example, the following functional forms can be fitted within the extended GSMs:

- ◇ linear forms,
- ◇ non-linear forms (e.g. represented by regression splines),
- ◇ categorizations,
- ◇ continuous-by-continuous interaction terms (e.g. tensor product [94]),
- ◇ multivariate regression splines [94],
- ◇ joint time-dependent and non-linear effects for age [94].

4.3 Introduction of a penalized likelihood framework for GSMs

Regarding the uses of pure and penalized regression splines, there are two estimation approaches to control the amount of smoothness of smooth regression components: (1) within *maximum likelihood estimation* methods, models first need to be fitted with fewer spline basis functions (e.g. $k \in \{3,4,5,6,7\}$), then the better model selected, with an

optimal k, according to the values of AIC; (2) within *maximum penalized estimation* methods, the maximum degrees of freedom need to be set for each smooth component and the *effective degrees of freedom* for each smooth component can be automatically selected by an AIC-like criterion through the so-called smoothing parameters. These two methods can be treated as two separate statistical procedures. However, we can also realize both estimation approaches under the following penalized likelihood framework.

Given n observations in from of $(u, \delta, \mathbf{z})$ and a proposed model (i.e. the model 4.1 with one smooth function of t), a penalized likelihood function can be formulated as

$$\begin{aligned} \log L_p(\beta|\lambda) &= \sum_{i=1}^{n} \{\delta_i \log\{h_i(u_i|\mathbf{z}_i;\beta)\} + \log\{S_i(u_i|\mathbf{z}_i;\beta)\}\} - \frac{\lambda}{2}\beta^T \mathbf{S}\beta \\ &= \sum_{i=1}^{n} \{\log L_i(\beta) - \frac{\lambda}{2n}\beta^T \mathbf{S}\beta\} \end{aligned} \tag{4.14}$$

where, $\log L_i(\beta)$=$\delta_i \log\{h_i(u_i|\mathbf{z}_i;\beta)\} + \log\{S_i(u_i|\mathbf{z}_i;\beta)\}$ is the full log-likelihood function, the penalty matrix $\mathbf{S}$=$\int_\Omega B''(t)^T B''(t)dt$ is a predefined positive semi-definite matrix. For example, for penalized B-splines (or P-splines [92]), it can be a difference parameter of the corresponding smooth function. λ is a unknown smoothing parameter that controls the amount of roughness of the smooth regression component. $S_i(u_i|\mathbf{z}_i)$ and $h_i(u_i|\mathbf{z}_i)$ are the related survival and hazard functions for each patient, given $\mathbf{z}_i$, and can be calculated from the equations 4.5. Note that the penalized log-likelihood function is conditional on the smoothing parameter λ for each smooth function.

Under the penalized likelihood framework 4.14, given the link function g and $\lambda = d$ ($d \geq 0$), define the objective function as $-\log L_p(\beta|d)$, then the M-estimator of β satisfies [118]

$$\frac{1}{n}\sum_{i=1}^{n} U_i(\beta|d) = \frac{1}{n}\sum_{i=1}^{n} \left\{ -\frac{\partial \log l_i(\beta)}{\partial \beta} + \frac{d}{n}\mathbf{S}\beta \right\} = 0. \tag{4.15}$$

That is the maximum penalized log-likelihood estimators (β) become the solution to the following optimization problem

$$\beta = \arg\min_{\beta} \{-\log L_p(\beta|d)\}$$

with the constraint of $h_i(u_i|\mathbf{z}_i;\hat{\beta}) > 0$ for each patient i [84, 94].

4.3.1 Separate procedures for estimation and model selection

Set d=0 to be the extreme case in the formula 4.15, that is, the second term in the above penalized likelihood function 4.14 becomes zero. The corresponding estimation approach then changes to the maximum likelihood estimation approach for unknown model parameters including spline coefficients, but vary the number of spline basis functions for representing each smooth function. A quasi-Newton algorithm can be used to estimate model parameters.

The strategy is computationally efficient for a model fit, but one still needs to choose the optimal number of spline basis functions (or degrees of freedom) from multiple fitted models by an information criterion (e.g. AIC), which may be complicated for a proposed model with multiple smooth regression components.

4.3.2 Integrated procedure for estimation and model selection

Ideally, an optimal value ($d = \lambda^{opt}$) for each smooth function can be automatically derived from any given data by minimizing the modified likelihood-based leave-one-out cross-validation criterion (LCV [32, 93, 119]),

$$\text{LCV}(\lambda) = -\sum_{i=1}^{n} \log l_i(\hat{\beta}(\lambda)) + \text{Trace}\{\mathcal{H}_{pl}^{-1}\{\hat{\beta}(\lambda)\}\mathcal{H}_l\{\hat{\beta}(\lambda)\}\},$$

where $\hat{\beta}(\lambda) = \arg\min_{\beta}\{-\log L_p(\beta|\lambda)\}$, $\mathcal{H}_{pl}$ and $\mathcal{H}_l$ are the Hessian matrices of the corresponding penalized log-likelihood and full log-likelihood functions, respectively; and $\text{Trace}(\mathcal{H}_{pl}^{-1}\mathcal{H}_l)$ is the total degrees of freedom for a proposed model, which depends on the smooth parameters. We select λ^{opt} that minimize $\text{LCV}(\lambda)$.

In general, given a link function, the process of model selection is continuous by varying the smoothing parameter that can be automatically optimized in the penalized estimation approaches. Furthermore, the optimally chosen smoothing parameters lead to an effective degrees of freedom, being a value between one and the given maximum degrees of freedom, for each smooth regression component. Hence, in the process of performing penalized estimation methods, we do not need to re-choose the optimal one from multiple model fits.

For example, if there are m smoothers in a proposed model, the number of parameters in linear terms is p and q for all smoothers, then the total degrees of freedom can be calculated as [40]:

$$\begin{aligned}
\text{EDF}_\lambda = \text{trace}(\mathcal{H}_{pl}^{-1}\mathcal{H}_l) &= \text{trace}(\mathcal{H}_{pl}^{-1}(\mathcal{H}_{pl} - \text{S}_\lambda)) \\
&= \text{trace}(\mathcal{H}_{pl}^{-1}\mathcal{H}_{pl}) - \text{trace}(\mathcal{H}_{pl}^{-1}\text{S}_\lambda)) \\
&= p + q - \text{trace}\left(\frac{\text{S}_\lambda}{\mathcal{H}_l + \text{S}_\lambda}\right)
\end{aligned}$$

where $\mathcal{H}_{pl}$ indicates $\mathcal{H}_{pl}(\hat{\theta}(\lambda))$, $\mathcal{H}_l$ indicates $\mathcal{H}_l(\hat{\theta}(\lambda))$, $\lambda = (\lambda_1, \ldots, \lambda_m)$, and

$$\text{S}_\lambda = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & \lambda_1\text{S}_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m\text{S}_m \end{pmatrix}$$

Given the optimal smoothing parameters, three properties can be provided:
1) When $\lambda^{opt} \to \mathbf{0}$, $\frac{\text{S}_\lambda}{\mathcal{H}_l + \text{S}_\lambda} \to \mathbf{0}_{p+q,p+q}$, then trace $\left(\frac{\text{S}_\lambda}{\mathcal{H}_l + \text{S}_\lambda}\right) \to 0$, EDF $\to p+q$
2) When $\lambda^{opt} \to \infty$, $\frac{\text{S}_\lambda}{\mathcal{H}_l + \text{S}_\lambda} \to \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{I}_q \end{pmatrix}$, then trace $\left(\frac{\text{S}_\lambda}{\mathcal{H}_l + \text{S}_\lambda}\right) \to q$, EDF $\to p$;
3) Others: $p <$ EDF $< p+q$.
where $\mathbf{0}$ is the vector with m zeros, $\mathbf{0}_{p+q,p+q}$ is a $(p+q)$*$(p+q)$ matrix with zeros, the

Hessian matrix $\mathcal{H}_l$ is also a real symmetric matrix, and $\mathbf{I}_q$ is a unit matrix. Based on the optimal smoothing parameters (or the corresponding effective degrees of freedom for each smooth regression component), one can select $\widehat{\beta}=\arg\min_{\beta}\{-\log L_p(\beta|\lambda^{opt})\}$.

Note that log-transformation of smoothing parameters, $\upsilon = \log(\lambda)$ was applied in the model fitting, which aims to: (1) convert a non-linear constricted optimization issue with $\lambda > 0$ to a common optimization issue without any condition; (2) reduce the range of each $\lambda \in (0, +\infty)$ to a narrow range of each converted variable υ . To some extent, the log-transformation imposes the effective degrees of freedom for each smooth regression component to be away from the minimum value 1 and a larger real value [67].

4.4 Implementation and assessment

4.4.1 Implementation integrated into the R package *rstpm2*

The *rstpm2* package in R was originally created and maintained by Mark Clements with contributions from Paul Lambert, mainly for fitting *flexible parametric survival models* [30].

All refinements and extensions in this thesis have been mainly integrated into *rstpm2*. Currently, there are two main functions: (1) *stpm2* is mainly used to fit so-called *parametric GSMs*, with either parametric forms or pure knot-based regression splines for representing smooth regression components and maximum likelihood methods for parameter estimation; and (2) *pstpm2* is first implemented to fit so-called *penalized GSMs*, with penalized regression splines for modeling smooth regression components and maximum penalized likelihood methods for parameter estimation. Furthermore, as proposed in Paper III, the penalized likelihood framework can include both parametric and penalized estimation methods for pure and penalized regression splines, respectively. This means that *pstpm2* has the ability to fit both parametric and penalized GSMs.

A description of the model formula syntax can be found in the R package rstpm2 [62], which has been made available on the comprehensive R archive network (CRAN). The following is the basis of the implementation:

(1) Firstly, construct likelihood components with all information on observations in matrix forms, such as the design matrix $\mathbf{X}(t, \mathbf{z})$, its first derivative matrix $X_D(t, \mathbf{z})$ with respect to time t, and a predefined penalty matrix for each smooth regression component (especially for penalized estimation methods).

(2) All optimization problems for model parameters and smoothing parameters can be solved by current mature implementation in compiled code (C/C++). The likelihood components first need to be converted from R to C++ (see [120]), then optimal model parameters and smoothing parameters are optimally estimated in C++, which would accelerate the estimation process especially for non-linear models.

4.4.2 Measures beyond the hazard ratio

For independent survival data, survival and hazard functions can be estimated based on all estimated parameters $\hat{\theta}$

$$\left.\begin{aligned} \widehat{S}(t|\mathbf{z};\widehat{\beta}) &= G\{\eta(t,\mathbf{z};\widehat{\beta})\} \\ \widehat{h}(t|\mathbf{z};\widehat{\beta}) &= -\frac{G'\{\eta(t,\mathbf{z};\widehat{\beta})\}}{G\{\eta(t,\mathbf{z};\widehat{\beta})\}}\frac{\partial\eta(t,\mathbf{z};\widehat{\beta})}{\partial t} \end{aligned}\right\} \tag{4.16}$$

where $G(\cdot)$ is the specific inverse link function, $X(t,\mathbf{z})$ is the design matrix and its first derivative $X_D(t,\mathbf{z})$ is also involved.

A: Ratios based on a estimated regression coefficient $\widehat{\beta}_{\mathbf{z}}$

For a comparison to reference group, GSMs can be expressed as

$$g\{S(t|\mathbf{z})\} - g\{S_0(t)\} = \beta_z^T\mathbf{z}, \tag{4.17}$$

within the proportional hazards model (or a GSM with the log−log link), regression coefficients β are interpreted as the related **log-hazard ratios**; but β denotes the **log-odds ratios** when a GSM with the −logit link.

More specifically, in the presence of non-proportionality, some alternative measures have been considered, e.g. the **cumulative hazard ratio** [121–123]. The measure in the form of $\beta_z(t)$ can be estimated from a GSM with the log−log link

$$g\{S(t|\mathbf{z})\} - g\{S_0(t)\} = \beta_z(t)^T\mathbf{z}, \tag{4.18}$$

where $\beta_z(t)$ can be interpreted as a cumulative effect.

B: Ratios based on all estimates $\widehat{\beta}$

In the presence of non-proportionality, a hazard ratio can be calculated from the estimated hazards for two subgroups.

For example, suppose the subgroup A with baseline covariates $\mathbf{z_a}$ and the subgroup B with baseline covariates $\mathbf{z_b}$, then the hazards ratio between the subgroup A and B is

$$HR_{ab} = \frac{\widehat{h}(t|\mathbf{z}_a;\widehat{\beta})}{\widehat{h}(t|\mathbf{z}_b;\widehat{\beta})},$$

which is based on all estimates $\widehat{\beta}$.

C: Absolute difference based on all estimates $\widehat{\beta}$

Similarly, we can estimate hazard or survival difference from the absolute measure 4.16. For the subgroup A and B comparison, the hazard difference is

$$\text{Hazard difference}_{ab} = \widehat{h}(t|\mathbf{z}_a;\widehat{\beta}) - \widehat{h}(t|\mathbf{z}_b;\widehat{\beta}),$$

and the survival difference is

$$\text{Survival difference}_{ab} = \widehat{S}(t|\mathbf{z}_a;\widehat{\beta}) - \widehat{S}(t|\mathbf{z}_b;\widehat{\beta}).$$

D: Population average (adjusted) survival based on all estimates $\widehat{\beta}$

In addition to fitting a marginal GSM (on all measured covariates) to estimate un-adjusted survival curves, adjusted survival curves have also been applied in the medical research [52] based on proportional hazards models given baseline covariates for each subject:

$$S(t) = \frac{1}{n}\sum_{i=1}^{n} S(t|\mathbf{z}_i;\widehat{\beta}) \tag{4.19}$$

Assuming the proposed model is correctly specified, all these measures based on estimated model parameters $\widehat{\beta}$ and smooth regression components, the variance-covariance matrices of functions of multiple parameters can be derived by the multivariate delta method.

Similarly, for correlated survival data, the population average survival probability can be derived as

$$S(t|\mathbf{z}) = E_b(S(t|\mathbf{z},b)),$$

which is averaged across all subjects with baseline covariates $\mathbf{z}$ and any frailty level b.

4.4.3 Assessment by comparison using simulations and examples

In this thesis, the performance of proposed methods has been assessed in two ways: (1) by comparing results from proposed methods to "truth" that is predefined to generate independent survival data sets for analysis; and (2) by analyzing real data sets to compare results from proposed methods to those from comparable approaches. At the same time, the corresponding implementation was also validated through these comparisons.

Simulation studies have been commonly applied to assess the performance of the proposed statistical methods. In general, a simulation study uses a numerical technique to mimic a specific "scenario" in which $S(t|\mathbf{z})$ is specified to include: (1) a known baseline hazard function; and (2) known functional forms of covariate effects. That is, all model parameters are predefined, i.e. the "truth". Observations of $\mathbf{Z}$ are generated from certain given distributions and are assumed to be fixed values for each data set.

Survival times (t) can be randomly generated from the survival probability function $S(t|\mathbf{z})$ using a previously proposed method [124]; censoring time c can be randomly simulated from an independent distribution, with observed times $min(t,c)$ and event indicators $I(t \leq c)$. Under these conditions, several hundred independent survival data sets are usually generated, which are analyzed by proposed methods and comparative approaches. Finally, the properties of estimates (of model parameters and smooth regression components) can be empirically summarized from all simulated data sets, usually including the bias, random error and coverage probability through comparison to the designed "truth".

Extensive simulation studies have been undertaken to assess the performance of proposed statistical methods and corresponding implementation in Study I-III. In

addition to comparing the proposed methods to predefined "truth", different strategies including the model specification and estimation approaches can be compared through the example data analysis. In Study I-III, we have demonstrated several comparisons between proposed and well-established approaches under a proportional hazards or odds model. Additionally, several novel features of GSMs have been illustrated in applications.

5 Summary of papers

Under the framework of extended GSMs, we have mainly studied several nested settings (see Table 5.1) to be able to analyze either independent or correlated survival data, with overall and net survival as measures of interest, respectively.

Table 5.1: Two types of survival data and two types of survival functions

	Overall survival (Recurrence-free survival)	Net survival (Relative survival)
Individual level $(T_o, \Delta_o)\|Z = z$ independent data	Paper I	Paper III
Cluster level $(T_{oi}, \Delta_{oi})\|Z = z$ correlated data	Paper II & IV	Future research

5.1 Paper I

In this paper, the outcome under study was time-to-death due to any cause (or time-to-any recurrence of disease); overall survival (or recurrence-free survival) will be the corresponding survival function, given baseline covariates.

Refinements and extensions of GSMs

Within the framework of GSMs in the form of

$$g\{S(t|\mathbf{z})\} = \mathbf{X}(t,\mathbf{z})\,\beta, \tag{5.1}$$

we had made several refinements and extensions:

- ◇ Reviewed link functions (e.g. log−log, −logit and −log) from a statistical perspective (4.1.1).
- ◇ Built GSMs using a linear predictor function (4.1.2).
- ◇ Specified smooth regression components of time on either a $\log(t)$ or t scale to match a specific link (4.1.3).
- ◇ Included penalized regression splines for functional forms of baseline functions and covariate effects, which are based on the available implementation (*mgcv* package in R) for fitting generalized additive models by Wood [2] (4.2.1).

◇ Introduced penalized likelihood estimation methods corresponding to penalized regression splines, with each smooth parameter optimally selected in a continuous manner for the corresponding smooth regression components (4.3).

The full log-likelihood function can be constructed as

$$\log L(\beta) = \sum_{i=1}^{n} \left\{ \delta_i \log h(t_{oi}|\mathbf{z}_i;\beta) + \log\{S(t_{oi}|\mathbf{z}_i;\beta)\} \right\} \tag{5.2}$$

where $h(\cdot)$ and $S(\cdot)$ are derived from Model 5.1. Similarly, the corresponding penalized log-likelihood function can be constructed as the $\log L(\beta)$ minus a roughness penalty function for each smooth regression component, as described in 2.19.

Note: for a comparison of different statistical procedures related to GSMs, we used parametric GSMs to indicate a statistical procedure using regression splines (e.g. natural splines based on B-spline basis functions) to represent each smooth regression component and applying maximum full likelihood methods for parameter estimation (termed *parametric procedures*); penalized GSMs are suitable for another statistical procedure with penalized regression splines (e.g. penalized thin plate regression splines based on radial basis functions) for modeling each smooth regression component and adopting maximum penalized full log-likelihood approaches to estimate all model parameters (*penalized procedures*).

Comparison of results in simulation studies

In this simulation study, a proportional hazards or proportional odds model with known baseline hazard function and regression coefficients was predefined. Then several hundred data sets were generated from this defined model using the method of Bender and colleagues [124]. Finally, we compared all results from the proposed approaches: (1) to those predefined "truth"; and (2) to established methods, e.g. estimates of regression coefficients based on partial likelihood estimation methods, and estimated smooth baseline hazard functions using smoothing splines [57] or penalized B-splines [125].

The performance of the proposed approaches, under the investigational settings, can be summarized as follows:

◇ Both parametric and penalized procedures can capture the predefined varying trends of baseline hazard function, measured by the integrated discrepancy [126] (or area between the estimated and predefined curves), the time-dependent empirical mean square error, and a probability-based symmetrized Kullback-Leibler distance [57].

◇ For a survival regression model with only one smooth regression component, both parametric and penalized procedures can provide very similar results.

◇ Both parametric and penalize procedures can provide similar estimates of regression coefficients to those from partial likelihood estimation methods.

◇ Both parametric and penalized procedures were comparable to another two established approaches: (1) using the penalized estimation method combined with

smoothing splines [57, 127]; and (2) smoothing under the framework of mixed models combined with the restricted maximum likelihood approach [58, 126], for estimating the baseline hazard function.

Comparisons and illustrations through examples

Using three examples, we compared estimates of regression coefficients from proposed methods to established approaches, with either a proportional hazards or proportional odds model. Additionally, we demonstrated model non-linear effects (or smooth regression components) for continuous covariates and time-dependent effects of categorical variables. Through these three applications, the following conclusions can be made:

- ◇ Penalized procedures based on a GSM with the log-log link can provide similar estimates of regression coefficients compared to those estimated from the Cox regression model in sections 5.2.1 and 5.3.1 (Paper I).
- ◇ Penalized procedures based on a GSM with the -logit link can also provide similar estimates of regression coefficients compared to other estimation approaches (see Table 3 in Paper I).
- ◇ In empirical data analyses, the link function could be selected by the AIC criterion, which had been demonstrated in the previous simulation and suggested in previous studies [81].
- ◇ The Cox-Snell residual plot can be used as a graphic method to assess the fit of the GSM.
- ◇ Based on a fitted GSM with either the log-log or -logit link, multiple relative or absolute measures can be provided, such as the hazards ratio, odds ratio, and survival difference.

5.2 Paper II

In this paper, the outcome of interest was clustered time-to-a specific event (or repeated event within the same subject). It was reasonable to consider that the subjects within a cluster may share some unmeasured environmental or genetic risk factors, which were commonly modeled by a random effect b (or frailty U) for each cluster.

In this context, the corresponding overall survival functions given covariates $\mathbf{Z}$ and a random effect b for each cluster can be modeled under the framework of GSMs. For instance, GSMs with baseline covariates $\mathbf{Z}$ and a normally distributed random effect b_i for cluster i can be expressed as follows:

$$g\{S(t|\mathbf{z_{ij}}, \mathbf{b_i})\} = \mathbf{X}(t, \mathbf{z_{ij}})\beta + \mathbf{b_i}. \tag{5.3}$$

Refinements and extensions for GSMs

In this paper, several novel features had been added so that GSMs are able to:

- ◇ Analyze correlated time-to-event data, in particular including proportional odds models with random effects, which was introduced in [45].
- ◇ use the functional ANOVA decomposition technique to estimate joint time-dependent and non-linear effects for age (4.2.2).
- ◇ Incorporate penalized marginal likelihood estimation methods, corresponding to penalized regression splines and clustered survival data.
- ◇ Incorporate multivariate regression splines that can be derived from radial basis functions (4.2.1).
- ◇ Perform continuous stratified analysis on *age* combined with age-varying effects for a specific treatment (4.2.2).

For clustered survival data, a full log-marginal likelihood function can be constructed as a sum of I cluster-level log-marginal likelihood with a sample size of n_i. This can be given as

$$\begin{aligned}\log L^M(\beta,\theta) &= \sum_{i=1}^{I} \log\{L_i^M(\beta,\theta)\} \\ &= \sum_{i=1}^{I} \log\left\{ \int_{-\infty}^{\infty} \left\{ \prod_{j=1}^{n_i} \{h(t_{oij}|\mathbf{z}_{ij},b_i)\}^{\delta_{ij}} S(t_{oij}|\mathbf{z}_{ij},b_i) \right\} p(b_i|\theta)\mathrm{d}\, b_i \right\}\end{aligned}$$

where $h(\cdot)$ and $S(\cdot)$ are derived from model 5.3, and $p(\cdot)$ was the density function of b with a mean of unity and the variance of θ.

Similarly, the corresponding penalized log-marginal likelihood function can be constructed as $\log L^M(\beta,\theta)$ minus a roughness penalty function for each smooth regression component 2.19. The remaining optimization problems for both parametric and penalized estimation procedures are the same as those for independent survival data.

As seen from the above marginal likelihood method, clustered survival data requires novel statistical procedures to: (1) construct a marginal (penalized) likelihood function integrating out the unknown random effect for each cluster; and (2) approximate the related integration for each cluster with different numerical techniques, including Gauss-Hermite quadrature or adaptive Gauss-Hermite methods.

Comparison of results in simulation studies

To compare the results from the proposed approaches and some predefined model components (with known model parameters), we assessed the performance of: (1) the numerical approximation methods (e.g. adaptive Gauss-Hermite method); (2) the (penalized) marginal likelihood estimation methods; and (3) the use of the AIC criterion for the choice of a specific link function. We were also interested in investigating the potential impact of model misspecifications on the estimates of baseline functions of time and the regression coefficient of interest, such as misspecified functional forms of covariates and the parametric distribution of random effects.

The simulation studies, undertaken in the investigational settings, demonstrated that:

- ◇ Both parametric and penalized procedures perform well for estimating β and the variance of random effect b when the fitted and the assumed models were the same. The empirical coverage probabilities of the regression coefficients and the variance were close to the nominal level of 95%.

- ◇ Estimates of regression coefficients were not sensitive to misspecification of parametric distributions of random effects, but the estimate of the transformed baseline functions of time) was affected. This may be due to the baseline hazard function absorbing the expectation of frailty $U_i = \exp(b_i)$ that may be not unity if $E(b_i) = 0$ (see the relationship of $E(U_i) = \exp(\theta/2)$ in Section 10.2.1 of [11]).

- ◇ For an extended proportional hazards or proportional odds model with one random effect, the proposed methods can provide similar results compared to other implementation, such as *survival:::coxph*, *coxme*, and *frailtypack* in R. Compared to these established methods, extended GSMs were able to include different links, and provide a flexible framework for modeling time-dependent and non-linear effects.

- ◇ The *frequency* of better models among all model fittings by the AIC [34] can be a useful measure to examine which model coincides with the data generating mechanism in simulation studies.

- ◇ The misspecification of the time-dependent effect of a binary variable leads to biased estimates of the corresponding effects, but the bias in other estimates (of regression coefficients and the variance) were not obvious in this settings. Further investigations were required to theoretically explain "where" the difference in $\beta_1(t)$-β_1 "goes" and what was affected by this misspecification. This could be related to the proposed dynamic frailty process as a random effect [128].

- ◇ However, the misspecification of the non-linear effects of a continuous variable will lead to biased estimates of corresponding variables; the estimates of the transformed baseline function and the variance of random effect b_i were also impacted. For example, theoretically the difference between the underlying functional form of age $f(age)$ and a misspecified functional form $g(age)$ will be decomposed into two parts: (1) the empirical average of these difference $\frac{1}{n}\sum_{i=1}^{n}\{f(age_i) - g(age_i)\}$ will be absorbed by the baseline function; and (2) the remaining part will be added to the random effect b_i, which did increase the magnitude of the underlying variance of b_i.

- ◇ The estimates of non-linear covariate effects from these proposed methods were similar to those from Cox regression with random effects.

Comparisons and illustrations through examples

Through two available examples (i.e. readmission to hospital for colorectal cancer patients [32] and diabetic retinopathy [129]), we aimed to compare our proposed features to well-established methods and demonstrated the novel features proposed in this paper. The following conclusions can be made:

- ◇ In the first example, the results from proposed approaches were similar to well-established methods for repeated events within the same patient, such as the implementation of *frailtypack*, *survival::coxph*, and *coxme* to fit a gamma frailty model with time-constant effects or a proportional hazards model with normally distributed random effects. Additionally, we fitted Andersen-Gill models using marginal approaches: a parametric GSM with the log-log link and the implementation of *survival::coxph*. Both of these provided similar estimates of the regression coefficients.
- ◇ In addition to time-dependent effects, continuous covariates (e.g. age) can be represented in various functional forms, such as the conversion to categorical variables or for constructing multivariate splines, and non-linear effects. All these functional forms can be investigated in a proposed GSM for clustered data.
- ◇ In the second example, we compared GSMs to other well-established conditional approaches, such as semi-parametric transformation models for both proportional hazards and proportional odds models with random effects. The estimates of regression coefficients of interest were also similar, within the time-constant effect setting.
- ◇ GSMs can perform a stratified analysis in a continuous manner. For example, the model proposed in Section 5.2 (Paper II) can be identified as a continuous stratified analysis by age.

5.3 Paper III

We extended GSMs for relative survival analysis, and the outcome of interest was time-to-death due to the disease of interest. Under the framework of GSMs, the corresponding relative (or net) survival function $S_E(t|\mathbf{z})$ can be modeled in matrix form as

$$g\begin{pmatrix} S_{E1}(t|\mathbf{z_1}) \\ S_{E2}(t|\mathbf{z_2}) \\ \vdots \\ S_{En}(t|\mathbf{z_n}) \end{pmatrix} = \mathbf{X}(t,\mathbf{z})\beta, \tag{5.4}$$

where g is a user-specified link-function.

For population-based cancer survival data, the full likelihood function can be constructed on both individual data and corresponding national life tables as

$$\log L(\beta) = \sum_{i=1}^{n} \{\delta_i \log\{h_{Ei}(u_i|\mathbf{z}_i;\beta) + h_{Pi}(u_i+a_i, u_i+y_i|\mathbf{x}_i)\} + \log\{S_{Ei}(u_i|\mathbf{z}_i;\beta)\}\} \tag{5.5}$$

where $\mathbf{x}_i$ denotes variables used to stratify mortality rates for those general population and were matched to all patients under study, and h_{Pi} indicate matched mortality rates

in the general population. S_{Ei} and h_{Ei} were the corresponding net survivals and excess hazards for the patient i, respectively, and can be calculated from model 5.4. Based on the formula 5.5, the corresponding penalized likelihood can be derived in the way as described in 2.19.

Refined model components and extended features for GSMs

In this paper, we had added the following features to GSMs:

- ◇ Connected smooth functions to regression splines (e.g. truncated power basis functions and radial basis functions).
- ◇ Incorporated knot-based and knot-free regression splines for GSMs.
- ◇ Included parametric (e.g. the Weibull proportional hazards model and log-logistic proportional odds model) and spline-based flexible parametric models for relative survival analysis.
- ◇ Introduced a penalized likelihood framework involving both maximum likelihood and maximum penalized likelihood estimation approaches; reinterpreted penalized likelihood estimation methods as a statistical procedure combining parameter estimation and model selection for a number of spline basis functions, only requiring the maximum degrees of freedom for each smooth regression component.

Comparison of results in simulation studies

The simulation studies, undertaken the investigational setting, demonstrated that:

- ◇ The coverage probabilities for the estimates of regression coefficients were close to the nominal level (95%); the coverage probabilities for the estimates of 10- and 15-year net survival using the proposed sandwich variance estimator also reached the nominal level.
- ◇ Model-based methods and the Pohar-Perme method can provide similar estimates along with the coverage probabilities, especially from the time since diagnosis to the third quartile of observed survival time. For longer-term follow-up (greater than the third quartile of observed survival time), the uncertainty of estimates from the Pohar-Perme method becomes larger than that of the model-based approach, but the coverage probabilities over that period from both approaches were close to the nominal level.
- ◇ Under the penalized likelihood framework, both the parametric and penalized estimation approaches can be performed for relative survival analysis. We observed the same results from both statistical procedures. However, multiple models need to be fitted within the parametric estimation approach, but only one fitted model within the penalized estimation approach.
- ◇ Both approaches can capture the underlying varying trends for continuous variables (e.g. time and age), and the underlying fixed-effect for the binary

variable. Additionally, the model-based approach can provide more information about some potential relationships between the prognostic factors of interest and cancer survival or excess hazards.

Comparisons and illustrations through examples

- ◇ The knot-free regression splines had good properties and were able to: (1) be an alternative tool to knot-based regression splines with a strategy using either equally-spaced or quantiles points for the locations of knots; and (2) select the optimal degrees of freedom of regression splines according to the AIC.
- ◇ Through comparing several knot-based regression splines to the knot-free regression spline, we found that the value of the AIC from the final chosen model was typically less than 3, under the time-constant effects setting.
- ◇ The Weibull parametric relative survival model provides a larger AIC value, however all estimates of covariates were similar or only slightly different to the results from the other three strategies.
- ◇ The penalized estimation approach was more easily applied in practice, as it does not require fitting multiple models with different numbers and locations of spline basis function (or spline knots).

5.4 Paper IV

Background and aims

Prostate cancer (PCa) is one of the leading causes of cancer death for men in many western countries. The PSA test for prostate cancer was commonly used as a simple and inexpensive way to find men with an increased risk of asymptomatic prostate cancer. Moreover, consecutive PSA test intervals may vary within and between individuals. Our objectives were: (1) to assess whether PSA testing had changed in recent years in Sweden; (2) to describe the probabilities of subsequent outcomes for men in the general population and those who have had a baseline PSA test; and (3) to describe the PSA retesting frequencies for men who have had a baseline PSA test.

Study population

The study population was defined as men living in Stockholm between January 1, 2003 and December 31, 2014 with no previous diagnosis of prostate cancer. Data were extracted from the Stockholm PSA and Prostate Biopsy Register. The PSA data included information on the date of PSA test and the total PSA level. which were linked to data from other health and population registers, including: (1) date and cause of death from the Cause of Death Register; (2) date of cancer diagnosis from the National Cancer Register; and (3) the population living in Stockholm in December of each year from the Population Register. The individual data set was integrated using an encrypted identifier, with data linkage performed by the National Board of Health and Welfare and Statistics

Sweden. These registers had been shown to have high quality [130]. The analysis of the linked data had been approved by the regional ethics committee in Stockholm.

Design and methods

The prevalence of PSA testing was modeled using a log-binomial regression for 10-year age groups. We assumed a piece-wise linear period effect with a knot at 2010. We did not model for the prevalence in 2014, as we expected that this would be affected by the STHLM3 diagnostic trial [131]. We tested for a change in slope at 2010 and reported the annual percentage change from 2010 to 2013.

The proportions of men in different testing and healthy states were estimated using Markov multi-state models [12, 132]. Furthermore, we used a GSM with a gamma frailty to estimate PSA retest rates and the cumulative proportion of men underwent a PSA retest.

Analysis and results

Over the study period, 1,253,309 men underwent PSA testing. There was evidence for a decline in PSA test rates during the period 2010 to 2013. The 10-year probability of having a PSA test for men aged 50-59, 60-69 and 70-79 years was 62.6%, 59.0% and 43.4%, respectively. Large proportions of men had a PSA test at the 10-year follow-up, which was associated with a markedly increased risk of being diagnosed with prostate cancer for those with an index PSA test value of 3 ng/mL and over. The 10-year risk of prostate cancer death was 0.1%, 0.3% and 1.2%, for men in the three age groups respectively, with a PSA less than 3 ng/mL. Using a shared frailty model, there was marked inter-individual variability in PSA retesting. The index PSA value was strongly associated with the PSA retesting interval and explained approximately 20% of the variability in retesting interval.

Table 5.1 provides the time-dependent cumulative risks (%) of being in subsequent outcomes following three initial states: (i) since study entry without an index PSA test; (ii) since study entry with an index PSA < 3 ng/mL; and (iii) since study entry with an index PSA of 3+ ng/mL, stratified by age groups, Stockholm males, 2003-2014.

Conclusions

There was evidence that PSA testing had decreased since 2010. However, large proportions of men have had a PSA test and there was a markedly increased risk of being diagnosed with prostate cancer. Our findings provide a detailed description of prostate cancer testing in one population, and provide useful evidence to clinicians when counselling men for PSA testing.

Table 5.1: Time-dependent cumulative risks (%) of being in subsequent outcomes following: (i) since study entry without an index PSA test; (ii) since study entry with an index PSA < 3 ng/mL; and (iii) since study entry with an index PSA of 3+ ng/mL, stratified by age groups, Stockholm males, 2003-2014

		Time-dependent cumulative risks of being subsequent outcomes (%)											
		Since study entry (without an index PSA)				Since index PSA < 3 ng/mL				Since index PSA ≥ 3ng/mL			
Age (years)	Follow-up time (years)	Having a PSA test	PCa diagnosis	PCa death	Other causes of death	Having a second PSA test	PCa diagnosis	PCa death	Other causes of death	Having a second PSA test	PCa diagnosis	PCa death	Other causes of death
50-59	1	14.6	0.2	0.0	0.4	17	0.2	0.0	1	45.7	18.4	0.2	1.2
	2	24.8	0.5	0.0	0.7	38.5	0.4	0.0	1.7	54.2	20.7	0.6	2.2
	5	46.0	1.7	0.1	2.2	68.8	1.4	0.0	4	58.9	26.1	1.3	4.4
	10	62.6	4.3	0.2	5.8	78.4	4.7	0.1	7.8	54.8	31.9	2.1	8.5
60-69	1	21.4	0.8	0.0	0.9	23.5	0.2	0.0	2.2	46.5	15.6	0.5	2.2
	2	32.5	1.8	0.1	1.9	46.7	0.4	0.0	3.8	55.9	18	1.1	4.1
	5	51.9	4.2	0.3	5.6	68.7	1.5	0.1	9.2	57.4	21.9	2.4	9.8
	10	59.0	7.6	0.9	14.2	68.6	3.8	0.3	18.3	50.0	23.9	3.8	19.1
70-79	1	24.8	1.0	0.2	2.5	27.9	0.2	0.1	6	47.2	8.2	1.0	6.0
	2	34.8	1.9	0.5	5.1	46.7	0.4	0.2	10.8	56.7	9.0	2.4	10.3
	5	49.5	3.8	1.4	15.1	56.6	0.8	0.6	24.7	52.5	9.8	5.1	24.1
	10	43.4	4.7	3.2	36.7	44.2	1.2	1.2	47.3	35	8.8	8.2	45.3

6 Discussion

6.1 Overall conclusion

The work presented in this thesis aimed to enrich the class of generalized survival models and corresponding estimation methods. We have provided a coherent framework to: (1) model independent and correlated time-to-event data; (2) incorporate both knot-based and knot-free regression splines, with different estimation strategies; (3) include corresponding parametric and penalized estimation procedures for parameter estimation, with model selection for the number of spline basis functions in either a non-continuous or continuous manner; (4) adopt full likelihood-based estimation methods to be able to estimate all model components, including regression coefficients and smooth regression components; (5) incorporate polynomials for representing functions of time as parametric models; and (6) introduce mature regression spline-based smooth techniques developed for generalized additive models into GSMs. The related implementation has been made available on the CRAN for R-users.

In conclusion, these proposed methods performed well in extensive simulation studies, with good point estimates and coverage probabilities. Through the analysis of real example data, similar results can also be observed between the proposed methods and some well-established approaches, under proportional hazards or proportional odds models settings. Moreover, novel features were also illustrated in both simulations and applications.

6.2 Strengths and limitations

In general, more than one statistical procedure can be used for a specific estimator in a proposed survival regression model. Given the same real data, it would: (1) be reasonable to obtain similar results (e.g. baseline hazard function, hazard and survival functions given $\mathbf{Z}$) from different statistical approaches, if the data sample is independent identically distributed and event times are not sparse; (2) be possible to fit a proposed model using different statistical procedures, e.g. either different ways to represent smooth regression components or distinct estimation methods; and (3) be desirable to have abilities to explore various investigations, e.g. distinct functional forms for effects of age.

By contrast, strengths of these extended GSMs and estimation methods include:

- ◇ The use of the finite difference method for the calculation of hazard functions is considerably less restricted in the choice of smooth functions (or the transformation) for time;
- ◇ Inclusion of the knot-free thin plate regression splines avoids selecting the locations of spline knots. Moreover, this provides a type of nested spline basis functions for survival models;
- ◇ Under the penalized likelihood framework, effective degrees of freedom (or dimension) can be optimally chosen for each smooth regression component

through the corresponding smoothing parameters. Users only need to set the maximum value of potential degrees of freedom for each smooth function.

Potential limitations includes:

- ◇ Use of only time-fixed baseline covariates;
- ◇ The process of the penalized estimation procedure could be time-consuming for the analysis of larger data sets or a proposed model with multiple smooth regression components. One alternative option is to set all smoothing parametric to be zeros and perform parametric estimation procedure for model fitting; another option is to apply efficient computational methods to larger data sets [133], which needs further investigation;
- ◇ In correlated survival data analysis, random effects and baseline covariates are assumed to be independent in the current extensions (Paper II), in which random effects (or frailty) either are treated as unmeasured heterogeneity or model within-cluster association. If the random effects represent the role of effects of unmeasured confounders, Sjölander introduced the idea of a between-within model for clustered data [134] into survival analysis [135], which could allow the dependence between frailty and covariates especially for twin research [136, 137].

6.3 Specific issue 1: About the constraint: $h_i(t|\mathrm{z_i})$>0

This section aims to examine the condition of $h_i(t|\mathrm{z_i})$>0 that we considered in the process of the maximum likelihood and maximum penalized likelihood estimation.

For example, for right-censored survival data, suppose all observations in a collection of $\{(u_i,\delta_i,\mathbf{z}_i), i = 1,2,\ldots,n.\}$, where $u_i = min\{t_i,c_i\}$, $\delta_i = I(t_i \leq c_i)$ and the vector of predefined baseline variables $\mathbf{z}_i$, and n is the total number of patients. The classical log-likelihood function is in the form of

$$\log L(\beta) = \sum_{i=1}^{n} \{\delta_i \log h(u_i|\mathbf{z}_i;\beta) - H(u_i|\mathbf{z}_i;\beta)\} \tag{6.1}$$

where $h(u_i|\mathbf{z}_i;\beta)$ and $H(u_i|\mathbf{z}_i;\beta)$ are the hazard and cumulative hazard functions for an individual i. The maximum likelihood estimators satisfy,

$$\widehat{\beta} = \arg\min_{\beta}\{-\log L(\beta)\} \qquad \text{subject to : } h_i(u_i|\mathbf{z}_i;\beta) > 0. \tag{6.2}$$

6.3.1 Solution 1: penalty method

One possible way to solve this constrained optimization problem is to apply the penalty method [138] to get proper estimators. That is, a quadratic penalty term

$$P_n(\beta) = \frac{\kappa}{2} \sum_{i=1}^{n} \left\{h^2(u_i|\mathbf{z}_i;\beta) I(h(u_i|\mathbf{z}_i;\beta) < 0)\right\}, \tag{6.3}$$

is subtracted from the above likelihood function, where κ is the penalty coefficient and $I(\cdot)$ is an indicator function. The un-constrained optimization problems with the augmented objective function as $f(\beta|\kappa) = -\log L(\beta) + P_n(\beta)$. The corresponding maximum likelihood estimators $\widehat{\beta}$ satisfy that

$$\widehat{\beta} = \arg\min_{\beta}\{f(\beta|\kappa)\}. \tag{6.4}$$

In the j^{th} iteration of the optimization process for $\widehat{\beta}_{(j)}$, there are two situations: (1) in which $h_i(u_i|\mathbf{z}_i;\widehat{\beta}_{(j)})>0$ for each subject i holds with κ=1, then $P_n(\beta)$=0 and an appropriate optimization method can be applied to get the MLE of β; and (2) in which some $h_i(u_i|\mathbf{z}_i;\widehat{\beta}_{(j)})<0$ with κ=1. We mainly illustrate the second situation with two types of optimization algorithms.

6.3.1.1 Optimization methods only using the objective function

There is only the objective function $f(\beta|\kappa)$ needed in the process of parametric estimation with the Nelder-Mead optimization algorithm. In the iteration j, there are $h_i(u_i|\mathbf{z}_i;\widehat{\beta}_{(j)})<0$ with κ=1. An iteration process of κ for the proper estimator of $\widehat{\beta}_{(j)}$ is to be performed in the estimation procedure. The $h_i(u_i|\mathbf{z}_i;\widehat{\beta}_{(j)})$ will be replaced by a small value (for example, ε=1.0e-16) in the log-likelihood function 6.1 and is kept in the formulation 6.3. With the initial value of κ being 1, the value of κ is doubled in next iteration for κ until $h_i(u_i|\mathbf{z}_i;\widehat{\beta}_{(j)})>0$ for each patient. Continue to the iteration j+1 for the proper estimator of $\widehat{\beta}_{(j+1)}$ until a convergence criteria is met for $\widehat{\beta}_J$, with the total J iterations.

6.3.1.2 Optimization with the Newton-Raphson algorithm

In addition to the objective function $f(\beta|\kappa)$, the gradient function of the objective function is usually required to obtain the MLE of β in the form of:

$$\begin{aligned}\frac{d\,f(\beta|\kappa)}{d\,\beta} = &\sum_{i=1}^{n}\left\{-\delta_i\frac{d\log h(u_i|\mathbf{z}_i;\beta)}{d\,\beta}I(h(u_i|\mathbf{z}_i;\beta)\geq\varepsilon)+\frac{d\,H(u_i|\mathbf{z}_i;\beta)}{d\,\beta}\right\}\\ &+\kappa\sum_{i=1}^{n}\left\{h(u_i|\mathbf{z}_i;\beta)\frac{d\,h(u_i|\mathbf{z}_i;\beta)}{d\,\beta}I(h(u_i|\mathbf{z}_i;\beta)<0)\right\}.\end{aligned}$$

6.3.2 Solution 2: direct use of monotonic regression splines

An alternative solution to this quadratic penalty method is to directly use monotonic regression splines for constrained smooth functions, such as monotone I-splines[139] applied in the R package *frailtypack* or using B-splines with constrained coefficients[140].

In general, hazard functions can be derived from a GSM. For instance, we use the

formula 4.5

$$\begin{pmatrix} h_1(t|\mathbf{z_1}) \\ h_2(t|\mathbf{z_2}) \\ \vdots \\ h_n(t|\mathbf{z_n}) \end{pmatrix} = -\frac{G'(\eta(t,\mathbf{z};\beta))}{G(\eta(t,\mathbf{z};\beta))}\frac{d\,\eta(t,\mathbf{z};\beta)}{dt} = -\frac{G'(\eta(t,\mathbf{z};\beta)}{G(\eta(t,\mathbf{z};\beta)}\mathbf{X}_D(t,\mathbf{z})\beta$$

with

$$-\frac{G'(\eta(t,\mathbf{z};\beta)}{G(\eta(t,\mathbf{z};\beta)} > 0$$

where the common inverse link functions are $G(\cdot)$=exp$-$exp$(\cdot)$ and $G(\cdot)$=$1/(1+\exp(\cdot))$.

To ensure $h_i(t|\mathrm{z_i})$>0 for each subject, we only need $\mathbf{X}_D(t,\mathbf{z})\beta$>0, which indeed coincides with two requirements: (1) If the definition of probability 4.3 holds, it requires $s(T)$ or $s(\log(T))$ to be an increasing transformation; (2) As described in 4.18, the time-dependent cumulative hazards ratio and odds ratio can be a type of cumulative effects, with increasing (or non-decreasing) first derivatives with respect to t or $\log(t)$.

6.4 Specific issue 2: Statistical inference for GSMs

6.4.1 Standard error and hypothesis testing

The M-estimator from both maximum likelihood and penalized likelihood estimation methods can be asymptotically normally distributed,[22, 141], if the condition of the proposed model is correct. In this setting, it may be reasonable to use $\mathcal{H}_l$ and $\mathcal{H}_{pl}$ evaluated at $\hat{\beta}$ to derive standard errors (SEs). Based on these SEs, hypothesis testing on regression coefficients or smooth regression components can be performed, which can be similar to related tests within generalized additive models [2].

In practice, we do not know whether or not the proposed model is correct. Generally, in the situation in which the proposed model is misspecified, the commonly used SEs approximated from the empirical Hessian matrix are invalid. Although one can apply the sandwich algorithm for SEs, the corresponding estimates may be biased [142]. In general, the bootstrap method based on re-sampling techniques can be applied to construct confidence intervals for estimated regression coefficients and smooth regression components, however, the proposed model remains unchanged. How to specify functional forms in multivariable analysis (for observational studies) is being investigated in an international collaborative study [33]. In this context, given baseline covariates $\mathbf{z}$, spline-based survival regression models could be a useful tool to approximate the underlying model components, with data-driven functional forms for baseline functions and covariate effects.

6.4.2 Model checking techniques

Based on a fitted survival model, a graphical method can be applied to check goodness-of-fit (e.g. Cox-Snell residual plots in Paper I), which is also related to the martingale

residual [105] based on estimated $H(t|\mathbf{z})$ and the final status δ for each subject. In general, the martingale residual can be expressed in a simple form for each subject [143]

$$M_i = \delta_i - \widehat{H}(t_{oi}|\mathbf{z_i}).$$

Cortese and Scheike proposed an adjusted martingale residual to check goodness-of-fit for relative survival models [144]. Collett proposed to explore effects forms for continuous covariates by martingale residual plots [11]. Especially for parametric regression with censored survival data, Lin and Spiekerman [145] proposed several model checking techniques, which could be further investigated for GSMs.

6.5 Extension to biomedical research

If the phenotype of interest is a time-to-event outcome, it would be interesting to extend the current statistical procedures to incorporate genomic data (or integrated omics data).

For instance, it is well known that when the outcome of interest is a binary variable over a fixed time period, classic logistic regression is commonly applied for risk prediction or classification. However, in practice, the sub-population might be a dynamic group allowing subjects to be included or excluded for various reasons. The study period could also be dynamic. This calls for, or can be extended to, the dynamic parametric log-logistic proportional odds model and a GSM with the $-$logit link, both of which are exact "dynamic" logistic regression models for censored survival data. Within survival analysis, these models can be an alternative tool for prediction and classification of patients, and also with similar interpretation to the "static" logistic regression model over a fixed time period.

The proposed model with both clinical and genomic data can be given in the form of

$$\log\left\{\frac{1-S(t|\mathbf{z}_c,\mathbf{z}_g)}{S(t|\mathbf{z}_c,\mathbf{z}_g)}\right\} = s(\log(t)) + \beta_c^T\mathbf{z}_c + \beta_g^T\mathbf{z}_g, \tag{6.5}$$

where $\mathbf{z}_c$ and $\mathbf{z}_g$ are the vectors of clinical and genomic variables, respectively, and $s(\log(t))$ is a parametric form of $\log(t)$ or flexible regression spline representation for estimating the transformed baseline survival function in the form:

$$s(\log(t)) = \log\left\{\frac{1-S_0(t)}{S_0(t)}\right\}, \tag{6.6}$$

where $S_0(t)$ is the survival function for the reference subgroup with categorical variables $\mathbf{z}_c$ and $\mathbf{z}_g$ be zero, with empirical average values for continuous variables $\mathbf{z}_c$ and $\mathbf{z}_g$.

Note that, based on model 6.5, one potential challenge to predict survival probability is to deal with the problem of *small n large p*, which indicates that there are p-dimensional covariates that are much larger than the number of subjects n in the study population. In statistics, LASSO (least absolute shrinkage and selection operator) [146] and Elastic net [147] are common regression analysis methods to handle this issue.

More generally, joint models of survival outcomes and longitudinal data (e.g. clinical biomarkers) have also been proposed in medical investigations, which simultaneously handle the problem of variable selection [148]. In this thesis, after adjustment for

measured covariates, censoring times are assumed to be independent of event times for each individual. However, in biomedical studies, dependent censoring might arise in the competing risks setting [149]. All these topics are of interest for future research.

7 Acknowledgements

I would like to express my heartfelt gratitude to all the people who provided support during my PhD study. In particular I would like to acknowledge:

My main supervisor, associated professor Mark Clements. I really appreciated those regular meetings, for sharing your knowledge and practical experiences on testing new ideas with R quickly. Thank you for all your guidance and support. Thank you for discussing about mathematical and statistical issues related to model building, which made me reflect until completely absorbed those knowledge and skills that are invaluable. Thank you for encouraging me to document with LaTex and become familiar using git for programming version control since the first day we met. I am amazed that you are so skilled at programming in C++, R, SAS, JavaScript, ... It is a pity that I did not have enough time to learn more programming with you. Under your supervision, I have felt free to do things in my own way with your guidance. It's an honor for me to be the first PhD student having you as the primary supervisor.

My co-supervisor and co-author, Professor Yudi Pawitan. I am very grateful for your input and cheering me up during the most difficult period while Paper I had not been accepted within the first 3 years of my PhD. Thank you for all the insightful scientific discussions, for the tips on writing manuscript and revision. I will never forget those weekly meetings, especially during preparation of Paper II on random effect models, and your patience in correcting my English pronunciation. My appreciation also extends to your perceptiveness by directly pointing out my issue of the mathematical way rather than the statistical way of thinking while I was studying the book *Statistical Modelling and Inference Using Likelihood*. I hope to learn more from you in my future research.

My co-supervisors and co-authors, Arvid Sjölander and Fredrik Wiklund. I appreciate you nice suggestions during my half time review, the discussions concerning my PhD projects and for always being positive. I regret that I have not conducted two planned studies with you on causal inference for dynamic treatment regimens and risk prediction with genetic data, respectively. I hope to have the opportunity to collaborate on related projects with both of you in the future.

My mentor Marie Reilly, for your warm-hearted and selfless support, for those nice chats during lunch or dinner, in your office or in the corridor.

All my co-authors on application papers, Henrik Grönberg, Martin Eklund, Jan Adolfsson, Tobias Nordström, Markus Aly, Thorgerdur Palsdottir, and Andreas Karlsson for your constructive suggestions. A special thanks to Henrik Grönberg for organizing the writing retreat when I drafted the preliminary version of Paper IV. Thanks to Andreas Karlsson for helping on the Ubuntu operating system, you are so smart with computer hardware and software. Thanks to all the others as well in this excellent group.

Paul Lambert, Michael Crowther and Paul Dickman for your input into our Papers II and III, respectively. Thanks to Anna Johansson for organizing several Survival Meetings and Therese Andersson, Caroline Weibull, Elisabeth Dahlqwist, Peter Ström for scientific

exchanges on a specific issue or sharing related knowledge on different topics. A special thanks to Paul Lambert for suggestions on my thesis.

The Biostat group, Juni Palmgren, Keith Humphreys, Sven Sandin, Rino Bellocco, Cecilia Lundholm, Robert Karlsson, Annika Tillander, Sandra Eloranta, Henrik Olsson, Henric Winell, Vu Nghia, Li Yin, Daniela Mariosa, Alessandra Grotta, and Gabriel Isheden, for all seminars over this period. Special thanks to Alexander Ploner for being committee member in my half time seminar and for your very helpful suggestions on my thesis.

All PhD-student colleagues for those interesting seminars. Special thanks to Bénédicte Delcoigne, Johan Zetterqvist, and Hannah Bower for collaborative learning of the book *Statistical Modelling and Inference Using Likelihood* by Yudi Pawitan. I also would like to thank Hannah Bower, Elisabeth Dahlqwist, Peter Ström for being member of the mock examination committee at my pre-dissertation seminar.

During this period of my PhD study, it is my pleasure to meet Odd O. Aalen and Sir David Cox. I took the *survival and event history analysis* course given by Odd O. Aalen, Ørnulf Borgan and Håkon K. Gjessing in Italy, 2014. In 2016, I presented our research on relative survival analysis in the *Population-Based Time-To-Event Analyses International Conference* (London) chaired by Sir David Cox. Thank you for sharing your knowledge and experience with us.

I would like to express my thankfulness to the wonderful working environment at MEB. All the research presented in this thesis was carried out here between 2013 and 2017. I would like to thank Camilla Ahlqvist and Gunilla Nilsson Roos for your support in my courses, my half time review and dissertation preparation. I also thank Marie Jansson for helping out on travels to abroad for attending course, seminar and conferences. Thanks to Gunilla Sonnebring for helping on dissertation preparation. Thanks to the MEB-IT support group for fixing my desk computer and software installation. Thanks to Åsa Agréus, Sofia Anderberg, and Erika Nordenhagen for helping on contract preparation and organizing my office, respectively. Thanks to my current and former roommates Linda Abrahamsson, Wenjiang Deng, Jinseub Hwang for chatting, sharing cookies and discussing on statistical issues.

Thanks to past and present Chinese colleagues at MEB: Chen Suo, Tong Gong, Qi Chen, Jiaqi Huang, Donghao Lu, Zhiwei Liu, Zheng Chang, Fei Yang, He Gao, Ruoqing Chen, Ci Song, Huan Song, Jie Song, Mei Wang, Xingdong Chen, Haiyun Wang, Xia Shen, Xu Chen, Tingting Huang, Bojing Liu, Zheng Ning, Qing Shen, Jiayao Lei, Yi Lu, Jiangrong Wang, Yunzhang Wang, Hong Xu, Wei He, Haomin Yang, Shuyang Yao, Jingru Yu, Yiqiang Zhan, Jianwei Zhu, and Shihua Sun for the friendship, a lot of fun during lunch, those gatherings for BBQ and Chinese Spring festival celebration.

I would like to extend my thanks to the Karolinska Institutet Board of Doctoral Education, Swedish Cancer Society and Swedish Research Council for supporting my doctoral education.

I would like to thank all contributors who had made the Latex template (`https://github.com/samuel-bohman/su-latex-phd-thesis-template`) available. I

adjusted it following KI PhD thesis requirement to produce this thesis frame.

最后，特别地感谢我最敬爱的杜副院长，谢谢您多年的谆谆教导和言传身教！最想感谢的是我的家人，Mei，在我读博期间，你贴心地照顾家庭，谢谢你这些年的陪伴，鼓励和支持！Leo，感谢你带给爸妈的诸多快乐和成长。感恩！

A Illustrative examples in R

A.1 R code for parametric models

```
## Load R Packages if you don't have them
library("rstpm2")
library("flexsurv")
library("SurvRegCensCov")
library("survival")

## Fit the Weibull AFT model
reg.weibull <- survreg(Surv(recyrs, censrec) ~ group, data = bc,
                       dist = "weibull")

## Convert to parametrs in the Weibull PH model
ConvertWeibull(reg.weibull, conf.level = 0.95)$vars

##               Estimate          SE
## lambda      0.03472474 0.005959304
## gamma       1.37965178 0.066787587
## groupMedium 0.84653938 0.171278007
## groupPoor   1.67243282 0.164243936

## Fit the Weibull proportional hazard model (parametric model)
summary(fit_ph <- stpm2(Surv(recyrs, censrec) ~ group, data = bc,
                        smooth.formula = ~log(recyrs),
                        link.type = "PH"))@coef

##               Estimate Std. Error    z value        Pr(z)
## (Intercept) -3.3603117 0.17161617 -19.580391 2.272690e-85
## groupMedium  0.8465194 0.17128019   4.942308 7.720332e-07
## groupPoor    1.6725060 0.16424387  10.183065 2.360090e-24
## log(recyrs)  1.3796448 0.06678709  20.657359 8.382576e-95

## Intercept = log(lambda)
## log(recyrs) = gamma


## Fit the log-logistic AFT model
reg.loglogistic <- survreg(Surv(recyrs, censrec) ~ group, data = bc,
                           dist = "loglogistic")

## Convert to parametrs in the log-logistic PO model
ConvertWeibull(reg.loglogistic, conf.level = 0.95)$vars

##               Estimate          SE
## lambda      0.02352749 0.004729315
## gamma       1.75456634 0.084878201
```

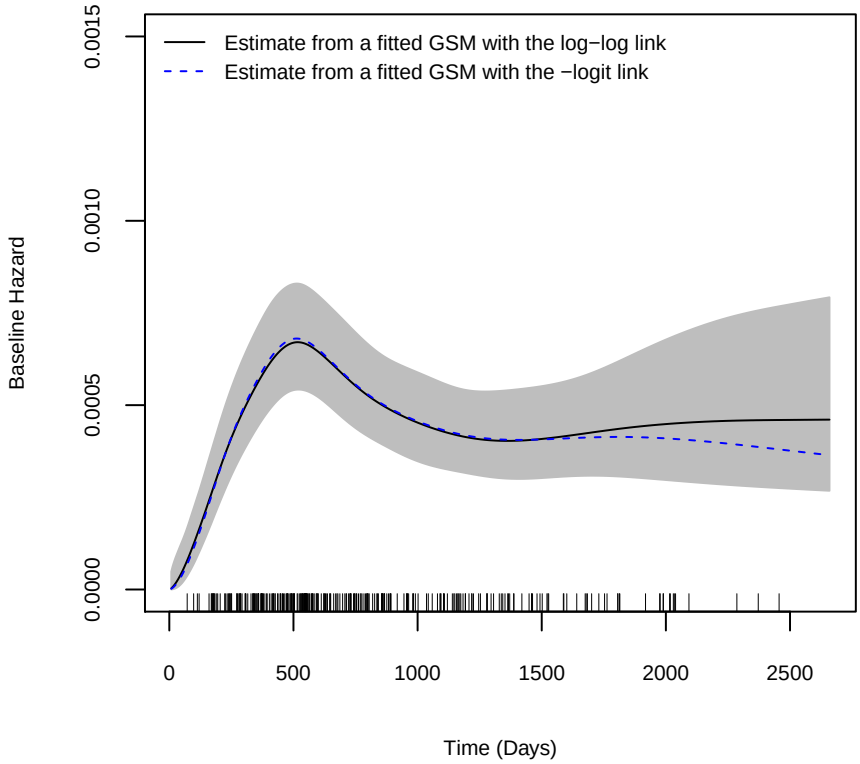
```
## groupMedium 1.09303127 0.208815258
## groupPoor   2.26035496 0.211622325

## Fit the log-logistic proportional odds model (parametric model)
summary(fit_po <- stpm2(Surv(recyrs, censrec) ~ group, data = bc,
                        smooth.formula = ~log(recyrs),
                        link.type = "PO"))@coef

##              Estimate Std. Error    z value        Pr(z)
## (Intercept) -3.749586  0.2010123 -18.653513 1.182470e-77
## groupMedium  1.093031  0.2088153   5.234439 1.654865e-07
## groupPoor    2.260355  0.2116223  10.681078 1.248138e-26
## log(recyrs)  1.754566  0.0848782  20.671578 6.244161e-95

## Intercept = log(lambda)
## log(recyrs) = gamma
```

A.2 R code for flexible parametric models

```
## use penalized estimation procedure to fit a spline-based GSM
## a GSM with the log-log link
pfit_ph = pstpm2(Surv(rectime, censrec==1) ~ 1, data=brcancer,
                 smooth.formula = ~s(log(rectime), k=10) +
                                   s(log(rectime), k=10, by=hormon),
                 link.type = "PH")

## the optimal smooth parameter
pfit_ph@sp

## [1]    1.152982 6576.648300

## the effective degees of freedom
pfit_ph@edf_var

##        s(log(rectime)) s(log(rectime)):hormon
##               4.132525               2.000525

## a GSM with the -logit link
pfit_po = pstpm2(Surv(rectime, censrec==1) ~ 1, data=brcancer,
                 smooth.formula = ~ s(log(rectime), k=10) +
                                    s(log(rectime), k=10, by=hormon),
                 link.type = "PO")

## the optimal smooth parameter
pfit_po@sp

## [1]   0.9062523 203.8098599
```

```
## the effective degees of freedom
pfit_po@edf_var

##        s(log(rectime)) s(log(rectime)):hormon
##               4.057106               2.012676

## Estimation of baseline hazard functions from two fitted GSMs
new.tim = seq(min(brcancer$rectime), max(brcancer$rectime),length=500)
par(cex.axis=0.7, cex.lab=0.7)
plot(pfit_ph, newdata=data.frame(rectime=new.tim, hormon=0),
     ylim=c(0,0.0015), cex.main=0.75,
     type="haz", xlab="Time (Days)", ylab="Baseline Hazard",
     main="Appendix A.2: Estimated baseline hazards from GSMs with different links")

## prediction from the fitted proportional odds model
haz0_po = as.vector(predict(pfit_po, newdata =
                      data.frame(rectime=new.tim, hormon=0), type="haz"))
lines(new.tim, haz0_po, lty=2,col="blue")
legend("topleft",
       legend=c("Estimate from a fitted GSM with the log-log link",
                "Estimate from a fitted GSM with the -logit link"),
       col=c("black","blue"),
       lty=c(1,2), cex=.70, y.intersp=1.1, lwd=c(1,1),
       bty="n")
```

Appendix A.2: Estimated baseline hazards from GSMs with different links

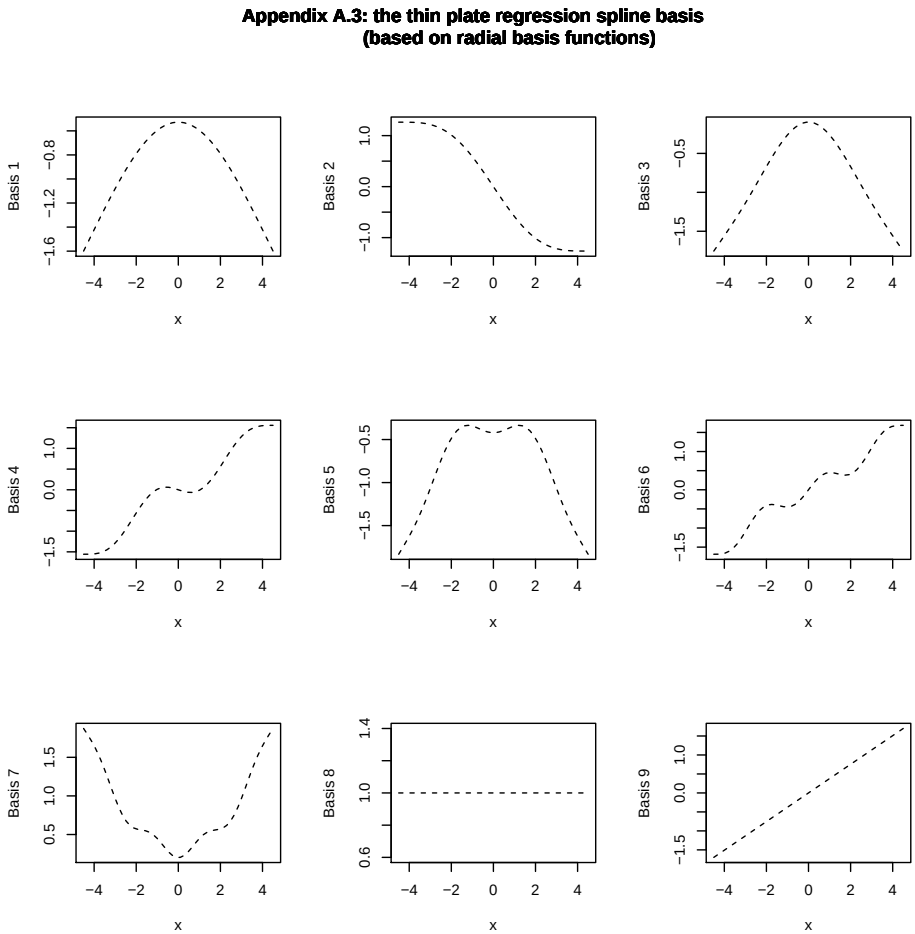


A.3 Univariate thin plate regression spline basis in R

```
## Following the procedures described in Section 4.2.1
m = 3 ## for cubic splines
set.seed(2017)
n_sample = 50
x_sam = seq(-6,3,length = n_sample)
x = sort(x_sam)
x = x - mean(x)
nn=length(x)
vec = vector()
for(i in 1:nn){
  vec = rbind(vec,abs(x[i]-x)^m)
}
E=as.matrix(vec,n_sample, n_sample)/12
```

```
## eigndecomposition
res = eigen(E, symmetric=TRUE)
D_old = diag(res$values)
U_old = res$vectors
# EE = U_old%*%D_old%*%t(U_old) ## eigen decompsition

## change the order to be |D_ii| > |D_i+1,i+1|
abs_value = abs(res$values)
ord = order(abs_value, decreasing = TRUE)
D = D_old[ord,ord]
U = U_old[,ord]
# EEE = U%*%D%*%t(U) ## eigen decompsition

## add bountary constraint TU_k\delta_k=0
k = 9
TT = matrix(rbind(rep(1,nn), x), 2, nn)
C12 = TT%*%U[,1:k]

## Get a basis for null space of the constraint
qrC = qr(t(C12))
Q = qr.Q(qrC,complete=TRUE)

## absorb constraint into basis
## E_k=U_k D_k U_k^T
## take the last k-2 columns
UDQ = (U[,1:k])%*%(D[1:k,1:k])%*%(Q[,(nrow(C12)+1):ncol(C12)])
UQ = (U[,1:k])%*%(Q[,(nrow(C12)+1):ncol(C12)])

## penalty matrix S=Q'DQ
S = t(Q[,(nrow(C12)+1):ncol(C12)])%*%(D[1:k,1:k])%*%
     (Q[,(nrow(C12)+1):ncol(C12)])

res_S = eigen(S, symmetric=TRUE)

## final basis without the sum-to-zero constraint
basis0 = UDQ
bas_full = as.matrix(cbind(UDQ, rep(1,nn), x))

## improve the numerical stability of the algorithm, Wood suggests to
## impose the mean square size of each column of bas_full to be 1
## applied in the mgcv package in R
W_vec = sapply(1:ncol(bas_full), function(i)
                sqrt(nrow(bas_full)/sum(bas_full[,i]^2)))
W = diag(W_vec)
bas_W = bas_full%*%W

## An example
par(mfrow=c(k/3,3),oma=c(0,0,2,0))
for(j in 1:k){
```

```
  plot(x, bas_W[,j], type="n", xlab="x", ylab=paste0("Basis ",j))
  lines(x, bas_W[,j], lty=2)
  title("Appendix A.3: the thin plate regression spline basis
            (based on radial basis functions)", outer=TRUE)
}
```



References

[1] P. Royston and W. Sauerbrei, "A new measure of prognostic separation in survival data," *Statistics in Medicine*, vol. 23, no. 5, pp. 723–748, 2004. vii, 1, 10, 19

[2] S. Wood, *Generalized Additive Models: An introduction with R*. CRC Press, 2006. vii, 1, 7, 8, 9, 11, 24, 25, 33, 46

[3] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958. 1

[4] O. Aalen, "Nonparametric inference for a family of counting processes," *The Annals of Statistics*, pp. 701–726, 1978.

[5] W. Nelson, "Hazard plotting for incomplete failure data(multiply censored data plotting on various type hazard papers for engineering information on time to failure distribution)," *Journal of Quality Technology*, vol. 1, pp. 27–52, 1969.

[6] M. Perme, J. Stare, and J. Estève, "On estimation in relative survival," *Biometrics*, vol. 68, no. 1, pp. 113–120, 2012. 1, 4, 5

[7] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society*, vol. 34, pp. 187–220, 1972. 1, 4, 9, 11, 12

[8] J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, 2011. 4, 11

[9] D. R. Cox and D. Oakes, *Analysis of Survival Data*, vol. 21. CRC Press, 1984.

[10] J. F. Lawless, "Regression methods for Poisson process data," *Journal of the American Statistical Association*, vol. 82, no. 399, pp. 808–815, 1987.

[11] D. Collett, *Modelling Survival Data in Medical Research*. CRC press, 2015. 4, 5, 6, 19, 37, 47

[12] P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding, *Statistical Models Based on Counting Processes*. Springer Science & Business Media, 2012. 41

[13] M. Bottai and J. Zhang, "Laplace regression with censored data," *Biometrical Journal*, vol. 52, no. 4, pp. 487–503, 2010. 1

[14] M. G. Hudgens, G. A. Satten, and I. M. Longini, "Nonparametric maximum likelihood estimation for competing risks survival data subject to interval censoring and truncation," *Biometrics*, vol. 57, no. 1, pp. 74–80, 2001. 1

[15] J. Sun, *The Statistical Analysis of Interval-censored Failure Time Data*. Springer Science & Business Media, 2007. 3

[16] P. Royston and M. K. Parmar, "Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects," *Statistics in Medicine*, vol. 21, pp. 2175–2197, 2002. 8, 9, 10

[17] V. Rondeau, D. Commenges, and P. Joly, "Maximum penalized likelihood estimation in a gamma-frailty model," *Lifetime Data Analysis*, vol. 9, no. 2, pp. 139–153, 2003. 1

[18] N. Reid, "A conversation with Sir David Cox," *Statistical Science*, pp. 439–455, 1994. 1

[19] O. Intrator and C. Kooperberg, "Trees and splines in survival analysis," *Statistical Methods in Medical Research*, vol. 4, no. 3, pp. 237–261, 1995. 1

[20] S. Durrleman and R. Simon, "Flexible regression models with cubic splines," *Statistics in Medicine*, vol. 8, no. 5, pp. 551–561, 1989. 8, 22, 24

[21] L. A. Sleeper and D. P. Harrington, "Regression splines in the Cox model with application to covariate effects in liver disease," *Journal of the American Statistical Association*, vol. 85, no. 412, pp. 941–949, 1990. 9

[22] R. J. Gray, "Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis," *Journal of the American Statistical Association*, vol. 87, pp. 942–951, 1992. 9, 11, 12, 46

[23] R. J. Gray, "Spline-based tests in survival analysis," *Biometrics*, vol. 50, no. 3, pp. 640–652, 1994.

[24] M. LeBlanc and J. Crowley, "Adaptive regression splines in the Cox model," *Biometrics*, vol. 55, no. 1, pp. 204–213, 1999.

[25] P. Bolard, C. Quantin, M. Abrahamowicz, J. Esteve, R. Giorgi, H. Chadha-Boreham, C. Binquet, and J. Faivre, "Assessing time-by-covariate interactions in relative survival models using restrictive cubic spline functions," *Journal of Cancer Epidemiology and Prevention*, vol. 7, no. 3, pp. 113–122, 2002.

[26] T. Cai and R. A. Betensky, "Hazard regression for interval-censored data with penalized spline," *Biometrics*, vol. 59, pp. 570–579, 2003.

[27] R. Giorgi, M. Abrahamowicz, C. Quantin, P. Bolard, J. Esteve, J. Gouvernet, and J. Faivre, "A relative survival regression model using B-spline functions to model non-proportional hazards," *Statistics in Medicine*, vol. 22, no. 17, pp. 2767–2784, 2003.

[28] G. Kauermann, "Penalized spline smoothing in multivariable survival models with varying coefficients," *Computational Statistics and Data Analysis*, vol. 49, no. 1, pp. 169–186, 2005.

[29] M. Costa and J. E. H. Shaw, "Parametrization and penalties in spline models with an application to survival analysis," *Computational Statistics and Data Analysis*, vol. 53, pp. 657–670, 2009.

[30] P. Royston and P. C. Lambert, *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox model*. Stata, 2011. 7, 8, 10, 24, 28

[31] P. Du, Y. Jiang, and Y. Wang, "Smoothing spline ANOVA frailty model for recurrent event data," *Biometrics*, vol. 67, no. 4, pp. 1330–1339, 2011. 9

[32] V. Rondeau, Y. Mazroui, and J. R. Gonzalez, "frailtypack: an R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation," *Journal of Statistical Software*, vol. 47, no. 1, pp. 1–28, 2012. 1, 11, 27, 37

[33] W. Sauerbrei, M. Abrahamowicz, D. Altman, S. Cessie, and J. Carpenter, "Strengthening analytical thinking for observational studies: the STRATOS initiative," *Statistics in Medicine*, vol. 33, no. 30, pp. 5413–5432, 2014. 1, 13, 14, 21, 46

[34] M. Abrahamowicz and T. MacKenzie, "Joint estimation of time-dependent and non-linear effects of continuous covariates on survival," *Statistics in Medicine*, vol. 26, no. 2, pp. 392–408, 2007. 1, 9, 14, 37

[35] L. Remontet, N. Bossard, A. Belot, and J. Esteve, "An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies," *Statistics in Medicine*, vol. 26, no. 10, pp. 2214–2228, 2007. 1, 9, 14

[36] D. Ruppert, M. Wand, and R. Carroll, *Semiparametric Regression*. Cambridge University Press, 2003. 1, 7, 8, 24

[37] S. N. Wood, "Thin plate regression splines," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 65, pp. 95–114, 2003. 1, 7, 9, 24, 25

[38] J. P. Klein and M. L. Moeschberger, *Survival Analysis: Techniques for censored and truncated data*. Springer Science & Business Media, 2005. 3

[39] G. J. Van den Berg and B. Drepper, "Inference for shared-frailty survival models with left-truncated data," *Econometric Reviews*, vol. 35, no. 6, pp. 1075–1098, 2016. 3

[40] T. M. Therneau and P. M. Grambsch, *Modeling Survival Data: Extending the Cox Model*. Springer, 2000. 4, 11, 12, 27

[41] L. Duchateau and P. Janssen, *The frailty model*. Springer Science & Business Media, 2007.

[42] A. Wienke, *Frailty Models in Survival Analysis*. CRC Press, 2010. 4

[43] S. Bennett, "Analysis of survival data by the proportional odds model," *Statistics in Medicine*, vol. 2, no. 2, pp. 273–277, 1983. 4, 6

[44] F. Vaida and R. Xu, "Proportional hazards model with random effects," *Statistics in Medicine*, vol. 19, no. 24, pp. 3309–3324, 2000. 4

[45] D. Zeng, D. Lin, and G. Yin, "Maximum likelihood estimation for the proportional odds model with random effects," *Journal of the American Statistical Association*, vol. 100, no. 470, pp. 470–483, 2005. 4, 10, 36

[46] "Measures of Cancer Survival." https://surveillance.cancer.gov/survival/measures.html. Accessed: 2017-10-12. 4

[47] H. Verheul, E. Dekker, A. Dunning, A. Moulijn, and P. Bossuyt, "Background mortality in clinical survival studies," *The Lancet*, vol. 341, no. 8849, pp. 872–875, 1993. 4

[48] "NCI Dictionary of Cancer Terms." https://www.cancer.gov/publications/dictionaries/cancer-terms?cdrid=655245. Accessed: 2017-10-12. 5

[49] P. C. Lambert, P. W. Dickman, and M. J. Rutherford, "Comparison of different approaches to estimating age standardized net survival," *BMC Medical Research Methodology*, vol. 15, no. 1, p. 64, 2015. 5

[50] P. Andersen and M. Vaeth, "Simple parametric and nonparametric models for excess and relative mortality," *Biometrics*, vol. 45, no. 2, pp. 523–535, 1989. 5

[51] J. Esteve, E. Benhamou, M. Croasdale, and L. Raymond, "Relative survival and the estimation of net survival: elements for further discussion," *Statistics in Medicine*, vol. 9, no. 5, pp. 529–538, 1990. 5

[52] W. A. Ghali, H. Quan, R. Brant, G. van Melle, C. M. Norris, P. D. Faris, P. D. Galbraith, M. L. Knudtson, A. A. P. P. for Outcome Assessment in Coronary Heart Disease) Investigators, *et al.*, "Comparison of 2 methods for calculating adjusted survival curves from proportional hazards models," *JAMA*, vol. 286, no. 12, pp. 1494–1497, 2001. 5, 30

[53] F. Harrell, *Regression Modeling Strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015. 5, 7

[54] S. Bennett, "Log-logistic regression models for survival data," *Applied Statistics*, pp. 165–171, 1983. 6

[55] C. de Boor, *A Practical Guide to Splines*, vol. 27. Springer-Verlag, 1978. 7, 8

[56] P. H. Eilers and B. D. Marx, "Flexible smoothing with B-splines and penalties," *Statistical Science*, pp. 89–102, 1996. 7

[57] C. Gu, *Smoothing spline ANOVA models*. Springer Science and Business Media, 2013. 7, 9, 34, 35

[58] P. H. Eilers and B. D. Marx, "Splines, knots, and penalties," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 6, pp. 637–653, 2010. 7, 8, 35

[59] P. L. Smith, "Splines as a useful and convenient statistical tool," *The American Statistician*, vol. 33, no. 2, pp. 57–62, 1979. 7

[60] C. Stone and C. Koo, "Additive Splines in Statistics Proceedings of the Statistical Computing Section," in *American Statistical Association*, pp. 45–48, 1985. 8, 22

[61] H. Binder and W. Sauerbrei, "Increasing the usefulness of additive spline models by knot removal," *Computational Statistics and Data Analysis*, vol. 52, no. 12, pp. 5305–5318, 2008. 8, 25

[62] M. Clements and X.-R. Liu, *rstpm2: Generalized Survival Models*, 2017. R package version 1.3.5. 8, 9, 17, 28

[63] M. J. Crowther and P. C. Lambert, "A general framework for parametric survival analysis," *Statistics in Medicine*, vol. 33, pp. 5280–5297, Dec. 2014. 9

[64] H. Charvat, N. Remontet, Land Bossard, L. Roche, O. Dejardin, B. Rachet, G. Launoy, and A. Belot, "A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates," *Statistics in Medicine*, vol. 35, no. 18, pp. 3066–3084, 2016. 9, 13

[65] A. Hennerfeind, A. Brezger, and L. Fahrmeir, "Geoadditive survival models," *Journal of the American Statistical Association*, vol. 101, no. 475, pp. 1065–1075, 2006. 9

[66] T. Kneib and L. Fahrmeir, "A mixed model approach for geoadditive hazard regression," *Scandinavian Journal of Statistics*, vol. 34, no. 1, pp. 207–228, 2007. 9

[67] S. N. Wood, N. Pya, and B. Säfken, "Smoothing parameter and model selection for general smooth models," *Journal of the American Statistical Association*, vol. 111, no. 516, pp. 1548–1563, 2016. 9, 28

[68] P. Royston and W. Sauerbrei, *Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables*, vol. 777. John Wiley & Sons, 2008. 9

[69] H. Binder, W. Sauerbrei, and P. Royston, "Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response," *Statistics in Medicine*, vol. 32, no. 13, pp. 2262–2277, 2013. 9

[70] U. S. Govindarajulu, E. J. Malloy, B. Ganguli, D. Spiegelman, and E. A. Eisen, "The comparison of alternative smoothing methods for fitting non-linear exposure-response relationships with Cox models in a simulation study," *The International Journal of Biostatistics*, vol. 5, no. 1, 2009. 9

[71] N. Younes and J. Lachin, "Link-based models for survival data with interval and continuous time censoring," *Biometrics*, pp. 1199–1211, 1997. 9, 12

[72] M. J. Crowther, M. P. Look, and R. D. Riley, "Multilevel mixed effects parametric survival models using adaptive Gauss–Hermite quadrature with application to recurrent events and individual participant data meta-analysis," *Statistics in Medicine*, vol. 33, no. 22, pp. 3844–3858, 2014. 10

[73] S. R. Hinchliffe and P. C. Lambert, "Flexible parametric modelling of cause-specific hazards to estimate cumulative incidence functions," *BMC Medical Research Methodology*, vol. 13, no. 1, p. 13, 2013. 10

[74] P. C. Lambert and P. Royston, "Further development of flexible parametric models for survival analysis," *Stata Journal*, vol. 9, no. 2, p. 265, 2009. 10

[75] T. M. Andersson, P. W. Dickman, S. Eloranta, and P. C. Lambert, "Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models," *BMC Medical Research Methodology*, vol. 11, no. 1, p. 96, 2011.

[76] S. Eloranta, P. C. Lambert, T. M. Andersson, K. Czene, P. Hall, M. Björkholm, and P. W. Dickman, "Partitioning of excess mortality in population-based cancer patient survival studies using flexible parametric survival models," *BMC Medical Research Methodology*, vol. 12, no. 1, p. 86, 2012. 10, 14

[77] D. O. Scharfstein, A. A. Tsiatis, and P. B. Gilbert, "Semiparametric efficient estimation in the generalized odds-rate class of regression models for right-censored time-to-event data," *Lifetime Data Analysis*, vol. 4, pp. 355–391, 1998. 10, 19

[78] T. Banerjee, M.-H. Chen, D. K. Dey, and S. Kim, "Bayesian analysis of generalized odds-rate hazards models for survival data," *Lifetime Data Analysis*, vol. 13, no. 2, pp. 241–260, 2007. 10

[79] J. Zhou, J. Zhang, and W. Lu, "An expectation maximization algorithm for fitting the generalized odds-rate model to interval censored data," *Statistics in Medicine*, vol. 36, no. 7, pp. 1157–1171, 2017. 10

[80] D. Zeng and D. Lin, "Maximum likelihood estimation in semiparametric regression models with censored data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, pp. 507–564, 2007. 10, 19

[81] D. Zeng and D. Lin, "Semiparametric transformation models with random effects for recurrent events," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 167–180, 2007. 10, 35

[82] D. Zeng, D. Lin, and X. Lin, "Semiparametric transformation models with random effects for clustered failure time data," *Statistica Sinica*, vol. 18, no. 1, pp. 355–377, 2008.

[83] D. Zeng and D. Lin, "Semiparametric transformation models with random effects for joint analysis of recurrent and terminal events," *Biometrics*, vol. 65, no. 3, pp. 746–752, 2009. 10

[84] X.-R. Liu, Y. Pawitan, and M. Clements, "Parametric and penalized generalized survival models," *Statistical Methods in Medical Research*, 2016. doi:10.1177/0962280216664760. 10, 26

[85] J. P. Fine and R. J. Gray, "A proportional hazards model for the subdistribution of a competing risk," *Journal of the American Statistical Association*, vol. 94, no. 446, pp. 496–509, 1999. 10

[86] T. A. Gerds, T. H. Scheike, and P. K. Andersen, "Absolute risk regression for competing risks: interpretation, link functions, and prediction," *Statistics in Medicine*, vol. 31, no. 29, pp. 3921–3930, 2012. 10

[87] F. Ambrogi and T. H. Scheike, "Penalized estimation for competing risks regression with applications to high-dimensional covariates," *Biostatistics*, vol. 17, no. 4, pp. 708–721, 2016. 10

[88] G. Bakoyannis, M. Yu, and C. T. Yiannoutsos, "Semiparametric regression on cumulative incidence function with interval-censored competing risks data," *Statistics in Medicine*, 2017. 10

[89] J. P. Klein, H. C. Van Houwelingen, J. G. Ibrahim, and T. H. Scheike, *Handbook of survival analysis*. CRC Press, 2016. 10, 11

[90] Y. Pawitan, *In All Likelihood: Statistical modelling and inference using likelihood*. Oxford University Press, 2001. 11

[91] S. N. Wood, *Core Statistics*, vol. 6. Cambridge University Press, 2015. 11

[92] P. H. Eilers and B. D. Marx, "Flexible smoothing with B-splines and penalties," *Statistical Science*, vol. 11, pp. 89–102, 1996. 11, 26

[93] P. J. Verweij and H. C. van Houwelingen, "Penalized likelihood in Cox regression," *Statistics in Medicine*, vol. 13, pp. 2427–2436, 1994. 11, 27

[94] X.-R. Liu, Y. Pawitan, and M. S. Clements, "Generalized survival models for correlated time-to-event data," *Statistics in Medicine*, 2017. doi:10.1002/sim.7451. 12, 25, 26

[95] T. Martinussen and T. H. Scheike, *Dynamic Regression Models for Survival Data*. Springer Science & Business Media, 2007. 12, 13

[96] P. K. Andersen and L. T. Skovgaard, *Regression with Linear Predictors*. Springer Science & Business Media, 2010. 12, 14, 20

[97] O. Aalen, *Statistical inference for a family of counting process*. PhD thesis, Ph. D. Dissertation, Univ. of California, Berkeley, 1975. 13

[98] T. H. Scheike and M.-J. Zhang, "Analyzing competing risk data using the R timereg package," *Journal of Statistical Software*, vol. 38, no. 2, 2011. 13

[99] R. Giorgi, M. Abrahamowicz, C. Quantin, P. Bolard, J. Esteve, J. Gouvernet, and J. Faivre, "A relative survival regression model using B-spline functions to model non-proportional hazards," *Statistics in Medicine*, vol. 22, no. 17, pp. 2767–2784, 2003. 13

[100] D. Stocken, A. Hassan, D. Altman, L. Billingham, S. Bramhall, P. Johnson, and N. Freemantle, "Modelling prognostic factors in advanced pancreatic cancer," *British Journal of Cancer*, vol. 99, no. 6, p. 883, 2008. 13

[101] S. Lagakos, "Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable," *Statistics in Medicine*, vol. 7, pp. 257–274, 1988.

[102] E. L. Turner, J. E. Dobson, and S. J. Pocock, "Categorisation of continuous risk factors in epidemiological publications: a survey of current practice," *Epidemiologic Perspectives & Innovations*, vol. 7, no. 1, p. 9, 2010.

[103] R. H. Groenwold, O. H. Klungel, D. G. Altman, Y. van der Graaf, A. W. Hoes, and K. G. Moons, "Adjustment for continuous confounders: an example of how to prevent residual confounding," *Canadian Medical Association Journal*, vol. 185, no. 5, pp. 401–406, 2013.

[104] K. Leffondré, K. J. Jager, J. Boucquemont, V. S. Stel, and G. Heinze, "Representation of exposures in regression analysis and interpretation of regression coefficients: basic concepts and pitfalls," *Nephrology Dialysis Transplantation*, vol. 29, no. 10, pp. 1806–1814, 2014. 13

[105] T. M. Therneau and P. M. Grambsch, *Modeling survival data: extending the Cox model*. Springer Science & Business Media, 2000. 14, 47

[106] D. Y. Lin, B. M. Psaty, and R. A. Kronmal, "Assessing the sensitivity of regression results to unmeasured confounders in observational studies," *Biometrics*, pp. 948–963, 1998. 14

[107] S. G. Hilsenbeck, P. M. Ravdin, C. A. de Moor, G. C. Chamness, C. K. Osborne, and G. M. Clark, "Time-dependence of hazard ratios for prognostic factors in primary breast cancer," *Breast Cancer Research and Treatment*, vol. 52, no. 1-3, pp. 227–237, 1998. 14

[108] W. Vach, *Regression models as a tool in medical research*. CRC Press, 2012. 14

[109] Y. Wang, M. Lee, P. Liu, L. Shi, Z. Yu, Y. A. Awad, A. Zanobetti, and J. D. Schwartz, "Doubly robust additive hazards models to estimate effects of a continuous exposure on survival," *Epidemiology*, vol. 28, no. 6, pp. 771–779, 2017. 14

[110] M. Zhang and M. Davidian, "Smooth semiparametric regression analysis for arbitrarily censored time-to-event data," *Biometrics*, vol. 64, no. 2, pp. 567–576, 2008. 19

[111] A. Komárek, E. Lesaffre, and J. F. Hilton, "Accelerated failure time model for arbitrarily censored data with smoothed error distribution," *Journal of Computational and Graphical Statistics*, vol. 14, no. 3, 2005. 19

[112] T. M. Apostol, "Calculus, Vol. 1: One-Variable Calculus, with an Introduction to Linear Algebra," *Jon Wiley & Sons*, 1967. 21

[113] P. J. Green and B. W. Silverman, *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach*. CRC Press, 1993. 22, 24

[114] "Draw a radial basis function diagram in LyX." https://tex.stackexchange.com/questions /174435/draw-a-radial-basis-function-diagram-in-lyx. Accessed: 2017-10-12. 23

[115] S. N. Wood, "Stable and efficient multiple smoothing parameter estimation for generalized additive models," *Journal of the American Statistical Association*, vol. 99, no. 467, pp. 673–686, 2004. 25

[116] W. Sauerbrei, C. Meier-Hirmer, A. Benner, and P. Royston, "Multivariable regression model building by using fractional polynomials: description of SAS, STATA and R programs," *Computational Statistics and Data Analysis*, vol. 50, no. 12, pp. 3464–3485, 2006. 25

[117] H. Binder and W. Sauerbrei, "Adding local components to global functions for continuous covariates in multivariable regression modeling," *Statistics in Medicine*, vol. 29, no. 7-8, pp. 808–817, 2010. 25

[118] A. W. Van der Vaart, *Asymptotic Statistics*. Cambridge University Press, 2000. 26

[119] P. Joly, D. Commenges, and L. Letenneur, "A penalized likelihood approach for arbitrarily censored and truncated data: application to age-specific incidence of dementia," *Biometrics*, vol. 54, no. 1, pp. 185–194, 1998. 27

[120] H. Wickham, *Advanced R*. CRC Press, 2014. 28

[121] R. L. Prentice, M. Pettinger, and G. L. Anderson, "Statistical issues arising in the Women's Health Initiative," *Biometrics*, vol. 61, no. 4, pp. 899–911, 2005. 29

[122] G. Wei and D. E. Schaubel, "Estimating cumulative treatment effects in the presence of nonproportional hazards," *Biometrics*, vol. 64, no. 3, pp. 724–732, 2008.

[123] X. Tang and A. S. Wahed, "Cumulative hazard ratio estimation for treatment regimes in sequentially randomized clinical trials," *Statistics in Biosciences*, pp. 1–18, 2013. 29

[124] R. Bender, T. Augustin, and M. Blettner, "Generating survival times to simulate Cox proportional hazards models," *Statistics in Medicine*, vol. 24, no. 11, pp. 1713–1723, 2005. 30, 34

[125] L. Fahrmeir, T. Kneib, and S. Konrath, "Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection," *Statistics and Computing*, vol. 20, no. 2, pp. 203–219, 2010. 34

[126] T. Cai and R. A. Betensky, "Hazard regression for interval-censored data with penalized spline," *Biometrics*, vol. 59, pp. 570–579, 2003. 34, 35

[127] C. Gu, "Smoothing Spline ANOVA Models: R Package gss," *Journal of Statistical Software*, vol. 58, no. 5, pp. 1–25, 2014. 35

[128] H. Putter and H. C. Van Houwelingen, "Dynamic frailty models based on compound birth–death processes," *Biostatistics*, vol. 16, no. 3, pp. 550–564, 2015. 37

[129] The Diabetic Retinopathy Study Research Group, "Preliminary report on effects of photocoagulation therapy," *American Journal of Ophthalmology*, vol. 81, no. 4, pp. 383–396, 1976. 37

[130] J. F. Ludvigsson, C. Almqvist, A.-K. E. Bonamy, R. Ljung, K. Michaëlsson, M. Neovius, O. Stephansson, and W. Ye, "Registers of the Swedish total population and their use in medical research," *European Journal of Epidemiology*, vol. 31, no. 2, pp. 125–136, 2016. 41

[131] H. Grönberg, J. Adolfsson, M. Aly, T. Nordström, P. Wiklund, Y. Brandberg, J. Thompson, F. Wiklund, J. Lindberg, M. Clements, *et al.*, "Prostate cancer screening in men aged 50–69 years (STHLM3): a prospective population-based diagnostic study," *The Lancet Oncology*, vol. 16, no. 16, pp. 1667–1676, 2015. 41

[132] A. Allignol, M. Schumacher, J. Beyersmann, *et al.*, "Empirical transition matrix of multi-state models: the etm package," *Journal of Statistical Software*, vol. 38, no. 4, pp. 1–15, 2011. 41

[133] S. N. Wood, Y. Goude, and S. Shaw, "Generalized additive models for large data sets," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 64, no. 1, pp. 139–155, 2015. 44

[134] J. M. Neuhaus and J. D. Kalbfleisch, "Between-and within-cluster covariate effects in the analysis of clustered data," *Biometrics*, pp. 638–645, 1998. 44

[135] A. Sjölander, P. Lichtenstein, H. Larsson, and Y. Pawitan, "Between–within models for survival analysis," *Statistics in Medicine*, vol. 32, no. 18, pp. 3067–3076, 2013. 44

[136] A. Sjölander, T. Frisell, and S. Öberg, "Causal interpretation of between-within models for twin research," *Epidemiologic Methods*, vol. 1, no. 1, pp. 217–237, 2012. 44

[137] M. Gerster, M. Madsen, and P. K. Andersen, "Matched survival data in a co-twin control design," *Lifetime Data Analysis*, vol. 20, no. 1, pp. 38–50, 2014. 44

[138] J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer, 2006. 44

[139] J. O. Ramsay, "Monotone regression splines in action," *Statistical Science*, pp. 425–441, 1988. 45

[140] N. Pya and S. N. Wood, "Shape constrained additive models," *Statistics and Computing*, vol. 25, no. 3, pp. 543–559, 2015. 45

[141] A. Van der Vaart, *Asymptotic Statistics*. Cambridge university press, 2000. 46

[142] D. Freedman, "On the so-called Huber sandwich estimator and robust standard errors," *The American Statistician*, vol. 60, no. 4, pp. 299–302, 2006. 46

[143] T. M. Therneau, P. M. Grambsch, and T. R. Fleming, "Martingale-based residuals for survival models," *Biometrika*, vol. 77, no. 1, pp. 147–160, 1990. 47

[144] G. Cortese and T. Scheike, "Dynamic regression hazards models for relative survival," *Statistics in Medicine*, vol. 27, no. 18, pp. 3563–3584, 2008. 47

[145] D. Lin and C. Spiekerman, "Model checking techniques for parametric regression with censored data," *Scandinavian Journal of Statistics*, vol. 23, pp. 157–177, 1996. 47

[146] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996. 47

[147] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005. 47

[148] Z. He, W. Tu, S. Wang, H. Fu, and Z. Yu, "Simultaneous variable selection for joint models of longitudinal and survival outcomes," *Biometrics*, vol. 71, no. 1, pp. 178–187, 2015. 47

[149] T. Emura and Y.-H. Chen, “Gene selection for survival data under dependent censoring: A copula-based approach,” *Statistical Methods in Medical Research*, vol. 25, no. 6, pp. 2840–2857, 2016. 48