Assessment of time–varying long–term effects of therapies and prognostic factors

Dissertation
by

Anika Buchholz

Submitted to
Fakultät Statistik,
Technische Universität Dortmund
in Fulfillment of the Requirements for the Degree of
Doktorin der Naturwissenschaften

Freiburg 2010

Referees:

Prof. Dr. Katja Ickstadt

Prof. Dr. Dieter Hauschke

Prof. Dr. Martin Schumacher

Date of Oral Examination:

July 5^{th}, 2010

Contents

Chapter 1

Introduction

In many medical applications, the common focus of analyses is to model the impact of prognostic factors and therapies on the time to a certain event such as relapse or death. A standard approach for such analyses is the Cox proportional hazards model (Cox, 1972), which evaluates the instantaneous risk for the event of interest. This model is based on assumptions which, for example, imply that the effects of prognostic factors and therapies are constant over time (proportional hazards assumption). However, with long-term follow-up this assumption may be questionable and erroneously assuming proportional hazards (PH), i.e. time-constant effects, results in incorrect models. However, mismodelling the shape of time-varying effects can likewise lead to incorrect models and false conclusions thereof. Hence, beyond detecting time-varying effects, appropriate modelling of their shape is at least as important.

Time-varying effects of prognostic factors have been detected in a variety of medical fields. For instance, the effects of oestrogen receptor and tumour size in breast cancer have been reported to change over time (Hilsenbeck et al., 1998; Coradini et al., 2000). Other examples include the effects of prothrombin time in primary biliary cirrhosis (Abrahamowicz et al., 1996), the Karnofsky performance status in ovarian cancer studies (Verweij and van Houwelingen, 1995) and diabetes on mortality after coronary artery bypass graft surgery (Gao et al., 2006).
Besides gaining insight into the impact of prognostic factors on survival, the accurate assessment of therapy effects is another important goal. As D'Agostino (2009) states: "To advance our understanding of treatments for diseases that progress slowly but that are ultimately debilitating, such as Alzheimer's disease, Parkinson's disease, rheumatoid arthritis, and chronic obstructive pulmonary disease, it is essential to evaluate the disease-modifying effects of administered treatments. It is also essential to separate these effects from the short-term beneficial effects on symptoms that such treatments may provide."

Recently, the importance to account for non-PH has also been recognised in microarray survival studies (Dunkler et al., 2010), where the PH assumption is unlikely to hold for each gene. Ignoring the violation of PH of some genes may lead to false conclusions about their importance.

The variety of methods to check for non-PH is broad, see e.g. Ng'Andu (1997) for an overview of different tests. However, significant non-proportionality does not necessarily involve the presence of time-varying effects. Spurious time-varying effects may also be introduced by mismodelling other parts of the data, such as omission of an important covariate, an incorrect functional form of covariates or an inappropriate survival model (Keiding et al., 1997; Therneau and Grambsch, 2000, chap. 6). All of these issues are important and mutually interact.
If the detected non-proportionality is due to a real time-varying effect, appropriate modelling of this effect is an even more important task, as usually the interest lies not only in the presence of a time-varying effect, but rather in the interpretation of its shape. Mismodelling the functional form of effects may lead to incorrect conclusions about prognostic factors and therapies, e.g. false believe in the benefit of a therapy may in the last resort lead to an increased mortality of patients. To cope with this task, several different approaches have been proposed, which extend the Cox model to allow for time-varying effects. So far, there is a lack of knowledge of properties of several of these approaches and advice on which techniques to use is rare.

Time-varying effects are not to be mistaken with time-dependent covariates, i.e. covariates that change values over time. In this context, only time-fixed covariates with time-varying effect are considered, i.e. the values of covariates are fixed and do not change over time, but their effects do. Of course, an extension to time-dependent covariates with time-varying effects is possible, but is not considered here.

The aim of this thesis is to assess the properties of the Fractional Polynomial Time (FPT) algorithm (Sauerbrei et al., 2007), which tests and models the time-varying effect of a single covariate based on fractional polynomials, and its multivariable extension in a large simulation study. Furthermore, to give guidance on different techniques, several recent approaches for modelling time-varying effects are compared in a data example and a simulated data set. The main focus in both investigations will be on the shape of the selected effects, to account for the importance of appropriate modelling of time-varying effects rather than mere testing. An investigation in this vein has, to our knowledge, not been accomplished before.

Chapter 2 gives a brief overview of different methods for modelling time-varying effects, which are based on techniques like parametric functions of time, piecewise constant effects, fractional polynomials, splines, cumulative effects, local linear estimation or neural networks. A special focus is thereby on the different approaches under investigation in the subsequent

chapters.
The assessment criteria for selected models and individual (time-varying) effects are introduced in Chapter 3. As graphical comparison of individual effects is limited in simulation studies, we additionally utilise a measure which quantifies the distance of estimated effects to the true effect by the weighted area between both functions. Furthermore, the prediction performance of the complete model is assessed in terms of the prediction error.
In Chapter 4, we compare five recent approaches in example data sets including the Rotterdam breast cancer series, a prognostic factor study. To gain further insights into their performance, all approaches are additionally applied to a simulated data set. We comment on the practical applicability of the investigated approaches and demonstrate assets and drawbacks to give guidance on which approaches seem to be most suitable.
In Chapter 5, we assess the stability of effects estimated by FPT in bootstrap samples of the Rotterdam breast cancer series. To account for model selection uncertainty and to enhance the reliability of time-varying effects, we propose a bootstrap based selection and modelling strategy.
The results of a large simulation study on the properties of the FPT algorithm and its multivariable extension are presented in Chapter 6. The performance is assessed in terms of type I and II error. As the definition of a type II error in the framework of time-varying effects is difficult, we consider two different versions. The usual definition is the failure to detect the time-dependency. Additionally we account for the shape of the selected effects by introducing a qualitative type II error. For evaluation of the effect estimates and complete models, we use the distance to the true effect and the prediction error, respectively. We had in mind to include a competitive approach in the simulation study. However, several practical problems arose in the application of this approach (Scheike and Martinussen, 2004) to the simulated data, which are briefly presented subsequent to the simulation study.
We conclude with a discussion summarising all findings and commenting on future perspectives in Chapter 7.

Chapter 2

Analysis of time-varying effects

2.1 The Cox model

The standard model for analysing survival data is the well-known Cox proportional hazards (CoxPH) model (Cox, 1972)

$$\lambda(t|X) = \lambda_0(t)\exp(\sum_{i=1}^{q} X_i\beta_i), \tag{2.1}$$

with covariates X_i, $i = 1,\ldots,q$, covariate matrix $X = (X_1 \ldots X_q)$, effects β_i and unspecified baseline hazard $\lambda_0(t)$.

The effects β_i are estimated based on the partial likelihood, which for n_e distinct ordered event times $t_{(1)},\ldots,t_{(n_e)}$ is equal to

$$PL(\beta) = \prod_{j=1}^{n_e} \frac{\prod_{k=1}^{d_j}\exp\left(\sum_{i=1}^{q} X_{ji}\beta_i\right)}{\{\sum_{l\in R(t_{(j)})}\exp\left(\sum_{i=1}^{q} X_{li}\beta_i\right)\}^{d_j}}, \tag{2.2}$$

where d_j is the number of events at time $t_{(j)}$ and $R(t_{(j)})$ the risk set at $t_{(j)}$, i.e. the set of all individuals still at risk just prior to $t_{(j)}$.
The maximum partial likelihood estimates for β_i are found by maximising (2.2), i.e. by solving the set of q score equations $U(\beta_i) = 0$, $i = 1,\ldots,q$. $U(\beta_i)$ are the partial derivatives of the log partial likelihood $\log(PL(\beta))$ with respect to β_i. The score equations are usually solved numerically using a Newton-Raphson algorithm, which starts with initial estimates $\beta_i^{(0)}$, e.g. $\beta_i^{(0)} = 0$, and iteratively updates them based on the score vectors and the Hessian matrix H of mixed second partial derivatives of the log partial likelihood as

$$\beta_i^{(j+1)} = \beta_i^{(j)} - H^{-1}(\beta_i^{(j)})U(\beta_i^{(j)})$$

until convergence, i.e. until the log partial likelihood stabilises (for details see e.g. Therneau and Grambsch, 2000, chap. 3 or Klein and Moeschberger, 2005, chap. 8). A variant of the Newton-Raphson algorithm is the Fisher scoring algorithm, which replaces the Hessian matrix by its expectation.

The baseline hazard $\lambda_0(t)$, i.e. the cumulative baseline hazard $\int_0^t \lambda_0(s)ds$, is usually estimated afterwards using the Breslow estimate

$$\hat{H}_0(t) = \sum_{t_{(j)} \leq t} \frac{d_j}{\sum_{l \in R(t_{(j)})} \exp\left(\sum_{i=1}^{q} X_{li}\hat{\beta}_i\right)}. \tag{2.3}$$

Estimation of the baseline hazard is, for example, required to obtain survival probabilities

$$S(t|X) = \exp\left\{-\int_0^t \lambda_0(s)ds \exp\left(\sum_{i=1}^{q} X_i\beta_i\right)\right\}. \tag{2.4}$$

The CoxPH model is based on several assumptions such as the proportional hazards (PH) assumption, i.e. the ratio of the hazards of two individuals is assumed to be constant. However, if the effects of covariates change over time, this assumption is violated and an extension of model (2.1) is required, which allows for time-varying effects.

2.2 Central issues of multivariable model building

When thinking about multivariable model building strategies, time-varying effects are not the only issue one has to deal with. In multivariable model building, three central issues have to be considered:

(i) Which covariates have to be included into the model?

(ii) Is the effect of continuous covariates linear as assumed in model (2.1) or is a non-linear functional form $f(X_i) \neq X_i$ more appropriate?

(iii) Do effects of covariates vary in time and if so, how do they look like?

All three issues are related to each other. Spurious time-varying effects may appear due to mismodelling of the first two aspects, i.e. omitting an important variable or assuming an incorrect functional form as discussed, for example, by Keiding et al. (1997), Therneau and Grambsch (2000, chap. 6.6) and Abrahamowicz and MacKenzie (2007). Vice versa, erroneously assuming PH in the presence of time-varying effects also results in mismodelling of the data.

Several approaches have been proposed which concentrate on either of the three issues, but only few consider all of them. One approach that addresses all three issues is the Multivariable Fractional Polynomial Time (MFPT) approach (Sauerbrei et al., 2007). It extends the Multivariable Fractional Polynomial (MFP) approach (Sauerbrei and Royston, 1999) for modelling of the functional form of continuous covariates (non-linear effects) by selecting and modelling time-varying effects based on fractional polynomials (FPs).

The **MFPT algorithm** consists of three steps, which gradually extend the model:

Step 1: Select covariates with influence on survival time and functional forms of continuous covariates by applying the MFP algorithm (Sauerbrei and Royston, 1999). The resulting model assumes PH, but relaxes the linearity assumption.

Step 2: Add covariates with a short-term effect only. Restrict the data to a short–term period [0, $\tilde{t}$] by censoring all observations at $\tilde{t}$, where $\tilde{t}$ may be defined by the first half of events. Rerun the MFP algorithm on this restricted data set, keeping covariates and transformations selected in Step 1 fix, but re-estimating the regression coefficients.

Step 3: Identify and model time-varying effects for covariates selected in Steps 1 and 2 (see Section 2.5.1).

Royston and Sauerbrei (2005) investigate the MFP approach (Step 1) by bootstrap resampling with respect to the stability of selected models and FP transformations of covariates. They show that MFP can find flexible functions for covariates if indicated by the data, without incurring major instability of functional forms or models.

A few of the methods introduced in the following section also allow for testing and/or modelling of non-linear covariate effects, but most approaches ignore these issues and focus on modelling time-varying effects only. They take a time-fixed model as their starting point, which may already include non-linear functional forms of covariates. In the analysis of a real life example in Chapter 4, we apply the first two steps of the MFPT algorithm to develop such a time-fixed model.
Since the focus of this work lies on the identification and modelling of time-varying effects, we will in the sequel concentrate on this topic.

2.3 Approaches for modelling time-varying effects

The variety of approaches for modelling time-varying effects is broad. Often the methodology is transferred from approaches for modelling non-linear functional forms of covariates. Unfortunately, several of the approaches are introduced in univariate settings and lack suitable strategies for multivariable model building, which limits their practical applicability.

In the sequel a selection of several approaches for modelling time-varying effects is presented, which is not claimed to be complete, but shall give an overview of the variety of different techniques used in this field by discussing some approaches exemplarily.
To assess the performance and practical applicability of the different techniques, we chose five recently proposed approaches representative for the different techniques and compare their performance in data examples. We decided in favour of promising flexible techniques, such as splines, FPs and non-parametric cumulative regression functions, including frequentist as well as Bayesian methods. These approaches are introduced in more detail in Section 2.5.

2.3.1 Parametric functions of time

The very first proposal for an extension of the CoxPH model is given by Cox (1972) in his original paper. He proposed to introduce time-dependent components based on a predefined function of time in case of non-proportional hazards. This corresponds to the inclusion of a time-dependent covariate $X_i f_i(t)$ representing an interaction between the predictor and a parametric function of time $f_i(t)$. Hence, model (2.1) is modified to

$$\lambda(t|X) = \lambda_0(t) \exp\left(\sum_{i=1}^{q} X_i f_i(t) \beta_i\right) = \lambda_0(t) \exp\left(\sum_{i=1}^{q} X_i \beta_i(t)\right), \tag{2.5}$$

with the time-varying effects $\beta_i(t) = \gamma_{i0} + \gamma_{i1} f_i(t)$. Model (2.5) simultaneously provides a check of the PH assumption by testing on $\gamma_{i1} = 0$.
This approach is still used, as for example by Putter et al. (2005), who choose $f(t) = \log(t+1)$. Their work, though, puts much emphasis on estimating the baseline hazard together with covariate effects to obtain a complete picture of the underlying structure.

The method is easy to implement in standard statistical software, but inference is highly dependent on the choice of the parametric function $f(t)$. As the shape of the estimated time-varying effect is determined by the specified function, an inappropriate choice of $f(t)$ may lead to incorrect interpretation of results. Often several alternatives for $f(t)$ are tried to overcome this problem. However, different choices of $f(t)$ may also influence the corresponding test on PH and may lead to different decisions, as the test depends on a specific departure from the null hypothesis. Furthermore, some functions may fit equally well or none might be able to describe the underlying effect, if its shape is too complex.

To enhance the flexibility in multivariable analyses, Quantin et al. (1999) allow different functions $f(t)$ for the time-varying effects of different covariates based on the best-fitting function selected from $\{t, \log(t), t^2, 1/t\}$ in univariate analyses. Furthermore, Therneau and Grambsch (2000, chap. 6) give some guidance on how to choose $f(t)$. Besides a choice based on

theoretical considerations, they propose to use smoothed Schoenfeld residuals to explore the shape of departure from PH.

2.3.2 Piecewise constant effects

Another straight-forward technique is partitioning of the time axis, also called piecewise constant effects. Based on the idea that the PH assumption holds at least over short time periods, separate effects are fitted for each period (under the PH assumption) resulting in a step-function for $\beta(t)$. Although piecewise constant effects are very popular, they have some drawbacks. The choice of the number of jump times is crucial. To produce reliable estimates, a sufficient number of events in each interval is required. Otherwise, estimated effects may be unstable or the algorithm may even fail to converge. This problem enhances in multivariable analyses, where the same additionally applies for each subgroup of covariate combinations.

Proposals on the number and position of jump times are manifold. Anderson and Senthilselvan (1982), for example, propose a stepwise regression model with as few steps as possible, e.g. two or three, to avoid estimation problems. To verify the PH assumption, they investigate the residuals. Moreau et al. (1985) generalise this approach to a larger number of intervals, whereas O'Quigley and Pessione (1991) consider a special case by limiting the change of coefficients in the two-step model to a mere change of sign, i.e. $\beta(t) = \beta_1 \left(I(t \leq \tau) - I(t > \tau)\right)$ for jump time τ.

Other approaches combine a larger number of jump times with penalised likelihood techniques which ensure smoothness of the piecewise constant effects by penalising too abrupt jumps. Verweij and van Houwelingen (1995), for example, suggest to estimate coefficients at each event time, using a penalised partial likelihood approach with first order difference penalty for adjacent values of coefficients. The smoothness parameters are determined based on the AIC.

Due to the potential instability of effect estimates and their sensibility to the number and position of jump times this class of approaches is not considered further.

2.3.3 Fractional polynomials

Often smooth functions are preferred to piecewise constant effects. One technique providing smooth estimates of $\beta(t)$ are fractional polynomials (Royston and Altman, 1994), which have been originally proposed for modelling of non-linear functional forms of covariates. Fractional polynomials (FPs) are an extension of conventional polynomials allowing for non-integer and negative powers. FP based approaches for modelling time-varying effects extend Cox's

idea of using a pre-defined function $f(t)$. Instead of choosing one function prior to fitting the model, they include multivariable selection procedures that determine the best-fitting function $f(t)$ out of a pre-defined set of functions for each variable in turn, including a test on PH. Two such approaches have been proposed by Berger et al. (2003) and Sauerbrei et al. (2007) and are described in more detail in Sections 2.5.2 and 2.5.1, respectively. FPs have the advantage of providing simple functional forms of estimated time-varying effects which are easy to interpret. As the class of FPs offers a broad variety of curve shapes, the potential drawback of fitting a global function is deemed to be of minor importance and may be outweighed by a better generalisability to other data sets.

2.3.4 Splines

Another large group of approaches for modelling time-varying effects is based on splines. Splines are a flexible non-parametric tool to identify functional relationships and produce visibly smooth curves. They are constructed from polynomial pieces joined at certain values (knots). The choice of the number and position of these knots is crucial, as they influence the fitted curve. Too many knots lead to overfitting of the data, while too few knots result in an underfitting. Solutions to this problem tend to two directions. Either relatively few knots are used or a relatively large number of knots is combined with a smoothness penalty, resulting in penalised splines (Eilers and Marx, 1996).

For example, Hess (1994) and Heinzl and Kaider (1997) use (unpenalised) natural cubic splines with 3 to 5 knots. Their proposals include a formal test of the PH assumption by testing on the spline coefficients being equal to zero. Abrahamowicz et al. (1996) prefer quadratic B-splines with no more than two knots.

Hastie and Tibshirani (1993) propose a penalised partial likelihood approach based on natural cubic splines, with knots at unique event times and second order penalty based on the squared second derivative of the time-varying effects. The values of the smoothing parameters are selected by specifying the degrees of freedom for the smooth, i.e. the effective number of parameters. Gray (1992) uses a similar method to determine the smoothing parameters, but bases the estimation of time-varying effects on B-splines of degree two and zero (i.e. piecewise constant effects) with a first order integral and first order difference penalty, respectively. The number of knots is limited to ten.

Brown et al. (2007) propose a mixed model approach, which assumes that some effects are random. They use linear B-splines or, equivalently, truncated polynomials penalised by a difference matrix or identity matrix, respectively, to approximate the time-varying effects. The number of knots are chosen to be $\min(\frac{n_e}{4}, 35)$, with n_e the number of distinct survival times as proposed by Ruppert et al. (2003, p. 126). The smoothing parameters relate to the

variance components in the mixed model framework and are estimated in a hybrid approach controlled by the AIC. The estimation procedure cycles between estimating the regression coefficients for given smoothing parameters and vice versa, checking the AIC at each iteration. The procedure stops if the AIC can no longer be improved.

2.3.5 Reduced Rank models

Another technique allowing for time-varying effects are Reduced Rank models (Perperoglou et al., 2006b). In this approach, time-varying effects are modelled as (pre-defined) covariate by time function interactions, e.g. based on splines. The main idea is to introduce a structure matrix containing the regression coefficients of all covariate by time function interactions. The rank of this structure matrix (i.e. the number of parameters to be estimated) is reduced by factorisation, resulting in more stable and parsimonious models. Details on this approach can be found in Section 2.5.5.

2.3.6 Cumulative regression effects

Scheike and Martinussen (2004) propose an approach which is mainly directed at successive testing of time-varying effects based on the non-parametric cumulative regression coefficients $B(t) = \int_0^t \beta(s)ds$. In multivariable analyses, time-varying effects can be selected via backward elimination. The time-varying effects $\beta(t)$ themselves can be obtained via a kernel estimator which smoothly downweights distant data points and requires specification of an appropriate bandwidth. The approach is discussed in more detail in Section 2.5.4.

2.3.7 Local linear estimation

Other approaches include local linear estimation techniques as proposed by Cai and Sun (2003). They use a weighted local partial likelihood in which a kernel function downweights distant data points. Time-varying effects for a time t are approximated by a linear function using a first order Taylor expansion around t. The partial likelihood estimate for the linear function is calculated using the observed event times within a window around each t. The estimated linear function at t is taken as the estimate of $\beta(t)$. Tian et al. (2005) further investigate this approach. They propose to choose the smoothing parameter (bandwidth) by cross-validation and construct confidence bands and tests for time-varying effects. In multivariable analyses time-varying effects may be selected in a backward elimination manner. Although this approach seems to be rather flexible, to our knowledge, no software tools are available for it. Therefore, we do not consider it further.

2.3.8 Neural networks

Furthermore, time-varying effects are considered in the framework of feed forward neural networks (Liestøl et al., 1994; Biganzoli et al., 1998). The proposed networks calculate linear combinations of the input nodes (covariates) with individual weights (regression coefficients) for each node. These linear combinations, also called hidden nodes, are then transformed by a so called activation function, usually the logistic function, and are again linearly combined to give the output node(s). Under the PH assumption, all weights (regression coefficients) for the same input node or hidden node are identical. Dropping this constraint yields time-varying effects. Penalty terms may be introduced to penalise deviations from PH. Furthermore, the degree of smoothing is also influenced by the number of hidden nodes. Both approaches, though, do not provide selection strategies for time-varying effects and require grouping of survival times and hence are not considered in our investigations.

2.3.9 Bayesian inference

Besides the frequentist approaches discussed so far, similar approaches have been developed in the Bayesian framework. While frequentist methods regard the covariate effects as fixed, but unknown, constants, Bayesian methods are based on the idea that all parameters whose true value is uncertain, such as the covariate effects, are random variables and have probability distributions (see e.g. Bland and Altman, 1998, for a short discussion of both concepts).

McKeague and Tighiouart (2000), for example, proposed a non-parametric Bayesian approach based on step functions for the time-varying effect and the baseline hazard, where the number and position of jump times are taken as random. The levels of the step functions follow a Gaussian Markov random field prior with a pairwise dependency structure imposed on adjacent values. Haneuse et al. (2008) extend this approach to allow for separate time-scales for the baseline hazard and the time-varying effect of a time-dependent covariate. They present an example, where the time scale of the baseline hazard is the usual one, with its origin at study entry. The time scale of the time-varying effect of the time-dependent exposure (transplant status), though, has its origin at the onset of the exposure (the time of transplantation) and runs in parallel to the baseline time scale.

Costa and Shaw (2009) use Bayesian penalised spline models based on cubic splines. They use a moderate number of knots (e.g. 10) placed at quantiles of event times and a first or second order integral based penalty. A Gaussian prior is assumed for the spline parameters with a conjugate gamma prior for the smoothing parameters. The authors further introduce a double penalty as the sum of the first and second order penalty, which is claimed to be useful in situations where the single penalty models do not attain the desired limit of smoothness.

As setting a prior for the pairs of smoothing parameters for this double penalty is not straightforward, an empirical Bayes method is applied. This method estimates the hyperparameters (i.e. the smoothing parameters) using the data at hand and inserts them in the prior. The spline parameters are obtained via Markov chain Monte Carlo (MCMC) techniques, which repeatedly sample the parameters from probability distributions by constructing a Markov chain that has the desired distribution as its stationary distribution. The AIC can be used to check whether inclusion of time-varying effects improves the model fit compared to time-constant effects.

Structured additive regression models including Bayesian penalised splines provide another alternative for modelling time-varying effects. Inference for these models can be performed either with a full Bayes (Hennerfeind et al., 2006) or an empirical Bayes approach (Kneib and Fahrmeir, 2007). For full Bayes inference, the regression parameters as well as their variance (or smoothing) parameters are considered as random variables and are provided with suitable priors and hyperpriors which express the prior beliefs about the parameters. All parameters are jointly estimated via MCMC simulation techniques. The empirical Bayes approach differentiates between the parameters of primary interest (the regression parameters) and the hyperparameters (variance or smoothing parameters). The latter are considered as constants and are estimated in advance from the data by restricted maximum likelihood. Since this approach is based on optimising likelihood-based criteria, it avoids potential problems with convergence and mixing of Markov chains (Kneib and Fahrmeir, 2007). Both approaches are based on penalised B-splines with a second order random walk penalty and a moderately large number of knots. Kneib and Fahrmeir (2007) show that both methods perform similar in terms of the posterior mode / posterior mean estimates. The empirical Bayes approach is introduced in more detail in Section 2.5.3 including a multivariable model building procedure recently proposed by Hofner et al. (2010).
Alternatively, He et al. (2010) suggest a linear Bayesian estimation approach for Bayesian dynamic survival models. Time-varying effects are modelled by first order random walks, i.e. piecewise constant effects. The major difference to MCMC estimation techniques is that the smoothing parameters must be pre-specified. The optimal smoothing parameters are chosen in terms of the minimum mean square error based on the changing pattern of the estimated coefficients.

2.3.10 Average hazard ratio

In general, models incorporating time-varying effects are much more complex than the standard CoxPH model. In situations with very small sample size or high-dimensional data this is problematic. For such cases, or when the shape of the underlying time-varying effect is of little interest, Schemper et al. (2009) propose average hazard ratios by weighted Cox regres-

sion. This method aims to provide a simple and interpretable multivariable analysis without introducing further parameters. The weights reflect the relative importance of hazard ratios over time and may, for example, be proportional to the number of individuals at risk. The authors note that under the PH assumption, weighted CoxPH approaches entail some loss of efficiency, but deem it to be small in practical applications. For non-PH on the contrary, average hazard ratios are claimed to provide an intuitive interpretation and, for converging hazards, improved power. Since we are particularly interested in modelling the shape of the time-varying effects, this approach is not suitable for our investigations.

2.4 Properties and comparisons

Although the literature on modelling time-varying effects is manifold, theoretical results are rare and we are not aware of larger simulation studies on properties of the approaches. Furthermore, sensible comparisons of different approaches or advice on which techniques to use are not supported by convincing studies.

Quantin et al. (1999) present a comparison of piecewise constant effects, different pre-specified parametric functions of time and regression splines in a data example. Due to the limited flexibility of the two former methods, they suggest to use the more flexible spline method. He et al. (2010) conduct a small simulation study limited to three spline based approaches, using splines of degree one, i.e. piecewise constant effects. They compare their linear Bayesian estimation approach to an MCMC approach (Hennerfeind et al., 2006) and a penalised partial likelihood approach (Gray, 1994) and conclude that both Bayesian approaches perform well while the penalised partial likelihood approach tends to over-smooth effects (using the default smoothing parameter). The MCMC approach is declared to be the best of the investigated methods, since it estimates all smoothing parameters. However, it is noted that the method may sometimes fail to converge when using splines of higher degrees.

Lehr and Schemper (2007) compare pre-defined functions, simple piecewise constant effects, penalised piecewise constant effects (Verweij and van Houwelingen, 1995), natural cubic splines (Hess, 1994; Abrahamowicz et al., 1996; Heinzl and Kaider, 1997) and FPs. The test procedure of the FP approach is identical to that proposed by Sauerbrei et al. (2007), apart from the default time transformation which is chosen to be t (as proposed by Sauerbrei and Royston, 1999, for non-linear functional forms) instead of $\log(t)$ (see Section 2.5.1 for details on the influence of the default transformation). They conduct a small simulation study with sample sizes up to 300 to investigate overfitting due to modelling time-varying effects. The results suggest that with respect to overfit FPs and penalised likelihood approaches are the techniques of choice, as they maintain the power of tests on the PH

assumption and simultaneously permit flexible modelling of time-varying effects.
However, inference about time-varying effects requires sufficiently large sample sizes and investigation of such small data sets would be inadvisable in practical applications, as it provides only limited insight in the presence and shape of time-varying effects with considerable uncertainty about conclusions.

So far, to our knowledge, no comprehensive comparisons of several recent approaches based on different techniques for modelling time-varying effects have been conducted. To assess the performance and practical applicability of different methods, we chose five selective approaches representative for different techniques, which are compared in a real life data example and a simulated data set with a special emphasis on the selected (time-varying) effects. Besides the Fractional Polynomial Time (FPT) approach proposed by Sauerbrei et al. (2007), on which the main focus of this thesis lies, an alternative FP approach (Berger et al., 2003), Reduced Rank models with splines (Perperoglou et al., 2006b), a Bayes approach based on penalised splines (Kneib and Fahrmeir, 2007) and a non-parametric approach based on cumulative regression functions (Scheike and Martinussen, 2004) are considered.

In addition to this comparison, the properties of the FPT approach are investigated in a large simulation study considering ten different (time-varying) effects with respect to type I and II error, the quality of estimated effects and the prediction performance.

2.5 Recent (multivariable) modelling strategies for time-varying effects

2.5.1 Fractional Polynomial Time (FPT) model

Fractional polynomial (FP) based approaches for modelling time-varying effects can be viewed as an extension of Cox's proposal of using a pre-defined function $f(t)$ by selecting the best $f(t)$ from a specified class of functions.
The Fractional Polynomial Time (FPT) procedure has been proposed by Sauerbrei et al. (2007) and is intended as a transfer of the MFP approach (Sauerbrei and Royston, 1999) for modelling the functional form of continuous covariates (i.e. non-linear effects) to time-varying effects. It corresponds to Step 3 of the MFPT approach introduced in Section 2.2. Hence, the FPT approach is based on a model of type

$$\lambda(t|X) = \lambda_0(t) \exp\left(\sum_{i=1}^{q} f_i(X_i)\beta_i(t)\right)$$

with potentially non-linear functional forms of covariates $f_i(X_i)$ and time-varying effects $\beta_i(t)$. Although FPT is a parametric approach based on global functions, the class of FPs is assumed to provide members that are capable of modelling situations with medium-term as well as long-term follow-up. That means, the functional form of the effect should remain the same when more information (longer follow-up) becomes available.

Selection and estimation of time-varying effects

Time-varying effects are modelled as covariate by time function interactions based on FPs of maximum degree 2. The class of FPs is defined by the set of powers

$$\mathcal{S} = \{\text{-2, -1, -0.5, 0, 0.5, 1, 2, 3}\}, \tag{2.6}$$

where t^0 is defined as $\log(t)$. In this case, an FP of degree 1 (FP1) is defined as

$$\beta_i(t) = \gamma_{i0} + \gamma_{i1} t^p, \quad p \in \mathcal{S},$$

and an FP of degree 2 (FP2) as

$$\beta_i(t) = \begin{cases} \gamma_{i0} + \gamma_{i1} t^{p_1} + \gamma_{i2} t^{p_2}, & p_1 \neq p_2; p_1, p_2 \in \mathcal{S} \\ \gamma_{i0} + \gamma_{i1} t^{p} + \gamma_{i2} t^{p} \log(t), & p_1 = p_2 = p \in \mathcal{S} \end{cases}$$

For the selection of a time-varying effect of a single covariate Sauerbrei et al. (2007) proposed the FPT algorithm:

- Calculate all possible FPs (8 FP1 and 36 FP2 functions for $\mathcal{S}$ as defined in (2.6))
- Determine the best FP1 and FP2 in terms of the deviance of the model
- Apply likelihood ratio tests to determine the best-fitting effect. The degrees of freedom (df) of the χ^2 test statistics are determined by the difference in complexity of fitted FPs. The hierarchical closed test procedure compares the deviance differences between
  (1) the best FP2 and a constant effect (4 df)
  (2) the best FP2 and a default function, i.e. an FP1 based on $\log(t)$ (3 df)
  (3) the best FP2 and the best FP1 (2 df)

This hierarchical closed test procedure aims to find a model which is as complex as necessary, but as parsimonious as possible. It successively checks, whether (1) a time-varying effect is needed at all, (2) the simple default time transformation $\log(t)$ already adequately describes the time-varying pattern or (3) the best FP1 function is sufficient. If a test on any of the three levels is not significant, the test procedure stops and chooses the more par-

simonious effect of the respective test. If all three tests are significant, the most complex modelling alternative, i.e. the best FP2, is used.

The choice of $\log(t)$ as a default time transformation is motivated by its practical plausibility. It allows the modelling of short-term effects and is a popular choice when using parametric functions of time. A logarithmic decrease (increase) of $\beta(t)$ relates to a uniform decrease (increase) in the hazard ratio over time. The default time transformation is selected unless the data gives strong evidence for a different shape.

Multivariable model building and properties

In a multivariable model, a forward selection procedure is applied. First, the best FP2 is selected and tested against a time-constant effect for each covariate in turn. Denote the p value of the most significant FP2 over all covariates as p_{min}. If $p_{min} \leq \alpha$ for a nominal significance level α, the final FP function for the corresponding covariate is determined using the FPT algorithm. This is repeated until $p_{min} > \alpha$, i.e. no possible time-varying effect is significant any more.

Adding a time-varying effect based on FPs to a model can also be regarded as adding a time-dependent covariate with constant effect, e.g. if the time-varying effect $\beta_i(t)$ is an FP1

$$X_i\beta_i(t) = X_i\gamma_{i0} + X_i t^p \gamma_{i1} = X_i\gamma_{i0} + \tilde{X}_i(t)\gamma_{i1}.$$

Consequently, the FPT model can be fitted using standard Cox regression tools for time-dependent covariates based on maximum partial likelihood methodology.

Properties of this algorithm are unknown and remain to be investigated. Evaluation of its performance is a major aim of this thesis.

2.5.2 Dynamic Cox model

Another approach based on FPs has been proposed by Berger et al. (2003). The main idea of this method is the same as for the FPT approach, with a hazard of the form

$$\lambda(t|X) = \lambda_0(t) exp\left(\sum_{i=1}^{q} X_i\beta_i(t)\right)$$

with $\beta_i(t) = \gamma_{i0} + \sum_{j=1}^{M} \gamma_{ij} t^{(p_j)}$ being an FP of maximum degree $M = 2$. γ_{ij} are the regression coefficients and $p_1 \leq \ldots \leq p_M$ the fractional polynomial exponents ($p_j \in \mathcal{S} = \{$-2, -1, -0.5, 0, 0.5, 1, 2, 3$\}$).

Selection of time-varying effects

The selection procedure, though, differs from that of the FPT approach. While the basic test procedure is also based on likelihood ratio tests, it is not hierarchical as in the FPT approach and thus does not distinguish between FPs of different degrees. Furthermore, a simultaneous test on the significance of the covariate ($H_0 : \beta(t) = 0$) may be performed. The best FP for a time-varying effect is selected in terms of the minimum p value of the likelihood ratio statistic with $2M$ degrees of freedom.

Multivariable model building and properties

A multivariable model (with q covariates) is derived in a backfitting-type procedure:

1. Select the best FP for the time-varying effect of covariate X_1 with time-constant effects for all other covariates. If the time-varying effect is significant (according to the likelihood ratio test), keep the FP powers fixed.
2. Repeat this procedure for all other covariates X_i, $i = 2, \ldots, q$: Select the best FP for X_i while the FP powers for covariates X_j, $j = 1, \ldots, i-1$, are kept fixed and assuming time-constant effects for all X_k, $k = i+1, \ldots, q$.
3. Update the FPs for each X_i, $i = 1, \ldots, q$, in turn, fixing the FP powers of all other covariates.
4. Repeat (3.) until the FP powers do not change any more.

This reveals another major difference to the FPT algorithm. While in the FPT algorithm the FP powers for time-varying effects remain fix once they have been estimated, the above algorithm enables updating of FPs. The order of the covariates should be irrelevant for independent covariates. In case of dependent covariates, or to assure reproducibility, the order of covariates may be fixed with respect to the p values of a full PH model.

Berger et al. (2003) investigate the test on time-varying effects in a simulation study. The results are promising, with high power for detecting time variation in the investigated settings. Comparison to standard tests shows a superiority of the FP test procedure. However, the functional form of time-varying effects is not investigated and may be completely wrong.
As Berger et al. (2003) denote this model by “Dynamic Cox model”, we will in the sequel use the same term.

2.5.3 Empirical Bayes model

An empirical Bayes approach based on structured additive regression is proposed by Kneib and Fahrmeir (2007). This approach simultaneously estimates the regression and variance

parameters using iteratively weighted least squares (IWLS) and restricted maximum likelihood (REML), respectively. Time-varying effects are modelled through cubic Bayesian penalised B-splines (P-splines) with second order random walk penalty.

The posterior mode estimates can in a frequentist setting be interpreted as penalised likelihood estimates. The penalisation of regression coefficients in the frequentist framework can from a Bayesian viewpoint be seen as specification of a prior for these coefficients. Furthermore, the variance parameters in the Bayesian approach are equivalent to the inverse smoothing parameters in a frequentist setting.

Model specification and priors

The empirical Bayes approach estimates the extended Cox model

$$\lambda_i(t|X) = \exp(\eta_i(t)), \qquad i = 1, \ldots, n,$$

where $\eta_i(t)$ is a structured additive predictor which partitions the covariates with respect to time-constant and time-varying effects and n is the number of observations. Thus,

$$\eta_i(t) = \beta_0(t) + \sum_{j=1}^{q^{tv}} x_{ij}^{tv} \beta_j(t) + x_i^{const^T} \alpha, \tag{2.7}$$

where $\beta_0(t) = \log(\lambda_0(t))$ is the log baseline hazard, $\beta_j(t)$ are the time-varying effects of covariates x_j^{tv} and α contains the time-constant effects of the covariates in x^{const}.

To derive a matrix notation of (2.7), the predictor vector is defined as $\eta = (\eta_1, \ldots, \eta_n)^T$ where $\eta_i = \eta_i(t_i)$ is the value of predictor (2.7) at the observed survival time $t_i, i = 1, \ldots, n$.

Similarly, $\beta_j = (\beta_j(t_1), \ldots, \beta_j(t_n))^T$ and $\beta_j^* = diag(x_{1j}^{tv}, \ldots, x_{nj}^{tv})\beta_j$ are the vectors of evaluations of $\beta_j(t)$ and $\beta_j(t)x_j^{tv}$, respectively. The latter can be expressed as the matrix product of an appropriately defined design matrix Z_j and a vector γ_j of regression coefficients

$$\beta_j^* = Z_j \gamma_j.$$

Hence, the predictor vector is equal to

$$\eta = Z_1\gamma_1 + \ldots + Z_{q^{tv}}\gamma_{q^{tv}} + X^{const}\alpha. \tag{2.8}$$

The fixed effects parameters α are assumed to follow a non-informative prior $p(\alpha) \propto const$. For random effects $\gamma_1, \ldots, \gamma_L$ Gaussian priors

$$p(\gamma_j|\tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\gamma_j^T P_j \gamma_j\right) \tag{2.9}$$

are assumed, where P_j is a penalty matrix that penalises too abrupt jumps between neighbouring parameters and τ_j^2 corresponds to the inverse smoothing parameter.

Modelling of time-varying effects

The unknown smooth functions β_j are modelled through P-splines, i.e. penalised B-splines. Hence, β_j is estimated by a polynomial spline of degree $p = 3$ defined on a set of $M + 1$ knots $t_{min} = \kappa_0 < \kappa_1 < \ldots < \kappa_{M-1} < \kappa_M = t_{max}$. The spline can then be written in terms of a linear combination of $p + M$ B-spline basis functions B_m, i.e.

$$\beta_j(t) = \sum_{m=1}^{p+M} \gamma_{jm} B_m(t).$$

Thus, the design matrices Z_j in (2.8) would be defined by $Z_j[i, m] = x_{ij}^{tv} B_m(t)$.

The number of knots is an essential choice. Here, the proposal of Eilers and Marx (1996) for non-linear functional forms of covariates is adopted to time-varying effects. They use a moderately large number (20 to 40) of equidistant knots, to obtain sufficient flexibility and impose a difference penalty to ensure smoothness of underlying functions. The Empirical Bayes approach uses the stochastic analogue of a second order difference penalty, a second order random walk (Kneib, 2006, pp. 32-39).
For simplicity, assume that κ_m are equidistant knots. For each κ_m, one parameter γ_{jm} is estimated using random walk priors. Second order random walks are

$$\gamma_{jm} = 2\gamma_{j,m-1} - \gamma_{j,m-2} + u_{jm}, \qquad m = 3, \ldots, p + M \tag{2.10}$$

with Gaussian errors $u_{jm} \sim N(0, \tau_j^2)$ and diffuse priors $p(\gamma_{j1})$ and $p(\gamma_{j2}) \propto const$. This second order random walk acts as a smoothness prior penalising deviations from the linear trend $2\gamma_{j,m-1} - \gamma_{j,m-2}$. The precision matrix of the joint distribution of γ_j is then of the form $P_j = D^T D$, where D is a second order difference matrix

$$D = \begin{pmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{pmatrix}.$$

In case of non-equally spaced survival times, the random walks must be modified to account for non-equal distances (for details see Kneib, 2006, p. 37).

Hence, the prior (2.9) corresponds to the difference penalty $\nu_j \gamma_j^T P_j \gamma_j$ with smoothing parameters $\nu_j = 1/(2\tau_j^2)$ in a penalised log-likelihood setting.

Reparametrisation and estimation

In most cases, the precision matrices P_j will be rank-deficient. Therefore, the random effect priors are partially improper. As standard mixed models require proper random effects priors, all vectors of regression coefficients γ_j are reparametrised into an unpenalised part (fixed effects) and a penalised part (random effects)

$$\gamma_j = Z_j^{unp}\gamma_j^{unp} + Z_j^{pen}\gamma_j^{pen}. \tag{2.11}$$

That is, γ_j^{unp} represents the part of γ_j that is not penalised by P_j and γ_j^{pen} represents the deviations of the parameters γ_j from the nullspace of P_j. From the general prior (2.9) for γ_j, it follows that

$$p(\gamma_j^{unp}) \propto const$$

and

$$\gamma_j^{pen} \sim N(0, \tau_j^2 I_{k_j}),$$

where k_j is the rank of the precision matrix P_j.

By defining the matrices $\tilde{U}_j = Z_j Z_j^{unp}$ and $\tilde{Z}_j = Z_j Z_j^{pen}$, the predictor (2.8) can be rewritten as

$$\eta = \sum_{j=1}^{q^{tv}} Z_j\gamma_j + X^{const}\alpha = \sum_{j=1}^{q^{tv}} (\tilde{U}_j\gamma_j^{unp} + \tilde{Z}_j\gamma_j^{pen}) + X^{const}\alpha = \tilde{U}\gamma^{unp} + \tilde{Z}\gamma^{pen},$$

with $\tilde{Z} = (\tilde{Z}_1\tilde{Z}_2\cdots\tilde{Z}_{q^{tv}})$, $\tilde{U} = (\tilde{U}_1\tilde{U}_2\cdots\tilde{U}_{q^{tv}}X^{const})$, $\gamma^{pen} = ((\gamma_1^{pen})^T, \ldots, (\gamma_{q^{tv}}^{pen})^T)^T$ and $\gamma^{unp} = ((\gamma_1^{unp})^T, \ldots, (\gamma_{q^{tv}}^{unp})^T, \alpha^T)^T$. This is a generalised linear mixed model (GLMM) with fixed effects γ^{unp} and random effects $\gamma^{pen} \sim N(0, \Sigma)$ where $\Sigma = blockdiag(\tau_1^2 I_{k_1}, \ldots, \tau_{q^{tv}}^2 I_{k_{q^{tv}}})$.

Thus, GLMM methodology for simultaneous estimation of the time-varying effects $\beta_j(t)$ and the variance parameters τ_j^2 can be applied. The estimation procedure iteratively updates

(i) the regression coefficients $\hat{\gamma}^{unp}$ and $\hat{\gamma}^{pen}$ given the current variance parameters using iteratively weighted least squares via a Newton-Raphson step and

(ii) the variance parameters given the current regression coefficients using restricted maximum likelihood via a Fisher scoring step

until convergence.

Multivariable model building and properties

The selection of time-varying effects is carried out according to the proposal of Hofner et al. (2010). For sole selection of time-varying effects (without variable selection and selection of non-linear effects), their proposal corresponds to a forward selection algorithm based on

the conditional AIC (AIC_c), which is composed of the conditional likelihood and the effective degrees of freedom as a complexity measure (Hofner et al., 2010). In the first iteration, all possible models with one time-varying effect are fitted. The best model in terms of the AIC_c is compared to the PH model. If it is better, the time-varying effect is kept and one further time-varying effect is added for each of the remaining covariates in turn. The best of these models is then compared to the best model of the previous iteration. The procedure stops, if no further improvement in terms of AIC_c is achieved. Hofner et al. (2010) also investigate the performance of the proposed model building strategy in combination with further modelling alternatives.

2.5.4 Semiparametric Extended Cox model

The Semiparametric Extended Cox model (Martinussen et al., 2002; Scheike and Martinussen, 2004; Martinussen and Scheike, 2006), as the authors name it, is based on cumulative parameter functions. Asymptotic properties of the predictors have been developed using the martingale structure of the model.

The intensity of a fully non-parametric model, where all covariate effects are allowed to vary with time, is

$$\lambda(t|X) = Y(t)\lambda_0(t)\exp\left(\sum_{i=1}^{q} X_i(t)\beta_i(t)\right), \tag{2.12}$$

where $Y(t)$ is the at risk process. Estimation and tests are based on the cumulative regression functions

$$B_i(t) = \int_0^t \beta_i(s)ds,$$

as they converge at a faster rate than $\beta_i(t)$ and lead to a uniform asymptotic description of the estimator which is necessary for hypothesis testing. Furthermore hypothesis testing about $\beta_i(t)$ can also be formulated in terms of $B_i(t)$.

Selection and modelling of time-varying effects

The test on a time-varying effect for covariate X_i is based on the hypothesis $H_0 : \beta_i(t) = \gamma_i$, or equivalently $H_0 : B_i(t) = \gamma_i t$. Test statistics for this hypothesis are based on the test process

$$\sqrt{n}\left(\hat{B}_i(t) - \hat{\gamma}_i t\right),$$

where $\hat{B}_i(t)$ is an estimator for $B_i(t)$ and $\hat{\gamma}_i$ is computed under the null hypothesis. Under the null, this process converges to a mean-zero Gaussian process. However, its limiting distribution is complicated and the distribution of the test statistics need to be simulated.

Scheike and Martinussen (2004) propose two test statistics, a Kolmogorov-Smirnov type test

$$T_S = \sqrt{n} \sup_{t \in [0,\tau]} |\hat{B}_i(t) - \hat{B}_i(\tau)\frac{t}{\tau}|$$

and a Cramér-von Mises type test

$$T_I = n \int_0^\tau (\hat{B}_i(t) - \hat{B}_i(\tau)\frac{t}{\tau})^2 dt.$$

Both test statistics are based on the idea that $\hat{B}_i(\tau)\frac{t}{\tau}$ is an estimate of the underlying constant effect under the null, i.e. $\hat{B}_i(\tau)\frac{t}{\tau} = \hat{\gamma}_i \tau \frac{t}{\tau} = \hat{\gamma}_i t$. A drawback of these tests is their dependence on the choice of the interval $[0, \tau]$ which defines the observation period of interest for the test. Furthermore, the supremum tends to be large at places with large variation. Yet, looking at places with small variation is sometimes more interesting. Modified versions of the above test statistics have been proposed which take the variance into account. These test statistics, though, may show erratic behaviour at the start and end of the time interval, because the test statistic, which is nearly zero, is divided by a standard error that is almost zero. In such cases, a third version ignoring the first and last two jumps can be used.

For calculation of p values a large number of resampling processes (e.g. 1000) are generated under the null. Then the test statistics for these resampled processes $\hat{\Delta}_i(t) - \hat{\Delta}_i(\tau)\frac{t}{\tau}$ and the test process $\hat{B}_i(t) - \hat{B}_i(\tau)\frac{t}{\tau}$ are calculated as

$$T_{test} = \sqrt{n} \sup_{t \in [0,\tau]} |\hat{B}_i(t) - \hat{B}_i(\tau)\frac{t}{\tau}|$$

and

$$T_{resampled} = \sqrt{n} \sup_{t \in [0,\tau]} |\hat{\Delta}_i(t) - \hat{\Delta}_i(\tau)\frac{t}{\tau}|$$

with the p value being equal to the probability

$$P(T_{resampled} \geq T_{test}).$$

Multivariable model building

In multivariable analyses, Scheike and Martinussen (2004) recommend (for testing purposes and not too many covariates) to start with the non-parametric model (2.12) and then simplify it in a backward elimination manner to a semiparametric model, where only some covariate effects vary with time while others are assumed to be constant:

$$\lambda(t|X) = Y(t)\lambda_0(t) \exp\left(\sum_{i=1}^{q^{tv}} X_i^{tv}(t)\beta_i(t) + \sum_{j=1}^{q^{const}} X_j^{const}(t)\gamma_j\right).$$

Hence, starting with the non-parametric model, where all effects are allowed to vary in time, p values for $H_0 : B_i(t) = \gamma_i t$ are calculated for each of the time-varying effects. If the largest p value p_{max} is smaller than a nominal significance level α, the current model is accepted. If $p_{max} > \alpha$, the corresponding effect is assumed to be constant and the p values are calculated for the model with one time-varying effect less. This procedure stops, when $p_{max} \leq \alpha$ or when all effects have been set to constant.

Estimation procedure

The likelihood based estimation procedure for $\beta_i(t)$ is based on finding a solution to the score equation $X_i^{tv}(t)\,(dN(t) - \lambda(t|X)dt)$, the first derivative of the log-likelihood with respect to $\beta_i(t)$. This has no solution, as the first term represents a pure jump process while the second is absolutely continuous. To obtain a solution, the cumulative parameter functions $B_i(t)$ are estimated and smoothness of the underlying coefficients is introduced through the estimation of $\beta_i(t)$, for which a kernel estimator

$$\hat{\beta}_i(t) = \int b^{-1} K\left(\frac{s-t}{b}\right) d\hat{B}_i(s)$$

is used with positive bandwidth b and a uniformly continuous kernel K with support [-1,1], satisfying $\int K(s)ds = 1$ and $\int sK(s)ds = 0$. An iteration procedure for estimating the effects γ_j and $\beta_i(t)$ is constructed based on initial estimates and iterated until convergence according to the following scheme. For iteration r

1. Compute the Breslow estimator of the cumulative baseline hazard based on preliminary estimates $\hat{\beta}_i^r(t)$ and $\hat{\gamma}_j^r$ and smooth this estimate to obtain $\hat{\lambda}_0^r(t)$.
2. Estimate $\hat{\gamma}_j^{r+1}$ based on the score equation using a Newton-Raphson algorithm.
3. Use the estimates $\hat{\gamma}_j^{r+1}$ to calculate $\hat{B}_i^{r+1}(t)$ based on the score equations using a Newton-Raphson algorithm.
4. Smooth $\hat{B}_i^{r+1}(t)$ to obtain $\hat{\beta}_i^{r+1}(t)$ and return to 1.

Properties

This approach leads to efficient estimates of γ_j and $B_i(t)$ (Scheike and Martinussen, 2004). Furthermore, the random vector $\sqrt{n}\left(\hat{\gamma}_j - \gamma_j\right)$ is asymptotically normal with mean zero and $\sqrt{n}\left(\hat{B}_i(t) - B_i(t)\right)$ converges towards a mean zero Gaussian process, both with variances/ covariances that can be consistently estimated.

To investigate the finite sample properties of the test statistics, Scheike and Martinussen (2004) conduct a small simulation study based on two covariates, one time-varying effect

and several sample sizes and correlation levels. They show that the procedure performs quite well in terms of type I error and power.

The complete approach, though, was mainly developed for testing and leads to an algorithm that is comparatively easy to study theoretically, but was not intended for estimating the time-varying effects $\beta_i(t)$. When the interest lies in the shape of $\beta_i(t)$ rather than the mere test result, an additional transformation is required by the user.

Technical remarks

Within the estimation algorithm, a simple kernel smoother with global bandwidth is used to obtain $\hat{\beta}_i(t)$. This can be improved by a local approach. Hence, we use the local polynomial regression approach (Loader, 1999) implemented in the `R` package `locfit` (Loader, 2007) with quadratic polynomials, tricube weight function and a nearest neighbour fraction of 0.7 (the default setting) to smooth the final estimate. The first derivative is provided as a local slope estimate.

2.5.5 Reduced Rank model

In the Reduced Rank model (Perperoglou et al., 2006a,b), time-varying effects are modelled as covariate by time function interactions. The main idea of this approach is to reduce the number of parameters in order to obtain more stable and parsimonious models depending on the rank of the model.

The full rank model is identical to the Cox non-PH model

$$\lambda(t|X) = \lambda_0(t) \exp\left(X\Theta F^T(t)\right). \tag{2.13}$$

The row vector $F(t) = (f_1(t), \ldots, f_s(t))$ contains the (pre-specified) time functions. The structure matrix Θ is composed of the regression coefficients of all covariate by time function interactions. For q covariates and s time functions, Θ is of dimension $q \times s$.

Estimation procedure

Model (2.13) can be estimated by standard software for time-dependent covariates by considering $X_i f_j(t)$ as time-dependent covariates. This estimation, though, can be unstable for many covariates and/or time functions. To avoid overfitting and instability, a rank restriction is put on the structure matrix Θ.

The idea behind this is that Θ can be factorised as $\Theta = B\Gamma^T$ in different ways, with B being

a $q \times r$ matrix and Γ a $s \times r$ matrix. Thus the rank r model is

$$\begin{aligned}\lambda(t|X) &= \lambda_0(t)\exp(XB\Gamma^T F^T(t)) \\ &= \lambda_0(t)\exp\left(\sum_{k=1}^{r}(X\beta_k)(F(t)\gamma_k)\right)\end{aligned}$$

where β_k is the kth column of B and γ_k the kth column of Γ. Thus, the original set of parameters is now reduced to a set of r linear combinations of time-functions $F(t)\gamma_k$ and r linear combinations of covariates $X\beta_k$, $k = 1, \ldots, r$.

This model is then estimated in an iterative procedure using the partial log-likelihood

$$PL(\beta,\gamma) = \sum_{i=1}^{D}\sum_{k=1}^{r}(X_i\beta_k)(F_i\gamma_k) - \sum_{i=1}^{D}\ln\left\{\sum_{j\in R(t_i)}\exp\Big(\sum_{k=1}^{r}(X_j\beta_k)(F_i\gamma_k)\Big)\right\},$$

where X_i is the covariate vector of the individual with an event at time t_i, F_i the row vector of time functions at time t_i and $R(t_i)$ the risk set at t_i.

The estimation procedure uses a Newton-Raphson algorithm and alternates between estimation of the β's and the γ's. It starts with the estimation of the γ's with given initial values for the r β vectors. For example, the estimates $\tilde{\beta}$ from a simple Cox model can be used as initial values for β_1. If $r > 1$, a random perturbation of the $\tilde{\beta}$'s can be used as initial values for $\beta_2, \ldots, \beta_r$.
The estimated effect of covariate X_i is then given by

$$\hat{\beta}_i(t) = \sum_{k=1}^{r}\hat{\beta}_k\left(\sum_{l=1}^{s}\hat{\gamma}_{lk}f_l(t)\right).$$

Model parameters and properties

The approach strongly depends on the choice of the rank r. To determine the optimal rank, Perperoglou et al. (2006b) propose a forward-type algorithm. This starts by fitting a rank 1 model and then increases the rank up to the maximum rank $r = min(q, s)$. The optimal rank is chosen based on the AIC.

For the time functions $F(t) = (f_1(t), \ldots, f_s(t))$ many choices are possible. Perperoglou et al. (2006b) propose to use B-splines with interior knots placed at convenient positions, ensuring a sufficient number of events in each interval. Any choice of $F(t)$ should meet the conditions $f_1(0) = 1$ and $f_k(0) = 0$ for $k = 2, \ldots, s$. Furthermore, $F(t)$ should include the constant to assure that the Reduced Rank model also contains the basic PH model, e.g. $f_1(t) = 1$.

The approach does not include selection of time-varying effects. An estimated effect for

covariate X_i can be time-constant only in the full rank model if $f_1(t)$ is set to $f_1(t) = 1$ and $\hat{\gamma}_{j2} = \ldots = \hat{\gamma}_{js} = 0$.

Perperoglou et al. (2006a) investigate the speed of their algorithm in a small simulation study with time-constant effects, but, to our knowledge, do not provide simulation studies on time-varying effects.

2.6 Predictions

To evaluate the prediction performance of approaches, predicted survival probabilities are required. However, information on evaluation of predictions for models allowing for time-varying effects is rare. Hence, this section presents theoretical considerations on the calculation of predicted survival probabilities for the approaches introduced in Sections 2.5.1-2.5.5.

2.6.1 FPT and Dynamic Cox model

To obtain predicted survival probabilities for both FP models, the baseline hazard $\lambda_0(t)$ has to be estimated. As Therneau and Grambsch (2000, chap. 10.2.4) and Kalbfleisch and Prentice (2002, chap. 6.4.1) state, the Breslow estimate is also applicable to time-dependent covariates, which is equivalent to our problem, as mentioned in Section 2.5.1. However, for time-varying effects it is

$$S(t|X) = \exp\left\{-\int_0^t \lambda_0(s)\exp\left(X\beta(s)\right)ds\right\} \neq \exp\left\{-\int_0^t \lambda_0(s)ds\,\exp\left(X\beta(s)\right)\right\}.$$

The cumulative baseline hazard for the FP models is estimated by

$$\hat{H}_0(t) = \begin{cases} \hat{H}_0^\star(t_{(k)}) & \text{for } t_{(k)} \leq t < t_{(k+1)} \\ \hat{H}_0^\star(t_{(n_e)}) & \text{for } t \geq t_{(n_e)} \end{cases}$$

with $\hat{H}_0^\star(t_{(k)})$ being the Breslow estimate of the cumulative baseline hazard as defined in (2.3).

Since $\hat{H}_0(t)$ is a step-function with jumps at event times $t_{(k)}$, $k = 1,\ldots,n_e$ (with n_e the number of distinct event times), a derivative of $\hat{H}_0(t)$ is

$$\hat{\lambda}_0(t_{(k)}) = \frac{\hat{H}_0(t_{(k+1)}) - \hat{H}_0(t_{(k)})}{\Delta_k}, \qquad \text{with } \Delta_i = t_{(k+1)} - t_{(k)}.$$

Taking into account that $\hat{\lambda}_0(t)$ (and thus $\hat{\lambda}_0(t)\exp\left(\sum_{i=1}^{q} X_i\hat{\beta}_i(t)\right)$) is a step function, the estimate for (2.4) reduces to

$$\begin{aligned}
\hat{S}(t|X) &= \exp\left\{-\int_0^t \hat{\lambda}_0(s)\exp\left(\sum_{i=1}^{q} X_i\hat{\beta}_i(s)\right)ds\right\} \\
&= \exp\left\{-\sum_{t_{(k)}\le t} \hat{\lambda}_0(t_{(k)})\exp\left(\sum_{i=1}^{q} X_i\hat{\beta}_i(t_{(k)})\right)\Delta_k\right\} \\
&= \exp\left\{-\sum_{t_{(k)}\le t} \frac{\hat{H}_0(t_{(k+1)}) - \hat{H}_0(t_{(k)})}{\Delta_k}\exp\left(\sum_{i=1}^{q} X_i\hat{\beta}_i(t_{(k)})\right)\Delta_k\right\} \\
&= \exp\left\{-\sum_{t_{(k)}\le t} \left(\hat{H}_0(t_{(k+1)}) - \hat{H}_0(t_{(k)})\right)\exp\left(\sum_{i=1}^{q} X_i\hat{\beta}(t_{(k)})\right)\right\}
\end{aligned}$$

The survival probabilities for new time points $t^\star$ can then be calculated by

$$\tilde{S}(t^\star|X) = \begin{cases} 1 & \text{for } t^\star < t_{(1)} \\ \hat{S}(t_{(k)}|X) & \text{for } t_{(k)} \le t^\star < t_{(k+1)} \end{cases}$$

2.6.2 Empirical Bayes model

For the Empirical Bayes model, information on baseline hazard and time-varying effects is available only at a finite number of time points. Hence, the predictor

$$\eta(t) = \beta_0(t) + \sum_{i=1}^{q^{tv}} X_j^{tv}\beta_j(t) + \sum_{j=1}^{q^{const}} X_j^{const}\alpha$$

reduces to a step function with $\beta_0(t)$ being the log baseline hazard. The predicted survival probabilities can be calculated using the numerical integral without loosing accuracy, giving:

$$\hat{S}(t|X) = \exp\left\{-\sum_{t_{(k)}\le t}\exp\left(\hat{\eta}(t)\right)\right\}.$$

The survival function is then defined as

$$\tilde{S}(t^\star|X) = \begin{cases} 1 & \text{for } t^\star < t_{(1)} \\ \hat{S}(t_{(k)}|X) & \text{for } t_{(k)} \le t^\star < t_{(k+1)} \end{cases}$$

Technical remarks: `BayesX` estimates the full likelihood and thus provides estimates for the log baseline hazard $\log(\lambda_0(t))$ and an additional constant offset α_0, i.e. $\hat{\beta}_0(t) = \log(\hat{\lambda}_0(t)) + \hat{\alpha}_0$. The baseline hazard and time-varying effects are estimated at all distinct survival times (event and censoring times).

2.6.3 Semiparametric Extended Cox model

The survival function for the Semiparametric Extended Cox model is given as

$$S(t|X) = \exp\left\{-\int_0^t \lambda_0(t)\exp\left(\sum_{i=1}^{q^{tv}} X_i^{tv}(t)\beta_i(t) + \sum_{j=1}^{q^{const}} X_j^{const}(t)\gamma_j\right) ds\right\} \tag{2.14}$$

along the lines of Martinussen and Scheike (2006, p. 226).

As the cumulative regression functions are available only for a finite number of time points k, $k = 1, \ldots, n_e$, the integral in (2.14) can be substituted by a numerical integral without loosing accuracy, giving the estimator:

$$\hat{S}(t|X) = \exp\left\{-\sum_{t_{(k)}\le t} \hat{\lambda}_0(t_{(k)})\exp\left(\sum_{i=1}^{q^{tv}} X_i^{tv}(t)\hat{\beta}_i(t_{(k)}) + \sum_{j=1}^{q^{const}} X_j^{const}(t)\hat{\gamma}_j\right)\left(t_{(k+1)} - t_{(k)}\right)\right\}.$$

The survival probabilities for new time points $t^\star$ are then calculated as

$$\tilde{S}(t^\star|X) = \begin{cases} 1 & \text{for } t^\star < t_{(1)} \\ \hat{S}(t_{(k)}|X) & \text{for } t_{(k)} \le t^\star < t_{(k+1)} \end{cases}$$

Technical remarks: The `timecox` function only fits the reparametrised model

$$\lambda(t|X) = \exp\left(\beta_0(t) + \sum_{i=1}^{q^{tv}} X_i^{tv}(t)\beta_i(t_{(k)}) + \sum_{j=1}^{q^{const}} X_j^{const}(t)\gamma_j\right), \tag{2.15}$$

providing cumulative regression functions for the log baseline hazard $\beta_0(t) = \log(\lambda_0(t))$ and the time-varying effects $\beta_i(t)$, estimated at all distinct event times.

The estimated cumulative effects must be smoothed to obtain the estimates for $\beta_0(t)$ and $\beta_i(t)$. For this smoothing step, though, numerous different possibilities exist. Here a local polynomial regression will be used for this purpose.

2.6.4 Reduced rank model

For the Reduced Rank model, the Breslow estimator of the cumulative baseline hazard can be calculated as

$$\hat{H}_0(t) = \sum_{t_{(k)} \leq t} \frac{d_k}{\sum_{j \in R(t_{(k)})} \exp\left(X_j \hat{\Theta} F^T(t_{(k)})\right)}$$

as outlined in Perperoglou et al. (2006a). Hence, an estimate for the baseline hazard is given by

$$\hat{\lambda}_0(t_{(k)}) = \frac{1}{\sum_{j \in R(t_{(k)})} \exp\left(X_j \hat{\Theta} F^T(t_{(k)})\right)}$$

The survival probabilities can be estimated by a numerical integral with the intervals being defined by the event times $t_{(k)}$, $k = 1, \ldots, n_e$:

$$\hat{S}(t|X) = \exp\left\{ - \sum_{t_{(k)} \leq t} \lambda_0(t_{(k)}) \exp\left(X_j \hat{\Theta} F^T(t_{(k)})\right) \right\}.$$

The predicted survival probabilities for new time points $t^\star$ are obtained by

$$\tilde{S}(t^\star|X) = \begin{cases} 1 & \text{for } t^\star < t_{(1)} \\ \hat{S}(t_{(k)}|X) & \text{for } t_{(k)} \leq t^\star < t_{(k+1)} \end{cases}$$

Technical remarks: The `coxvc` package offers the function `calc.h0` to calculate the estimated cumulative baseline hazard $\hat{H}_0(t)$.

Chapter 3

Assessment of time-varying effects

When assessing Cox models with time-varying effects, our interest lies in (i) the performance of the complete (multivariable) model and (ii) the fit of a selected time-varying effect, i.e. whether it adequately reflects the true effect or not. This chapter shortly presents two methods addressing these issues.

3.1 Prediction error curves

Prediction error curves (Gerds and Schumacher, 2007) are used to assess and compare prediction rules. Because resampling methods are used, the prediction rules can be assessed in the same data they are developed. Due to its time-dependency, the prediction error is also applicable to models with time-varying effects.

Let $Q_n = \{X_1, \ldots, X_n\}$ be survival data of n individuals with $X_i = (\tilde{T}_i, \Delta_i, Z_i)$, where $\tilde{T}_i = \min(T_i, C_i)$ is composed of the event time T_i and censoring time C_i, $\Delta_i = I(T_i \leq C_i)$ is the event indicator and Z_i a q-dimensional covariate vector. Additionally, $Y_i(t) = I(T_i > t)$ is the event status, $r_n(t|Z_i)$ the predicted survival probability for individual i at time t and $r_n = r(Q_n)$ the prediction rule trained on the data Q_n. The true prediction error of a prediction rule r at time t is defined as the expectation of the process of squared residuals

$$Err(t; r, Q_n) = E\{Y(t) - r_n(t|Z)\}^2.$$

It measures how well r predicts the individual event status, given Q_n (where Y and Z are replicates that are not in the sample). In practical applications, usually only time points before a certain time τ can be used due to censoring, where τ is chosen such that

$$G(\tau, z) = P(\Delta_i = 1|\tilde{T}_i = \tau, Z_i = z) > 0,$$

e.g. τ could be a time just before the maximum on study time.

The apparent error rate is a measure of the cumulative prediction error over $[0, \tau]$:

$$\overline{err}(t; r_n, \hat{G}_n) = \frac{1}{n} \sum_{i=1}^{n} \{Y_i(t) - r_n(t|Z_i)\}^2 W(t, \hat{G}_n, X_i)$$

with weights $W(t, \hat{G}_n, X_i) = \frac{I(\tilde{T} \leq t)\Delta_i}{\hat{G}_n(\tilde{T}_i - |Z_i)} + \frac{I(\tilde{T}_i > t)}{\hat{G}_n(t|Z_i)}$. Because the apparent error evaluates the prediction rule on the training set, it may result in a seriously negative biased estimate for the true prediction error.

Exactly the opposite direction of bias can be observed for the bootstrap based estimate of prediction error. When bootstrap samples $Q_1^\star, \ldots, Q_B^\star$ of size n are drawn with replacement from the data Q_n, the bootstrap cross-validation estimate is obtained as

$$\widehat{Err}_{B0}(t; r) = B^{-1} \sum_{b=1}^{B} n^{-1} \sum_{i: X_i \in Q_b^\star} \{Y_i(t) - r_b^\star(t|Z_i)\}^2 W(t, \hat{G}_n, X_i)$$

where $Q_b^0 = \{X_i : X_i \notin Q_b^\star\}$ is the out-of-bag sample and $r_b^\star$ is trained on $Q_b^\star$.

To overcome the bias problems, Efron (1983) proposes to use a linear combination of the downward biased apparent error and the upward biased bootstrap cross-validation estimator

$$\widehat{Err}_\omega(t) = \{1 - \omega(t)\}\overline{err}(t; r_n, \hat{G}_n) + \omega(t)\widehat{Err}_{B0}(t; r)$$

with weight $\omega(t) = 0.632 = P(X_i \in Q_b^\star)$. This choice reduces the bias (Efron, 1983) and is motivated by the fact that bootstrap samples are supported on approximately $0.632n$ of the original data points. This estimator can be further improved by choosing ω depending on the prediction rule. For this purpose, Efron and Tibshirani (1997) introduce the no-information error rate

$$NoInf(t; r_n) = n^{-2} \sum_{j=1}^{n} \sum_{i=1}^{n} \{Y_i(t) - r_n(t|Z_j)\}^2 W(t, \hat{G}_n, X_i),$$

which assesses the performance of r in a situation where the survival status is independent of the covariates, i.e. where the Z_i are reallocated systematically on Y_j for all $j = 1, \ldots, n$. This no-information error rate also contributes to the weights ω.

A specific adaptation of the estimator for survival data is proposed by Gerds and Schumacher (2007) with

$$\omega^\star(t) = \begin{cases} 1 & \text{if } NoInf(t; r_n) \leq \widehat{Err}_{B0}(t; r_n) \\ 0.632 & \text{if } NoInf(t; r_n) \leq \overline{err}(t; r_n) \text{ or } \widehat{Err}_{B0}(t; r) \leq \overline{err}(t; r_n) \\ \frac{0.632}{1 - 0.368\hat{R}(t)} & \text{otherwise} \end{cases}$$

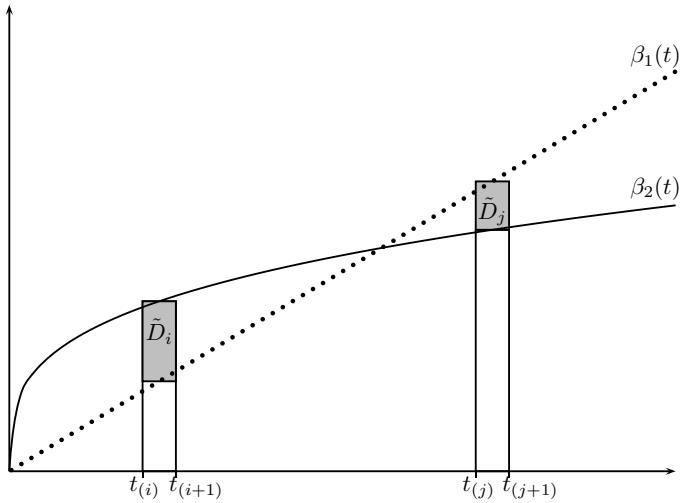


Figure 3.1: Scheme for calculating the area between curves of time-varying effects (ABCtime).

---

and the relative overfitting rate

$$\hat{R}(t) = \frac{\widehat{Err}_{B0}(t;r) - \overline{err}(t;r_n,\hat{G}_n)}{NoInf(t;r_n) - \overline{err}(t;r,\hat{G}_n)}.$$

Hence, the improved bootstrap 0.632+ estimator is given by

$$\widehat{Err}_{\omega^\star}(t) = \{1 - \omega^\star(t)\}\overline{err}(t;r_n,\hat{G}_n) + \omega^\star(t)\widehat{Err}_{B0}(t;r).$$

As a summary measure of prediction error, the integrated prediction error (IPEC) can be calculated using the Riemann integral implemented in the `R` package `peperr` (Porzelius and Binder, 2009). The IPEC is presented as the difference to the Kaplan-Meier estimate (dIPEC) following Porzelius et al. (2010).

3.2 Area between curves of time-varying effects (ABCtime)

To measure the distance between time-varying effects, we adapt an approach proposed by Govindarajulu et al. (2007) to calculate the area between smoothed curves of exposure. The area between curves for time-varying effects (ABCtime) should be applicable to various types of time-varying effects. This aim requires an adaptation of the original approach.

The ABCtime is based on weighted numeric integration. The area under a curve is calculated using 500 successive, non-overlapping rectangles with equal width. The area between two curves $\beta_1(t)$ and $\beta_2(t)$ is determined as the difference $\tilde{D}_s$ of pairs of rectangles as

sketched in Figure 3.1. Only the absolute difference between the two curves is considered, ignoring the sign. To determine the height of rectangles, Govindarajulu et al. (2007) use the right endpoint of intervals. For ABCtime, neither the right endpoint of intervals nor the left endpoint seem very suitable, as both could introduce systematic bias if left- and right-continuous step functions are compared. Instead, we prefer the midpoint of intervals to determine the function value, in order to achieve adequate applicability for all types of time-varying effects.

To account for the varying precision of estimates across the range of exposure, the area between curves is calculated as a weighted sum. Govindarajulu et al. (2007) use bootstrap based weights which can be very time-consuming. Therefore, we use the less computer-intensive logrank like weights which upweight time points where many patients are at risk and decrease for later time points, with less patients at risk:

$$w(t_{(s)}) = \frac{R(t_{(s)})}{\sum_{i=1}^{S} R(t_{(i)})}$$

with S being the number of intervals (here $S = 500$). Other choices of weights are possible. Besides equal weights, weights based on the inverse variance of the reference (or true) function or the inverse mean variance over competitive approaches could be used.

The ABCtime is then calculated as the weighted sum of rectangles $\tilde{D}_s$

$$\widehat{ABCtime} = \frac{\sum_{s=1}^{S} w(t_{(s)})\tilde{D}_s}{\sum_{s=1}^{S} w(t_{(s)})}.$$

Interpretation of the absolute value of ABCtime is not straight-forward. Therefore, we calculate the percentage of ABCtime on the weighted area under the reference function (pABCtime). The weights for this area under the reference function are equivalent to the weights $w(t_{(s)})$ for the area between the curves. The reference function may, for example, be the true time-varying effect in simulation studies or piecewise constant effects or smoothed Schoenfeld residuals in real-life applications. An pABCtime value of zero means that the effect under investigation is in perfect agreement with the reference.

As pABCtime has no upper bound, interpretation of a single value for one time-varying effect may provide limited information on the agreement. For comparison of several alternatives, though, pABCtime is a simple tool for identification of the best effect, i.e. the effect that is most similar to the reference function.

Chapter 4

Comparison of different approaches

In this chapter, the five approaches introduced in Section 2.5 are compared in a prognostic factor study, the Rotterdam breast cancer series. However, conclusions about selected effects are somewhat limited in real data sets, because the true shape is unknown. To obtain some more detailed insights, all approaches are additionally applied to a simulated data set.

4.1 The Rotterdam breast cancer series

4.1.1 The data

The Rotterdam breast cancer series includes data on patients treated at the Erasmus MC Daniel den Hoed Cancer Center for primary breast cancer between 1978 and 1993 (Foekens et al., 2000; Sauerbrei et al., 2007).
Data from 2982 patients are available for analysis, with the follow-up time ranging from 1 to 231 months and a median follow-up time of 107 months (estimated with the reverse Kaplan-Meier method). The final endpoint is event-free survival time (EFS) which is defined as time from primary surgery to the first occurrence of locoregional or distant recurrence, contralateral tumour, secondary tumour or death from breast cancer. Times to death from other causes are treated as censored, resulting in 1518 events for EFS.

The data set contains the covariates age, menopausal status, tumour size, tumour grade, number of positive lymph nodes, progesterone receptor, oestrogen receptor, hormonal therapy and chemotherapy (see Table 4.1). We stick to the proposal of Sauerbrei et al. (2007) to modify some of the variables. For tumour grade, they collapse grades 1 and 2 and use MICE (van Buuren et al., 1999) to replace missing values. The variable tumour size is split

Covariate	Code	Median (min, max) or percent
Age (years)	X_1	54 (24, 90)
Menopausal status	X_2	44% pre, 56% post
Tumour size	X_3	
$>$ 20mm	X_{3a}	47% no, 53% yes
$>$ 50mm	X_{3b}	90% no, 10% yes
Tumour grade	X_4	2% 1, 25% 2, 73% 3
Grade 1 and 2 collapsed	X_{4b}	27% 2, 73% 3
No. of positive lymph nodes	X_5	1 (0, 34)
transformed to $\exp(-0.12X_5)$	X_{5e}	
Progesterone receptor	X_6	41 (0, 5004)
Oestrogen receptor	X_7	61 (0, 3275)
Hormonal therapy	X_8	89% no, 11% yes
Chemotherapy	X_9	81% no, 19% yes

Table 4.1: Covariates in the Rotterdam breast cancer series and their distribution.

into two dummy variables for tumour size $>$20mm and tumour size $>$50mm. Furthermore, the preliminary transformation X_{5e} = $\exp$(-0.12X_5) is applied as proposed by Sauerbrei and Royston (1999). This modified version of the data is available on `http://www.imbi.uni-freiburg.de/biom/Royston-Sauerbrei-book/index.html#datasets`. Since our focus lies on investigation of approaches for time-varying effects and not on medical questions, we follow Sauerbrei et al. (2007) and consider the treatment variables in the same way as the prognostic factors.

The Rotterdam breast cancer series contains continuous and binary variables. For continuous covariates, estimating the baseline at value zero is sometimes not sensible. For variables like age, for example, estimation of baseline hazard and prediction at the mean covariate value is more suitable. Some prediction routines, as that for `coxph` in `R` even centre all variables around their mean by default. To enable better interpretability and comparability of all approaches, all variables of the Rotterdam breast cancer series are centred around their mean prior to analysing the data set.

4.1.2 Selection of a time-fixed model

Mismodelling the data by omitting important variables or erroneously assuming linearity for continuous covariates also influences the selection of time-varying effects. Most methods for assessing time-varying effects focus on this aspect only and take a time-fixed model as their starting point which may already include non-linear functional forms. We consider the first two steps of the MFPT approach as introduced in Section 2.2 as a sensible approach

	Step 1		Step 2	
Variable	$\hat{\beta}$	SE	$\hat{\beta}$	SE
X_1	-0.013	0.002	-0.013	0.002
X_{3a}	0.288	0.057	0.249	0.059
X_{3b}	-	-	0.171	0.080
X_{4b}	0.390	0.064	0.354	0.065
X_{5e}^2	-1.742	0.083	-1.710	0.085
$\log(X_6)$	-	-	-0.032	0.012
X_8	-0.387	0.085	-0.390	0.085
X_9	-0.456	0.073	-0.444	0.073

Table 4.2: Selection of covariates and non-linear covariate effects in the Rotterdam breast cancer series (variables adjusted by their mean) using the first two steps of the MFPT algorithm.

to derive such a time-fixed model. However, in general any prespecified model may be used as a starting point for investigating time-varying effects (Royston and Sauerbrei, 2008, chap. 11).

In the first step of the MFPT procedure, six covariates are selected (first column of Table 4.2), one of them with a non-linear functional form. The short-term analysis in step 2 adds two further covariates, again one of them with a non-linear transformation. Hence, the final model under the PH assumption consists of eight covariates as shown in the second column of Table 4.2. This model is the starting point for the following analysis of time-varying effects and builds a common basis for all five approaches under investigation.

We are aware that basing inference on this model, i.e. assuming that the model is given a priori, ignores model selection uncertainty of selected components such as included variables and non-linear effects. Focusing on a single model neglects that there usually exist other equally appropriate models. Different choices of the final PH model may also affect the analysis of time-varying effects. However, to ensure comparability of all approaches with respect to time-varying effects, this decision is necessary. As applied data analysis always requires decisions on model building strategies, we accept that all inference on time-varying effects is conditional on this decision of the final PH model.

4.1.3 Selection of time-varying effects

For all of the approaches under investigation several options must be set which may have a strong influence on the results. In the sequel, we give some information on these options and on the selection procedures.

FPT model

The FPT algorithm is provided as an add-on function in `Stata` (StataCorp, 2007) written by Patrick Royston and is available at `http://www.imbi.uni-freiburg.de/biom/Royston-Sauerbrei-book/index.html#progs`. From a technical side, a potential drawback of this method is the need for a split at each event time to get the correct risk sets for estimation of time-varying effects. For medium-sized data sets as the Rotterdam example, such an enlargement of the data is unproblematic. For extremely large data sets, though, technical problems may occur which require categorisation of survival times as discussed in Appendix B.

The results of the FPT algorithm with nominal significance level $\alpha = 0.01$ for selection of time-varying effects are shown in Table 4.3. In each iteration of the procedure, the deviance difference between the model with time-varying effect for the covariate under investigation is compared to the model with time-constant effect for the same covariate. In iteration 1, $\log(X_6)$ has the largest deviance difference for the best FP2 compared to constant effect. The deviance difference of 83.218 corresponds to a p value < 0.0001. Thus, the first test of the hierarchical test procedure (best FP2 vs. constant effect) is significant. The following test on the best FP2 vs. default $\log(t)$ is not significant ($p = 0.1160$). Consequently, the default transformation is used as basis for the time-varying effect. In the same way the default transformation is selected for X_{5e}^2, X_9 and X_8 in iterations 2 to 4. In iteration 5, the largest deviance difference is observed for X_1, which is not significant ($p = 0.0426$). Consequently the algorithm stops and the final model includes time-varying effects for X_{5e}^2, $\log(X_6)$, X_8 and X_9.

The decision for X_{5e}^2 in iteration 2 is close ($dev = 17.419$). X_{3a} has an only marginally smaller deviance difference of 17.396 with identical p-value. This close decision influences all following selection steps. If X_{3a} would have been selected instead of X_{5e}^2, the selection procedure would stop in iteration 3, because the smallest p value ($p = 0.0222$ for X_9) is not significant at the 1% level. Hence, the final model would include time-varying effects for $\log(X_6)$ and X_{3a} only.

The standard FPT algorithm uses $\log(t)$ as default transformation of time. When omitting the use of a default transformation, the FPT algorithm selects time-varying effects for the same covariates, because this decision is independent of the default function. Yet, the selected FP transformations change. In the Rotterdam data, decisions in the first two iterations are identical to those shown in Table 4.3, because for both X_{5e}^2 and $\log(X_6)$ the $\log$ transformation is the best FP1 transformation. In iteration 3, though, the best FP1 power -0.5 is chosen for X_9. Hence, deviances and p values of all consecutive iterations change slightly. In the fourth iteration, the time-varying effect for X_8 is selected with the best FP1 power 0.5. Consequently, the coefficients of the final model differ slightly from those derived with default

	best FP2		best FP1		default	p values of test vs. best FP2		
Variable	powers	deviance	power	deviance	deviance	constant	default	best FP1
Iteration 1								
X_1	0.5 0.5	9.450	-1	3.654	1.131	0.0508	0.0399	0.0551
X_{3a}	-2 0	26.295	0.5	25.241	24.863	0.0000	0.6981	0.5903
X_{3b}	-2 -0.5	13.383	0	12.189	12.189	0.0095	0.7546	0.5507
X_{4b}	-2 2	7.748	-1	6.668	5.115	0.1013	0.4518	0.5824
X_{5e}^2	-2 -1	21.071	0	18.694	18.694	0.0003	0.4980	0.3047
$\mathbf{\log(X_6)}$	1 2	**83.218**	0	77.307	77.307	**0.0000**	0.1160	0.0521
X_8	-1 -1	6.453	-2	3.200	0.349	0.1678	0.1067	0.1966
X_9	0.5 0.5	12.310	-1	7.648	4.600	0.0152	0.0524	0.0972
Iteration 2								
X_1	0.5 0.5	9.937	-1	3.920	1.330	0.0415	0.0350	0.0494
X_{3a}	-2 0	17.396	0.5	16.369	15.982	0.0016	0.7024	0.5983
X_{3b}	-2 -0.5	10.286	0	8.920	8.920	0.0359	0.7135	0.5051
X_{4b}	-2 -2	2.821	-2	2.677	0.481	0.5882	0.5049	0.9304
$\mathbf{X_{5e}^2}$	-2 -1	**17.419**	0	14.268	14.268	**0.0016**	0.3689	0.2069
X_8	-1 -1	7.306	-2	2.158	1.327	0.1206	0.1126	0.0762
X_9	0.5 0.5	10.810	-1	5.931	3.102	0.0288	0.0524	0.0872
Iteration 3								
X_1	0.5 0.5	8.755	3	2.658	0.471	0.0675	0.0405	0.0474
X_{3a}	-2 0	10.113	0.5	9.247	8.569	0.0386	0.6722	0.6486
X_{3b}	-2 -0.5	5.461	0	4.167	4.167	0.2432	0.7306	0.5237
X_{4b}	-2 -1	2.451	-2	2.095	0.138	0.6534	0.5101	0.8370
X_8	-1 -1	10.887	0.5	5.187	4.535	0.0279	0.0957	0.0578
$\mathbf{X_9}$	0.5 0.5	**14.077**	-0.5	10.377	7.740	**0.0071**	0.0963	0.1572
Iteration 4								
X_1	0.5 0.5	8.790	3	6.057	0.399	0.0666	0.0386	0.2550
X_{3a}	-2 0	9.948	0.5	9.096	8.272	0.0413	0.6424	0.6532
X_{3b}	-2 -0.5	5.404	0	4.105	4.105	0.2483	0.7292	0.5221
X_{4b}	-2 -1	2.478	-2	2.107	0.130	0.6486	0.5033	0.8307
$\mathbf{X_8}$	-1 -1	**13.378**	0.5	7.407	6.985	**0.0096**	0.0940	0.0505
Iteration 5								
X_1	0.5 0.5	**9.873**	3	6.375	0.146	0.0426	0.0210	0.1740
X_{3a}	-2 0	9.710	0.5	8.823	7.972	0.0456	0.6283	0.6416
X_{3b}	-2 -1	5.053	0	3.736	3.736	0.2819	0.7251	0.5176
X_{4b}	-2 -1	2.587	-2	2.281	0.195	0.6292	0.4953	0.8580

Table 4.3: Forward selection based on deviance differences for FPT in the Rotterdam breast cancer series. Covariates for which time-varying effects are selected are in bold.

time-transformation. However, selected time-varying effects for X_{5e}^2 and $\log(X_6)$ are virtually identical. For X_8 and X_9, where a different functional form is selected, time-varying effects are at least similar in shape (see Figure A.1 in the Appendix).

Variable	best power(s)	p value
Iteration 1		
$\mathbf{X_1}$	**-1**	**<0.001**
$\mathbf{X_{3a}}$	**0**	**<0.001**
X_{3b}	0	0.105
X_{4b}	-2	0.101
$\mathbf{X_{5e}^2}$	**-0.5**	**0.006**
$\mathbf{\log(X_6)}$	**0**	**<0.001**
X_8	1 2	0.039
X_9	-2 -2	0.012
Iteration 2		
X_1	0.5 0.5	0.111
$\mathbf{X_{3a}}$	**0.5**	**0.006**
X_{3b}	0	0.365
X_{4b}	-2	0.395
$\mathbf{X_{5e}^2}$	**-1**	**0.003**
$\mathbf{\log(X_6)}$	**0**	**<0.001**
X_8	1 2	0.088
$\mathbf{X_9}$	**-2 -2**	**0.008**
Iteration 3		
X_1	3	0.134
$\mathbf{X_{3a}}$	**0.5**	**0.003**
X_{3b}	0	0.308
X_{4b}	-2	0.411
$\mathbf{X_{5e}^2}$	**-0.5**	**<0.001**
$\mathbf{\log(X_6)}$	**0**	**<0.001**
X_8	1 2	0.012
$\mathbf{X_9}$	**-2 -2**	**0.006**
Iteration 4		
X_1	3	0.121
$\mathbf{X_{3a}}$	**0.5**	**0.007**
X_{3b}	0	0.330
X_{4b}	-2	0.391
$\mathbf{X_{5e}^2}$	**-0.5**	**<0.001**
$\mathbf{\log(X_6)}$	**0**	**<0.001**
X_8	1 2	0.012
$\mathbf{X_9}$	**-2 -2**	**0.006**

Table 4.4: Backfitting algorithm based on likelihood ratio tests for the Dynamic Cox model in the Rotterdam breast cancer series. Covariates with significant time-varying effects are in bold.

Dynamic Cox model

The Dynamic Cox model is available as an add-on library for `S-Plus` written by Ursula Berger. In older `S-Plus` versions, the functions can easily be executed. For execution in newer versions, updating of the functions would be needed.

Another major drawback of the program are potential technical problems due to the required enlargement of the data. With data sets of more than a few hundred observations, i.e. distinct survival times, splitting at each event time fails. A possible solution to this problem is categorisation of survival time, which is discussed in Appendix B in more detail. In the Rotterdam data, survival time is categorised into one month intervals to reduce the enlarged data set to a manageable size.

For a nominal significance level of $\alpha = 0.01$ for selection of time-varying effects, the backfitting algorithm of the Dynamic Cox model stops after four iterations (see Table 4.4). Variable $\log(X_6)$ has a highly significant time-varying effect with power 0 in all iterations. The time-varying effects of X_{3a} and X_{5e}^2 also remain significant throughout all iterations. Their FP powers, though, are subject to changes. The powers of X_9, on the contrary, are stable over all iterations, but are significant only from iteration 2 on. The opposite is observed for X_1,

Variable	AIC_c
Iteration 0	
PH	9568.14
Iteration 1	
X_1	9561.94
X_{3a}	9547.22
X_{3b}	9559.76
X_{4b}	9556.91
X_{5e}^2	9554.57
$\mathbf{log(X_6)}$	**9491.38**
X_8	9569.14
X_9	9563.62
Iteration 2	
X_1	9485.18
$\boldsymbol{X_{3a}}$	**9478.75**
X_{3b}	9485.77
X_{4b}	9493.40
X_{5e}^2	9482.30
X_8	9491.10
X_9	9488.04
Iteration 3	
$\boldsymbol{X_1}$	**9473.19**
X_{3b}	9477.40
X_{4b}	9480.88
X_{5e}^2	9475.70
X_8	9477.72
X_9	9475.21

Variable	AIC_c
Iteration 4	
X_{3b}	9472.07
X_{4b}	9475.30
$\boldsymbol{X_{5e}^2}$	**9470.18**
X_8	9470.92
X_9	9472.50
Iteration 5	
X_{3b}	9470.39
X_{4b}	9472.32
$\boldsymbol{X_8}$	**9465.97**
X_9	9466.72
Iteration 6	
X_{3b}	9466.31
X_{4b}	9468.12
$\boldsymbol{X_9}$	**9460.20**
Iteration 7	
X_{3b}	**9460.56**
X_{4b}	9462.29

Table 4.5: Forward selection procedure based on the AIC_c for the Empirical Bayes model in the Rotterdam breast cancer series. Covariates for which time-varying effects are selected are in bold.

which looses significance of the time-varying effect after iteration 1. In the final model it is included with a constant effect, which is also the case for X_{3b}, X_{4b} and X_8.

Empirical Bayes model

The Empirical Bayes model is fitted using `remlreg` in `BayesX` (Belitz et al., 2009). The random effects priors for the time-varying effects are chosen as cubic P-splines with second order random walk penalty and 20 equidistant knots, which is the default setting. For fixed effects, diffuse priors are used.

As the approach does not automatically involve selection of time-varying effects, a manual

forward selection type procedure comparing the models based on the AIC_c is applied following Hofner et al. (2010). In the Rotterdam data this forward selection procedure stops after 7 iterations (see Table 4.5) with time-varying effects for X_1, X_{3a}, X_{5e}^2, $\log(X_6)$, X_8 and X_9. Thus, it results in a quite complex model with six time-varying effects and two time-constant effects for X_{3b} and X_{4b}.
Selection of time-varying effects based on BIC_c instead of AIC_c results in a more parsimonious model. With this criterion, the selection procedure stops in iteration three with time-varying effects for $\log(X_6)$ and X_{3a} only.

Semiparametric Extended Cox model

The Semiparametric Extended Cox model is implemented in an `R` package (Scheike, 2009) which is available on CRAN (`http://cran.r-project.org/`). The functions are easy to use but require some specifications and further programming by the user. Selection of time-varying effects is not included in the package and is in the sequel realised by the backward elimination strategy proposed in Scheike and Martinussen (2004). The final cumulative regression coefficients are smoothed using local polynomial regression with quadratic polynomials, tricube weight function and a nearest neighbour fraction of 0.7 provided by the `R` package `locfit` (Loader, 2007).

The choice of a suitable bandwidth for the kernel smoother used within the estimation procedure is also important. For the moment, we stick to the default value of 0.5. The influence of the bandwidth is discussed in more detail in Section 6.8.

For testing on a single time-varying effect, we use the variance weighted Kolmogorov-Smirnov type test ($w = 1$) with a nominal significance level of $\alpha = 0.01$ in the backward elimination procedure. In each step, the largest p value is considered. If it is larger than the nominal significance level, the time-varying effect of the covariate is changed to a constant effect.
Table 4.1 shows the results of the selection procedure. In iteration 1, the largest p value is 0.719 for X_{4b}, which is not significant. In iterations 2 to 7, time-varying effects for covariates X_9, X_8, X_{3b}, X_1, X_{5e}^2 and X_{3a} are set to constant in the same way. The p value for $\log(X_6)$ in iteration 8 equals $p = 0.000$ and thus the time-varying effect is kept in the model and the procedure stops.

The model fitting procedure provides estimates for the cumulative time-varying effects $B(t)$. The first derivative $\hat{\beta}(t)$ is obtained as a local slope estimate of a local polynomial regression on $\hat{B}(t)$. Both the estimated cumulative effect $\hat{B}(t)$ and the time-varying effect $\hat{\beta}(t)$ for $\log(X_6)$ are shown in Figure 4.2.

The selection of time-varying effects strongly depends on the test that is applied. The vari-

Variable	p value
Iteration 1	
X_1	0.361
X_{3a}	0.142
X_{3b}	0.489
$\boldsymbol{X_{4b}}$	**0.719**
X_{5e}^2	0.302
$\log(X_6)$	0.004
X_8	0.319
X_9	0.432
Iteration 2	
X_1	0.330
X_{3a}	0.081
X_{3b}	0.416
X_{5e}^2	0.324
$\log(X_6)$	0.001
X_8	0.367
$\boldsymbol{X_9}$	**0.422**
Iteration 3	
X_1	0.260
X_{3a}	0.114
X_{3b}	0.335
X_{5e}^2	0.183
$\log(X_6)$	0.000
$\boldsymbol{X_8}$	**0.683**

Variable	p value
Iteration 4	
X_1	0.263
X_{3a}	0.098
$\boldsymbol{X_{3b}}$	**0.356**
X_{5e}^2	0.191
$\log(X_6)$	0.000
Iteration 5	
$\boldsymbol{X_1}$	**0.293**
X_{3a}	0.057
X_{5e}^2	0.150
$\log(X_6)$	0.000
Iteration 6	
X_{3a}	0.041
$\boldsymbol{X_{5e}^2}$	**0.204**
$\log(X_6)$	0.000
Iteration 7	
$\boldsymbol{X_{3a}}$	**0.028**
$\log(X_6)$	0.000
Iteration 8	
$\log(X_6)$	**0.000**

Figure 4.1: Backward elimination based on the p value of the variance weighted Kolmogorov-Smirnov type test for the Semiparametric Extended Cox model in the Rotterdam breast cancer series. Covariates for which time-varying effects are eliminated from the model are in bold.

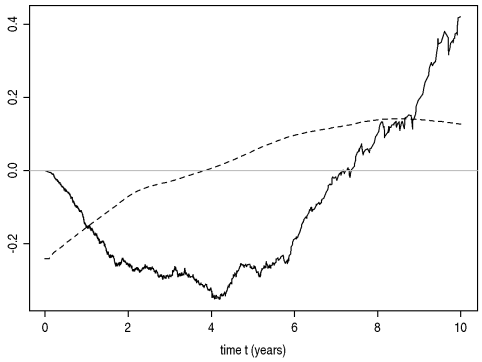


Figure 4.2: Estimated cumulative regression function $\hat{B}(t)$ (—) and smoothed time-varying effect $\hat{\beta}(t)$ (**- -**) of covariate $\log(X_6)$. $\hat{\beta}(t)$ is obtained as the local slope estimate of a local polynomial regression on $\hat{B}(t)$ with a nearest neighbour fraction of 0.7.

ance weighted Kolmogorov-Smirnov type test with end points removed ($w = 2$) gives slightly different p values and excludes time-varying effects in a different order, but results in an identical final model. The unweighted test ($w = 0$), though, results in a PH model without any time-varying effects.

Reduced Rank model

The Reduced Rank approach is implemented in the `R` package `coxvc` (Perperoglou, 2005), which requires several decisions that influence estimation results. These include decisions on the choice of the optimal rank and the time functions for modelling of time-varying effects.

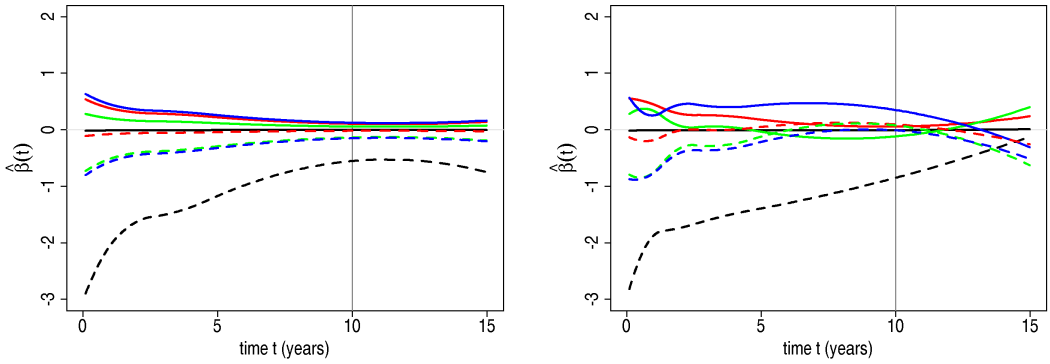
The optimal rank is chosen based on the AIC, following the recommendation of Perperoglou et al. (2006b). The time functions are quadratic B-Splines with interior knots at the quartiles of uncensored event times.
With these choices, the optimal model is of rank 3. As no selection of time-varying effects is applied, all effects vary with time. Figure 4.3 gives an overview of all possible models of rank 1 to 6. The effect functions are relatively smooth for lower ranks, but become more wiggly for larger ranks.

4.1.4 Investigation of the scaled Schoenfeld residuals

To get a first impression of the nature and extent of the time-varying behaviour of effects, we use the PH test based on the scaled Schoenfeld residuals (Grambsch and Therneau, 1994) for different functions of time.
The p values of PH tests (Table 4.6) indicate time-varying effects for X_{5e}^2 and $\log(X_6)$, which are significant at the 1% level, irrespective of the time transformation. Furthermore, there

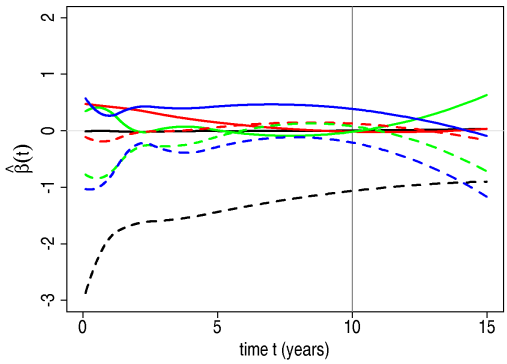
	p value			
Variable	t	$rank(t)$	$\log(t)$	$\sqrt{t}$
X_1	0.1880	0.4744	0.5835	0.3179
X_{3a}	**0.0070**	0.0183	0.0268	0.0108
X_{3b}	0.3379	0.1531	0.1311	0.2136
X_{4b}	0.9420	0.8956	0.8939	0.9711
X_{5e}^2	**0.0037**	**0.0005**	**0.0003**	**0.0009**
$\log(X_6)$	**0.0000**	**0.0000**	**0.0000**	**0.0000**
X_8	0.0396	**0.0023**	**0.0079**	0.0120
X_9	0.0387	**0.0020**	**0.0017**	**0.0079**

Table 4.6: P values for test on PH for different time transformations in the Rotterdam breast cancer series. Bold numbers mark significant time-varying effects at the 1% level.

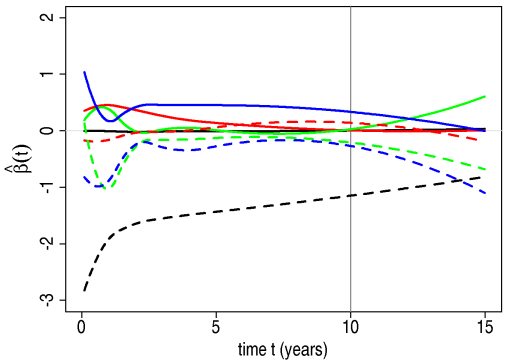


Rank 1: AIC=22197.24

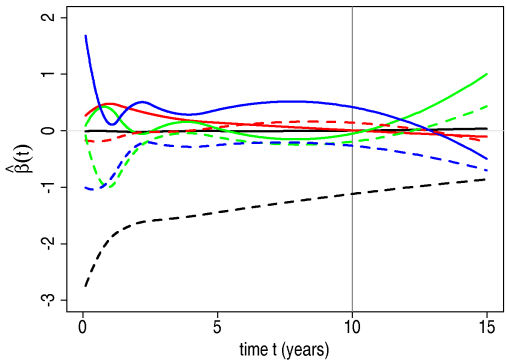
Rank 2: AIC=22141.56



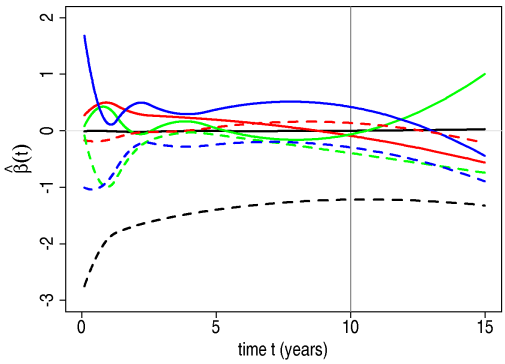
Rank 3: AIC=22141.02



Rank 4: AIC=22146.88



Rank 5: AIC=22152.72



Rank 6: AIC=22157.54

Figure 4.3: Estimated effects for different ranks of the Reduced Rank model in the Rotterdam breast cancer series for covariates X_1 (—), X_{3a} (—), X_{3b} (—), X_{4b} (—), X_{5e}^2 (- -), $\log(X_6)$ (- -), X_8 (- -) and X_9 (- -). The rank 3 model has the smallest AIC and $r = 3$ is therefore chosen as the optimal rank.

is some evidence for time-varying effects of X_8 and X_9 which are significant (or close to significant) for three of the tests.

In the sequel, we plot the smoothed scaled Schoenfeld residuals of the CoxPH model and their 95% pointwise confidence intervals as reference functions together with the (time-varying) effects estimated by the five competitive approaches and the CoxPH model. Other choices for the reference function such as piecewise constant effects would also be possible. Yet, due to the discretisation the latter approach is very sensitive to the number and position of jump times and estimates may be extremely instable. To avoid these problems we decide in favour of the smoothed Schoenfeld residuals. Although they also depend on the choice of a suitable bandwidth, they provide a continuous representation of the time-varying behaviour by smoothing over adjacent values. Hence, for not too small bandwidth, we deem the smoothed Schoenfeld residuals a sensible method to reflect the general tendency of (potential) time-varying effects. To obtain the smoothed curves and confidence intervals, the `loess` smoother implemented in `R` with span 0.7 is applied.

4.1.5 Comparison of approaches

The models selected by the five approaches differ considerably in terms of the number and shape of time-varying effects (Table 4.7). The only agreement is observed for $\log(X_6)$, which is modelled as time-varying by all approaches. While the Semiparametric Extended Cox Model does not select any further time-varying effects, FPT and the Dynamic Cox Model select four time-varying effects in total, the Empirical Bayes Model even six. The Reduced Rank Model, which does not incorporate selection of time-varying effects, models the effects of all eight covariates as time-varying.

Estimated effects

The estimated effects for the five different approaches are shown in Table 4.7. It is striking that the time-constant effects of all non-PH models are relatively close to the CoxPH effects, irrespective of the (time-varying) effects selected for the other covariates. Since not all approaches yield a functional form for time-varying effects, the selected effects are additionally compared graphically. The smoothed Schoenfeld residuals and their 95% pointwise confidence intervals are used to represent the 'raw' data, i.e. the true time-varying pattern of covariate effects. Furthermore, the CoxPH estimate is included in all figures representing the standard analysis.

For most covariates, the 95% pointwise confidence intervals of the smoothed Schoenfeld residuals are rather wide and cover all estimated effects, time-varying or not. Only X_{5e}^2 and $\log(X_6)$ show strong indications against time-constant effects. This fact is not discussed fur-

Variable	Estimated effects						
	CoxPH model	FPT		Dynamic Cox model	Empirical Bayes model	Semi-parametric Extended Cox model	Reduced Rank model
		with default transformation	w/o default transformation				
X_1	-0.013	-0.013	-0.013	-0.013	$\hat{\beta}_{X_1}(t)^*$	-0.013	$\hat{\beta}_{X_1}(t)^*$
X_{3a}	0.249	0.254	0.253	**0.701-0.254$t^{0.5}$**	$\hat{\beta}_{X_{3a}}(t)^*$	0.253	$\hat{\beta}_{X_{3a}}(t)^*$
X_{3b}	0.171	0.165	0.166	0.148	0.153	0.162	$\hat{\beta}_{X_{3b}}(t)^*$
X_{4b}	0.354	0.375	0.374	0.375	0.372	0.370	$\hat{\beta}_{X_{4b}}(t)^*$
X_{5e}^2	-1.710	**-2.028+0.436log(t)**	**-2.030+0.440log(t)**	**-1.126-0.733$t^{-0.5}$**	$\hat{\beta}_{X_{5e}^2}(t)^*$	-1.710	$\hat{\beta}_{X_{5e}^2}(t)^*$
$\log(X_6)$	-0.032	**-0.133+0.114log(t)**	**-0.133+0.114log(t)**	**-0.121+0.107log(t)**	$\hat{\beta}_{\log(X_6)}(t)^*$	$\hat{\beta}_{\log(X_6)}(t)^*$	$\hat{\beta}_{\log(X_6)}(t)^*$
X_8	-0.390	**-0.608+0.265log(t)**	**-0.994+0.371$t^{0.5}$**	-0.422	$\hat{\beta}_{X_8}(t)^*$	-0.412	$\hat{\beta}_{X_8}(t)^*$
X_9	-0.444	**-0.640+0.237log(t)**	**0.063-0.702$t^{-0.5}$**	**-0.274-0.266t^{-2} -0.125t^{-2}log(t)**	$\hat{\beta}_{X_9}(t)^*$	-0.453	$\hat{\beta}_{X_9}(t)^*$

** No simple functional form available*

Table 4.7: Comparison of selected effects for different approaches in the Rotterdam breast cancer series. Selected time-varying effects are in bold.

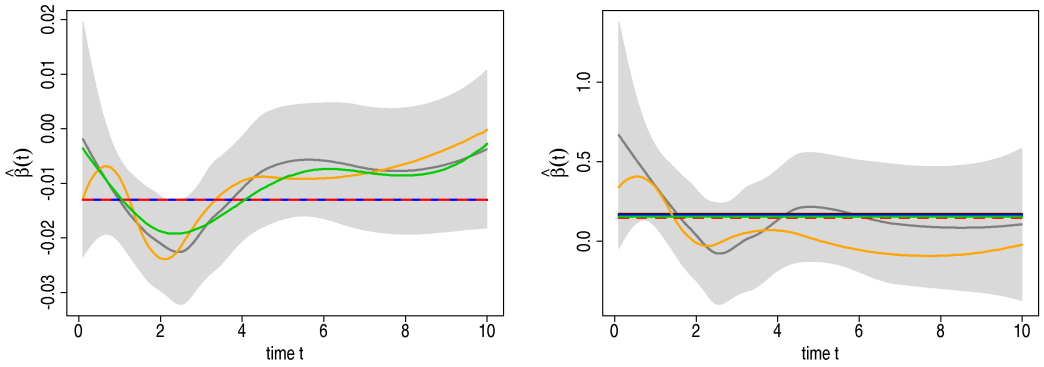


Figure 4.4: Effects for X_1 (left) and X_{3b} (right) in the Rotterdam breast cancer series estimated by the CoxPH model (—), FPT (—), the Dynamic Cox model (- -), the Empirical Bayes model (—), the Semiparametric Extended Cox model (—), the Reduced Rank model (—) and the smoothed Schoenfeld residuals (—) with their 95% pointwise confidence intervals ( ).

---

ther for individual covariates.

The estimated effects for X_1 are displayed in Figure 4.4 (left panel). FPT, the Semiparametric Extended Cox model and the Dynamic Cox model estimate constant effects identical to the CoxPH model. The Reduced Rank and Empirical Bayes model, on the contrary, give time-varying effects with a clear minimum between 2 and 3 years and some further local extrema which reflect the shape of the smoothed Schoenfeld residuals.
A similar behaviour is observed for the effect of X_{3b} (Figure 4.4, right panel), where only the Reduced Rank model estimates a time-varying effect whose shape resembles the smoothed Schoenfeld residuals.

For X_{3a} (Figure 4.5, left panel) the Reduced Rank, Dynamic Cox and Empirical Bayes models estimate decreasing effects which are close to linear. This trend is also reflected by the smoothed Schoenfeld residuals. The estimates of the remaining models are time-constant.

For X_{4b}, the Empirical Bayes model, FPT and the Semiparametric Extended Cox model estimate time-constant effects close to the CoxPH effect (Figure 4.5, right panel). The smoothed Schoenfeld residuals fluctuate around these constant effects with some local extrema especially at the beginning, a pattern that is closely followed by the Reduced Rank estimate.

The smoothed Schoenfeld residuals for X_{5e}^2 (Figure 4.6, left panel) indicate an increasing effect. As the effect is negative, this means that the protective effect of X_{5e}^2 is strong initially but diminishes in the first five years. All approaches but the Semiparametric Extended Cox model detect this time-varying pattern. The Empirical Bayes estimate, though, is in the first years considerably flatter than the other effects.

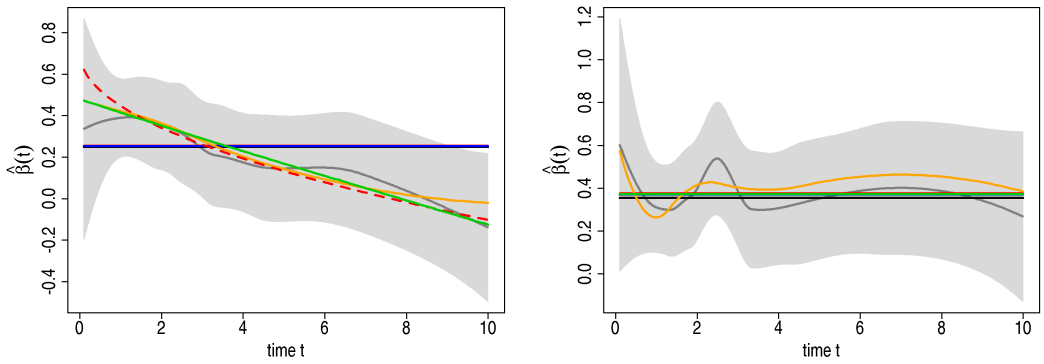


Figure 4.5: Effects for X_{3a} (left) and X_{4b} (right) in the Rotterdam breast cancer series estimated by the CoxPH model (—), FPT (—), the Dynamic Cox model (- -), the Empirical Bayes model (—), the Semiparametric Extended Cox model (—), the Reduced Rank model (—) and the smoothed Schoenfeld residuals (—) with their 95% pointwise confidence intervals ( ).

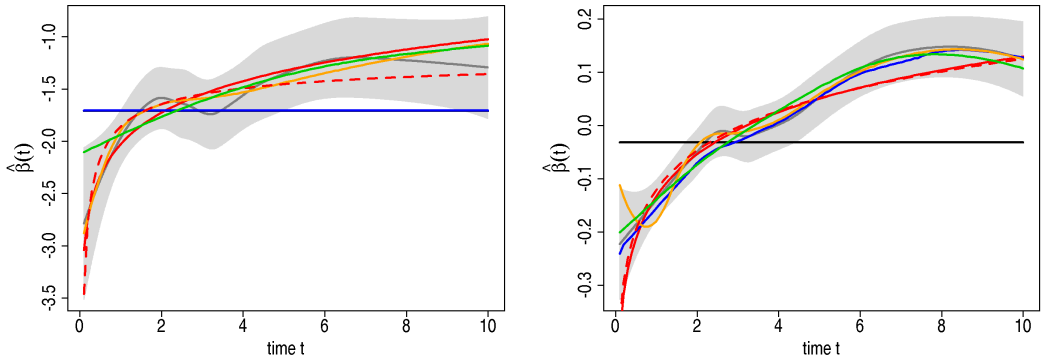


Figure 4.6: Effects for X_{5e}^2 (left) and $\log(X_6)$ (right) in the Rotterdam breast cancer series estimated by the CoxPH model (—), FPT (—), the Dynamic Cox model (- -), the Empirical Bayes model (—), the Semiparametric Extended Cox model (—), the Reduced Rank model (—) and the smoothed Schoenfeld residuals (—) with their 95% pointwise confidence intervals ( ).

A similarly strong time-varying pattern is observed for the effect of $\log(X_6)$ (Figure 4.6, right panel). The smoothed Schoenfeld residuals show a clearly increasing effect, which crosses zero at about 3 years, i.e. the effect inverts. This is reflected by all estimated effects. However, the Reduced Rank estimate again shows several local extrema, especially at the beginning.

For X_8 (Figure 4.7, left panel), the smoothed Schoenfeld residuals and their 95% pointwise confidence intervals hint at a time-varying effect of increasing nature, which flattens off or even decreases later in time. The effects estimated by FPT, the Reduced Rank model

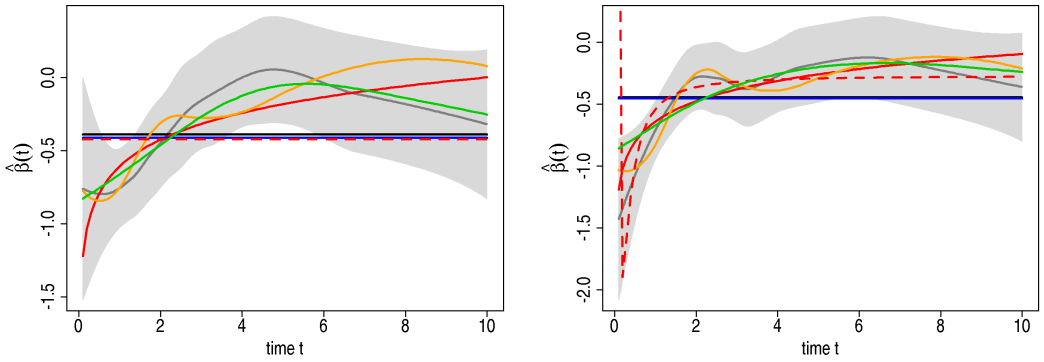


Figure 4.7: Effects for X_8 (left) and X_9 (right) in the Rotterdam breast cancer series estimated by the CoxPH model (—), FPT (—), the Dynamic Cox model (- -), the Empirical Bayes model (—), the Semiparametric Extended Cox model (—), the Reduced Rank model (—) and the smoothed Schoenfeld residuals (—) with their 95% pointwise confidence intervals ( ).

and the Empirical Bayes model describe a time-varying behaviour of this type, although the Reduced Rank estimate shows even more local extrema than the smoothed Schoenfeld residuals.

A similar pattern is observed for X_9 (Figure 4.7, right panel), for which the smoothed Schoenfeld residuals suggest an initial increase levelling off to a plateau near zero at about 2 years. That is, the initially strong protective effect of X_9 vanishes over time. Besides FPT, the Reduced Rank model, the Empirical Bayes model and the Dynamic Cox model also estimate an increasing time-varying effect. Yet, the FP2 function selected by the latter shows an artefact for very small times.

Summing up, the Semiparametric Extended Cox model, the Empirical Bayes model, the Dynamic Cox model and the FPT model give flexible estimates in reasonable agreement with the smoothed Schoenfeld residuals. Due to the missing selection of time-varying effects in the Reduced Rank approach, all effects show a slightly time-varying behaviour, which might also influence the estimates for stronger time-varying effects. Furthermore, most estimates are rather wiggly, with several local extrema, which may result in overfitting of the data at hand.

Prediction error curves

Although the presented models appear rather different in terms of selected effects, differences in the prediction error are less pronounced. The apparent error for the five approaches, the CoxPH model and the Kaplan-Meier estimate are shown in the left panel of Figure 4.8. The apparent error of the FP models is only marginally better than that of the CoxPH model. The other three approaches are even worse than the CoxPH model. Due to

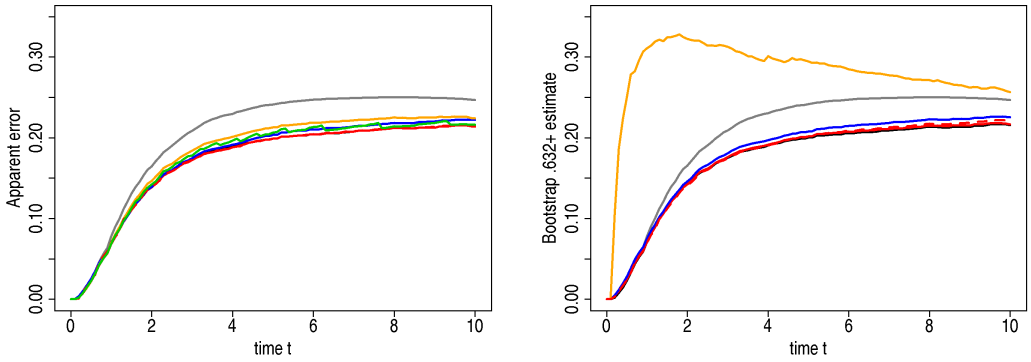


Figure 4.8: Apparent error (left) and Bootstrap .632+ estimate of prediction error (right) in the Rotterdam breast cancer series for the Kaplan-Meier estimate (—), the CoxPH model (—), FPT (—), the Dynamic Cox model (- -), the Empirical Bayes model (—), the Semiparametric Extended Cox model (—) and the Reduced Rank model (—).

extremely long execution times, predictions for the Empirical Bayes model are restricted to the 5% quantiles of uncensored event times. This restriction causes the jagged prediction error curve.

Since the apparent error tends to underestimate the true prediction error, the bootstrap .632+ estimate of prediction error is estimated for all approaches but the Empirical Bayes model, for which the manual selection procedure and the time-consuming computations are not feasible for 50 bootstrap replications. The bootstrap .632+ error of the two FP approaches is again similar but not better than the CoxPH model, while the Semiparametric Extended Cox model has a slightly worse prediction error (Figure 4.8, right panel). The Reduced Rank model shows an extremely large bootstrap .632+ error. As this approach does not incorporate selection of time-varying effects, it estimates all effects as time-varying and considerably overfits the data at hand. The resulting models are poorly transferable to new data and result in an extremely poor bootstrap cross-validation error which also dominates the bootstrap .632+ estimate.

4.2 A simulated data set

Conclusions about the estimated effects in the Rotterdam breast cancer series are limited, as the true effects are unknown. Therefore, all five approaches are additionally applied to a simulated data set. This data set contains 1000 observations with about 50% censoring and five standard normal distributed covariates. Two of the covariates are provided with time-varying effects while the others are constant. These effects correspond to those used

in the multivariable simulation study (see Section 6.1.2):

$$\begin{aligned}
\beta_{X_1}(t) &= 0.32 + \frac{1.42}{\exp(t)} - 0.02t^{0.7} \\
\beta_{X_2}(t) &= 0.1 + 0.8t^{0.3} \\
\beta_{X_3} &= 0.3 \\
\beta_{X_4} &= 0.5 \\
\beta_{X_5} &= 0.7
\end{aligned}$$

The shape of these effects is shown in the subsequent Figures 4.9 and 4.10.

4.2.1 Selection of time-varying effects

The FPT approach selects time-varying effects for exactly X_1 and X_2 (see Table A.1 in the Appendix for details). In the first iteration, all three tests of the hierarchical closed test procedure on the time-varying effect for X_1 are significant. Consequently, the best FP2 function (powers -1 and -1) is selected. The time-varying effect for X_2 is the default $\log$ transformation. Without the default, power 0.5 would be selected for X_2.
The Dynamic Cox model also starts with time-varying effects for X_1 and X_2 in its first iteration, but adds a time-varying effect for X_3 in iteration 2, leading to a model with one false positive time-varying effect. The final time-varying effect for X_1 is based on power -0.5, while for both X_2 and X_3 a $\log$ transformation is selected (Table A.2).
The Empirical Bayes model even selects four time-varying effects for X_1, X_2, X_3 and X_4 (Table A.3), where the true time-varying effects for X_1 and X_2 are selected in the first two iterations. With BIC_c, the forward selection procedure would stop in iteration four with three time-varying effects for X_1, X_2 and X_4.
Time-varying effects for the same covariates are selected by the Semiparametric Extended Cox model with variance weighted Kolmogorov-Smirnov type test (Table A.4). The other two test versions, though, select completely different models. The unweighted test ($w = 0$) selects time-varying effects for X_2 and X_4, missing the strong time-varying effect for X_1. On the contrary, the weighted test with removed tails ($w = 2$), selects a rather complex model with time-varying effects for X_1, X_2, X_4 and X_5. Although none of the three models is correct in terms of selected time-varying effects, in this example the weighted test ($w = 1$) seems to be the best alternative as it includes both time-varying effects with only one false positive time-varying effect.
For the Reduced Rank model, estimated time-varying effects for all five covariates of models with rank one to three are subject to changes (especially for times larger 10), while for larger ranks the effect estimates remain similar. The rank 3 model gives the best AIC in this

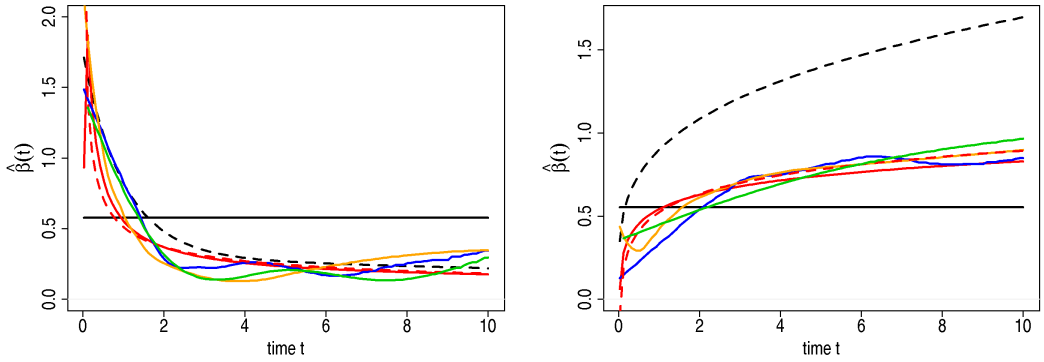


Figure 4.9: Effects for X_1 (left) and X_2 (right) in a simulated data set. Shown are the true effect (- -) and the effects estimated by the CoxPH model (—), FPT (—), the Dynamic Cox model (- -), the Empirical Bayes model (—), the Semiparametric Extended Cox model (—) and the Reduced Rank model (—).

example and seems to be a good compromise between flexibility and parsimony.

4.2.2 Comparison of approaches

A comparison of the selected models shows that only FPT is correct in terms of inclusion of time-varying effects, while the remaining approaches tend to select more (false positive) time-varying effects (Table 4.8). Like in the Rotterdam breast cancer series, estimated time-constant effects of all approaches are quite similar, irrespective of the effects of the other covariates in each model. Comparison to the true effects reveals that the time-constant effects for X_3 and X_5 are underestimated, while estimates for X_4 are reasonably close to the true effect.

Estimated effects

Graphical comparison of true and estimated effects shows that all approaches recover the strong initial decrease in the effect of X_1 (Figure 4.9, left panel). The FP2 function selected by the FPT approach shows an artefact at small times, while the FP1 effect estimated by the Dynamic Cox model is similar in shape, but does not show this erratic behaviour. Both estimated effects reflect the shape of the true effect, but show a slightly steeper decrease at the beginning. The time-varying effects of the other three approaches are rather wiggly and show some local extrema after the initial decrease.

The general shape of the increasing effect for X_2 is also reflected by most of the approaches (Figure 4.9, right panel). The curvature of both FP models mimics the true effect well. The Reduced Rank model is in general of similar curvature, but produces a local minimum in the

Variable	True model	Estimated effects $\hat{\beta}(t)$						
		CoxPH model	FPT		Dynamic Cox model	Empirical Bayes model	Semi-parametric Extended Cox model	Reduced Rank model
			with default	w/o default				
X_1	$\mathbf{0.32 + \frac{1.42}{\exp(t)} - 0.02t^{0.7}}$	0.577	$\mathbf{0.105 + 0.450t^{-1} + 0.112t^{-1}\log(t)}$	$\mathbf{0.106 + 0.448t^{-1} + 0.111t^{-1}\log(t)}$	$\mathbf{0.030 + 0.481t^{-0.5}}$	$\hat{\beta}_{X_1}(t)^*$	$\hat{\beta}_{X_1}(t)^*$	$\hat{\beta}_{X_1}(t)^*$
X_2	$\mathbf{0.1 + 0.8t^{0.3}}$	0.554	$\mathbf{0.543 + 0.125\log(t)}$	$\mathbf{0.253 + 0.226t^{0.5}}$	$\mathbf{0.524 + 0.160\log(t)}$	$\hat{\beta}_{X_2}(t)^*$	$\hat{\beta}_{X_2}(t)^*$	$\hat{\beta}_{X_2}(t)^*$
X_3	0.3	0.158	0.161	0.156	$\mathbf{0.226 - 0.109\log(t)}$	$\hat{\beta}_{X_3}(t)^*$	0.144	$\hat{\beta}_{X_3}(t)^*$
X_4	0.5	0.466	0.462	0.466	0.467	$\hat{\beta}_{X_4}(t)^*$	$\hat{\beta}_{X_4}(t)^*$	$\hat{\beta}_{X_4}(t)^*$
X_5	0.7	0.550	0.579	0.584	0.553	0.576	0.559	$\hat{\beta}_{X_5}(t)^*$

** No simple functional form available*

Table 4.8: Comparison of selected effects for the different approaches in a simulated data set. Selected time-varying effects are in bold.

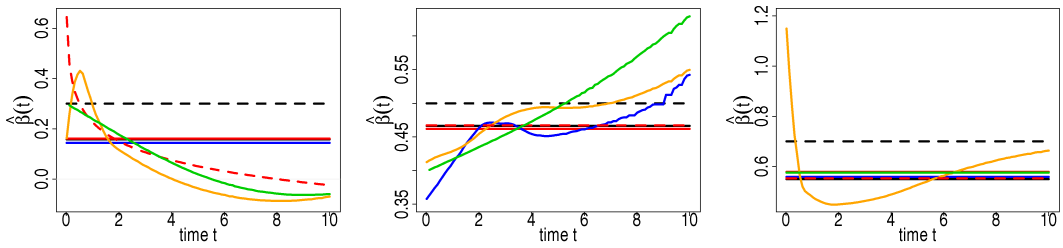


Figure 4.10: Effects for X_3 (left), X_4 (centre) and X_5 (right) in a simulated data set. Shown are the true effect (- -) and the effects estimated by the CoxPH model (—), FPT (—), the Dynamic Cox model (- -), the Empirical Bayes model (—), the Semiparametric Extended Cox model (—) and the Reduced Rank model (—).

---

first year. The Semiparametric Extended Cox model, on the contrary, shows the general tendency but gives a rather ragged estimate. The effect estimated by the Empirical Bayes model, though, is rather flat throughout and fails to recover the steep initial increase. Nevertheless, all five effects considerably underestimate the size of the effect and perform quite poor in this respect. This fact can be explained by a frailty effect and is discussed in more detail in the simulation study on p. 97.

The estimated effects of X_3 to X_5 are depicted in Figure 4.10. In general, the estimated effects tend to underestimate the true effect. For X_3, only FPT and the Semiparametric Extended Cox model estimate time-constant effects, the other three approaches select time-varying effects of decreasing nature. For X_4 on the contrary, the Empirical Bayes model, the Reduced Rank and the Semiparametric Extended Cox model decide on increasing time-varying effects. The latter is rather ragged for $t > 8$. However, the scale of the y axis is small and consequently absolute differences between estimated effects are not very large. Finally, the effect of X_5 is correctly modelled as time-constant by all approaches but the Reduced Rank model which shows a time-varying effect with a relatively strong initial decrease.

Prediction error curves

In this example, four of the approaches achieve a visible improvement in terms of the prediction error over the CoxPH model (Figure 4.11). The Empirical Bayes model and both FP models perform about equally well, while the Reduced Rank model shows a slightly larger prediction error after time 6. Hence, in this simulated data set the true prediction error seems rather unaffected by the missing selection of time-varying effects in the Reduced Rank approach.

Only the prediction error of the Semiparametric Extended Cox model is worse than that of the CoxPH model. This seems surprising, as the effect estimates are inconspicuous. Yet, the estimate of the cumulative baseline hazard severely differs from the true cumulative

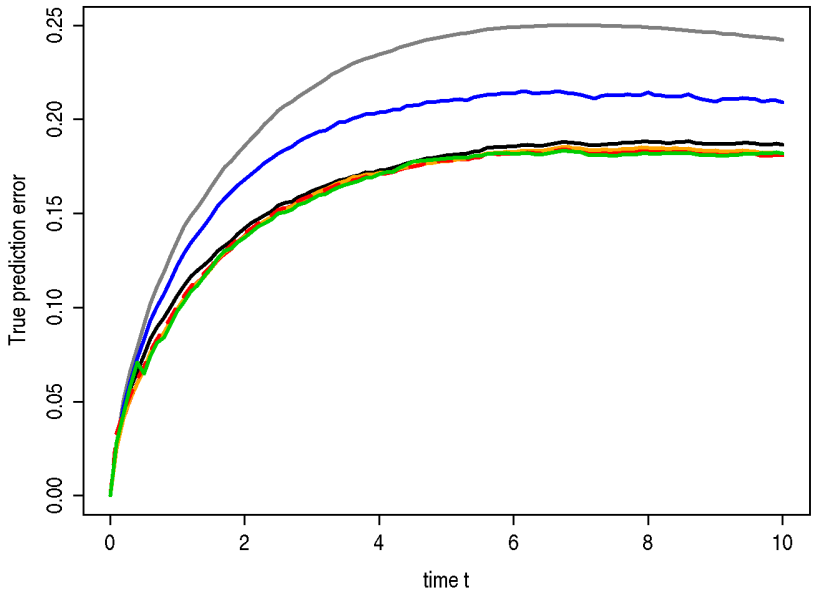


Figure 4.11: True prediction error in a simulated data set for the Kaplan-Meier estimate (—), the CoxPH model (—), FPT (—), the Dynamic Cox model (- -), the Empirical Bayes model (—), the Semiparametric Extended Cox model (—) and and the Reduced Rank model (—).

baseline hazard (see Figure A.3 in the Appendix), which is most likely the reason for the inflated prediction error.

4.3 Summary and concluding remarks

4.3.1 Performance of approaches

Investigations in the two examples reveal that FPT is very good at detecting time-varying effects, while the Reduced Rank approach results in far too complex models, as it does not include selection of time-varying effects. The other three approaches show a tendency towards false positive decisions in the selection process, but also some probably false negative decisions. In particular the Semiparametric Extended Cox model seems to miss an important time-varying effect in the Rotterdam breast cancer series .

The time-varying effects estimated by the Reduced Rank model show many local extrema and seem to overfit the data. The effect estimates of all other approaches seem to be reasonably good. In the Rotterdam breast cancer series they reflect the general pattern of the smoothed Schoenfeld residuals or select time-constant effects which are covered by the 95% pointwise confidence intervals. In the simulated data set, on the contrary, all approaches underestimate the increasing time-varying effect, due to a frailty effect, but perform

quite well in estimating the decreasing effect.

Although the selected models of all five approaches are rather different in terms of selected (time-varying) effects, their prediction error is relatively similar. In the Rotterdam breast cancer series, the two FP approaches show a prediction performance similar but not better than the CoxPH model, while the Semiparametric Extended Cox model and the Empirical Bayes model are slightly inferior. The Reduced Rank model with time-varying effects for all eight covariates considerably overfits the data, which results in a considerably inflated prediction error. Hence, the time-varying effects in the data seem to be not strong enough to seriously influence the prediction performance if they are ignored.
In the simulated data set, the prediction performance of FPT, the Dynamic Cox model, the Empirical Bayes model and the Reduced Rank model are similarly good and show a clear improvement compared to the CoxPH model. Only the Semiparametric Extended Cox model shows an inflated prediction error.

4.3.2 Practical applicability

Although evidence about the performance of the approaches in these two examples is somewhat limited, we gained valuable insights into their practical applicability and usability.

The FPT approach is easy to use. The standard class of FPs is very flexible, but can easily be adjusted if required. The same applies for the default time transformation. Furthermore, FPs provide simple functional forms that are intuitive and easy to interpret. One drawback of the implementation, though, is the required enlargement of the data set, which may cause problems with extremely large data sets but is of less consequence to the small and medium sized data sets of common applications.

Although the Dynamic Cox model is equally flexible and easy to use, this approach suffers from two severe drawbacks. First, the enlargement of data sets already causes problems when the number of observations exceeds a few hundred. In these cases, categorisation of survival times is necessary to analyse the data. Second, the algorithm has been implemented in an older S-Plus version and is incompatible to current S-Plus versions.

The Empirical Bayes model is also a very flexible tool. This model can easily be applied also by users less experienced in Bayesian modelling, as it does not require, for example, decisions on the mixing of Markov chains. The splines depend, like all spline approaches, on the number and position of knots, but the default settings in combination with penalisation work pretty well in our examples. As Ruppert (2002) shows that their impact on the fit of penalised splines is small, modification of the default settings may be rarely required in practise. However, the implementation does not include automated selection of time-varying effects and manual selection based on information criteria can be very time consuming in

applications with many variables.

The cumulative time-varying effects $B(t)$, on which the Semiparametric Extended Cox model is based, provide asymptotic properties of tests but complicate investigation of $\beta(t)$, as transformation of the cumulative estimates is left to the user. Furthermore, the choice of the bandwidth is critical (Cortese et al., 2010), especially considering a potential sensibility of the tests to this decision.

The Reduced Rank model is theoretically a flexible tool for investigating time-varying effects. Yet, the implemented algorithm unfortunately does not include selection of time-varying effects which may also be the reason for the extremely bad bootstrap prediction error in the Rotterdam breast cancer series, as it limits transferability of selected models. In the simulated data example, though, the true prediction error did not show the same effect. Another potential drawback of the approach is that the choice of suitable time functions is left completely to the user. Usage of the recommended splines further introduces the typical problem of finding a suitable number and position of knots.

4.3.3 Recommendations

These (limited) comparisons suggest that the Reduced Rank model and the Dynamic Cox model are less suitable for practical applications due to the theoretical and technical reasons mentioned above. The Semiparametric Extended Cox model is not the method of choice for estimation of time-varying effects $\beta(t)$ either, because of its drawbacks with respect to the choice of a suitable bandwidth and its sensibility to the very same. Nevertheless, it is a flexible nonparametric tool for testing the PH assumption with known asymptotic properties, which may be of value when the interest lies on testing alone (depending still on a suitable bandwidth). When the interest lies in the shape of time-varying effects, the FPT approach or the Empirical Bayes approach seem to be preferable, as they provide flexible effects and easy to use programs.

Chapter 5

Beyond a single model: bootstrapping

5.1 Stability of FPT

To assess the stability of time-varying effects estimated by FPT, bootstrap resampling is applied to the Rotterdam breast cancer series and results are compared to those obtained for the original data.

To promote stability and to partly protect against overfitting, the FPT procedure utilises a default time transformation. Although this makes the approach less flexible in the time-varying function selected, the default transformation $\log(t)$ allows modelling of short-term effects and is a popular choice for modelling time-varying effects. To investigate the influence of this default time transformation on the shape of estimated time-varying effects, an additional version of FPT without the default time transformation is examined. This version differs only in the testing procedure, where the second test (best FP2 vs. default FP) is omitted.

To explore (in-)stability of the estimated time-varying effects, we draw $n_B = 1000$ bootstrap samples of size 2982 with replacement from the original data. Time-varying effects are selected at a nominal significance level of $\alpha = 0.01$ and $\alpha = 0.157$. The latter corresponds to the asymptotic significance level of the AIC for the inclusion of one additional time-varying effect. The effects estimated in the bootstrap samples are used to check the difference of selected functional forms to the reference function (fitted in the original data) and to summarise the variation among the bootstrap samples.

A large amount of bootstrap samples makes such investigations extremely computationally intensive. Hence, we follow the proposal of Sauerbrei et al. (2007) to 'categorise' survival times in half-year periods up to year 15 and a final period $>$15 years for computational

reasons. For details on categorisation of survival time see Appendix B.

The reference function

Since in a multivariable model the selected time-varying functions may depend on the sequence of time-varying effects entering the model, we will concentrate on the prognostic index (PI) as a univariate summary of the underlying multivariable model. Application of the first two steps of the MFPT procedure to the Rotterdam data as in Sauerbrei et al. (2007) leads to a PI of

$$\begin{aligned} PI = &- 0.013 X_1 + 0.249 X_{3a} + 0.171 X_{3b} + 0.354 X_{4b} \\ &- 1.681 X_{5e}^2 - 0.032 \log(X_6 + 1) - 0.389 X_8 - 0.443 X_9 \end{aligned}$$

in the final CoxPH model.

Categorising the survival times as described above, application of the FPT algorithm discovers a logarithmically decreasing time-varying effect

$$\hat{\beta}(t) = 1.335 - 0.361 \log(t)$$

for the PI for both significance levels and with / without default time transformation $\log(t)$.

Standardisation of effect estimates

Since the Cox model has no intercept term, the resulting time-varying function is standardised to have mean zero when averaged over the empirical distribution of the categorised time T. That is, the estimated time-varying effect in the original data (reference function) $\hat{\beta}_{orig}(t)$ is standardised to

$$\tilde{\beta}_{orig}(t) = \hat{\beta}_{orig}(t) - \frac{1}{n_T} \sum_{i=1}^{n_T} \hat{\beta}_{orig}(t_i),$$

where n_T is the number of survival times, i.e. the number of observations in the original data set.

In each bootstrap sample, FP powers (p_1 and p_2) and regression coefficients ($\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$) estimated by FPT are collected. With this information, the time-varying effects $\hat{\beta}(t)$ of the bootstrap samples are recalculated for the same survival times as in the original analysis to enable comparisons. These functions are then standardised in the same way as the reference function. The estimated time-varying effect of the bth bootstrap replication $\hat{\beta}_b(t)$ is thus standardised to

$$\tilde{\beta}_b(t) = \hat{\beta}_b(t) - \frac{1}{n_T} \sum_{i=1}^{n_T} \hat{\beta}_b(t_i).$$

We use Breiman's bagged estimator (Breiman, 1996) as a summary of all standardised bootstrap functions

$$\hat{\beta}_{bag}(t) = \frac{1}{n_B} \sum_{b=1}^{n_B} \tilde{\beta}_b(t),$$

where n_B is the number of bootstrap replications.

Measures of stability

To quantify the stability of time-varying effects, the total variation of bootstrap functions around the reference function $\mathcal{T}$, the within subset variance $\mathcal{V}$ and the squared deviation between bagged and reference curve $\mathcal{D}^2$ are applied, following Royston and Sauerbrei (2003). The total variation of bootstrap functions around the reference function $\mathcal{T}$ can be decomposed into the within-subset variance $\mathcal{V}(t)$ and the squared deviation between the bagged and reference curves $\mathcal{D}^2(t)$:

$$\begin{array}{ccccccc} \frac{1}{n_B} \sum_{b=1}^{n_B} [\tilde{\beta}_b(t) - \tilde{\beta}_{orig}(t)]^2 & = & \frac{1}{n_B} \sum_{b=1}^{n_B} [\tilde{\beta}_b(t) - \hat{\beta}_{bag}(t)]^2 & + & [\hat{\beta}_{bag}(t) - \tilde{\beta}_{orig}(t)]^2 \\ \mathcal{T}(t) & = & \mathcal{V}(t) & + & \mathcal{D}^2(t) \end{array}$$

A large $\mathcal{D}^2(t)$ points to a difference in shape between the reference and bagged curves. Large values of $\mathcal{V}(t)$, on the other hand, reflect a large random variation and other contributions to variability of the individual bootstrap curves $\tilde{\beta}_b(t)$ around the bagged function $\hat{\beta}_{bag}(t)$. In practise, the summary measures of $\mathcal{V}$ and $\mathcal{D}^2$ are calculated by averaging over the empirical distribution of time T in the original data as

$$\mathcal{V} = \frac{1}{n_T} \sum_{i=1}^{n_T} \mathcal{V}(t_i) \qquad \text{and} \qquad \mathcal{D}^2 = \frac{1}{n_T} \sum_{i=1}^{n_T} \mathcal{D}^2(t_i).$$

FP types

Fitted FPs are divided into different types according to the criteria in Table 5.1. For FP2 functions we distinguish between unimodal and monotonic curves. In most applications, one would probably expect a monotonic behaviour of time-varying effects, such as a monotonically decreasing effect of a prognostic factor over time (i.e. a diminishing influence of the factor on survival). FP1 functions are always monotonic and are further subdivided into linear, logarithmic and other FP1 functions and the sign of their coefficient β_1.

The background for this classification is that a certain curve shape may be described almost equally well by different FP functions. Furthermore, with bootstrap resampling, recovering the general shape of the time-varying effect seems to be sufficient and exact re-selection of the reference function fitted in the original data is not required. Hence, identical powers for

FP type	description	form of $\hat{\beta}(t)$
PH	time-constant effect	$\hat{\beta}(t) = \hat{\beta}_0$
log + (-)	log-transformation with positive (negative) coefficient	$\hat{\beta}(t) = \hat{\beta}_0 + (-)\ \hat{\beta}_1 \log(t)$
lin + (-)	linear transformation ($p_1 = 1$) with positive (negative) coefficient	$\hat{\beta}(t) = \hat{\beta}_0 + (-)\ \hat{\beta}_1 t$
other FP1	other FP1 transformations with $p_1 \notin \{0, 1\}$	$\hat{\beta}(t) = \hat{\beta}_0 \pm \hat{\beta}_1 t^{p_1}$
FP2 monotonic	FP2 with different powers $p_1 \neq p_2$ and $sign(\hat{\beta}_1\hat{\beta}_2)sign(p_2) = sign(p_1)$	$\hat{\beta}(t) = \hat{\beta}_0 \pm \hat{\beta}_1 t^{p_1} \pm \hat{\beta}_2 t^{p_2}$
FP2 unimodal	FP2 with different powers $p_1 \neq p_2$ and $sign(\hat{\beta}_1\hat{\beta}_2)sign(p_2) \neq sign(p_1)$ or FP2 with equal powers $p_1 = p_2$	$\hat{\beta}(t) = \hat{\beta}_0 \pm \hat{\beta}_1 t^{p_1} \pm \hat{\beta}_2 t^{p_2}$, $\hat{\beta}(t) = \hat{\beta}_0 \pm \hat{\beta}_1 t^{p} \pm \hat{\beta}_2 t^{p} \log(t)$
FP with asymptote	FP1s with $p_1 < 0$ and FP2s with $p_1 < 0$ and $p_2 < 0$	
FP w/o asymptote	FP1s with $p_1 \geq 0$ and/or FP2s with $p_1 \geq 0$ and $p_2 \geq 0$	

Table 5.1: Classification of time-varying effects into different FP types.

the time-varying effect in the original data and the effects selected in the bootstrap samples are not necessarily required.

Stability of selected effects

To illustrate the (in-)stability of selected time-varying functions, the reference function selected in the original data is plotted with its 95% pointwise analytical confidence intervals, all curves selected in the 1000 bootstrap samples and their bagging estimate. Figure 5.1 shows the four settings for FPT with and without default time transformation $\log(t)$ and with $\alpha = 0.01$ and 0.157, respectively.

For significance level $\alpha = 0.01$ and FPT with default time transformation (upper left panel), nearly all time-varying effects selected in the bootstrap samples are close to the reference effect. Only few functions show a different shape. This is not surprising, as the $\log$ transformation with negative coefficient, which describes a decreasing effect over time, is selected in 99% of the bootstrap samples (Table 5.2). In about half of these samples, it is the best FP1 function, while it is chosen as default time transformation in the other 50%. In these samples, FPT without default time transformation selects other FP1 functions instead. Without the default time transformation, the amount of effects with different shape slightly increases and leads to more variability. Most frequently FP1 functions with power 0.5 or -0.5 are selected instead of the $\log$ function, as shown in Table 5.3. Both powers in combination with negative and positive coefficients, respectively, result in similar curve shapes than the $\log$ function. Figure 5.2 shows the fitted time-varying effects of two exemplary bootstrap samples, for which a $\log$ transformation is selected with default time transformation and powers

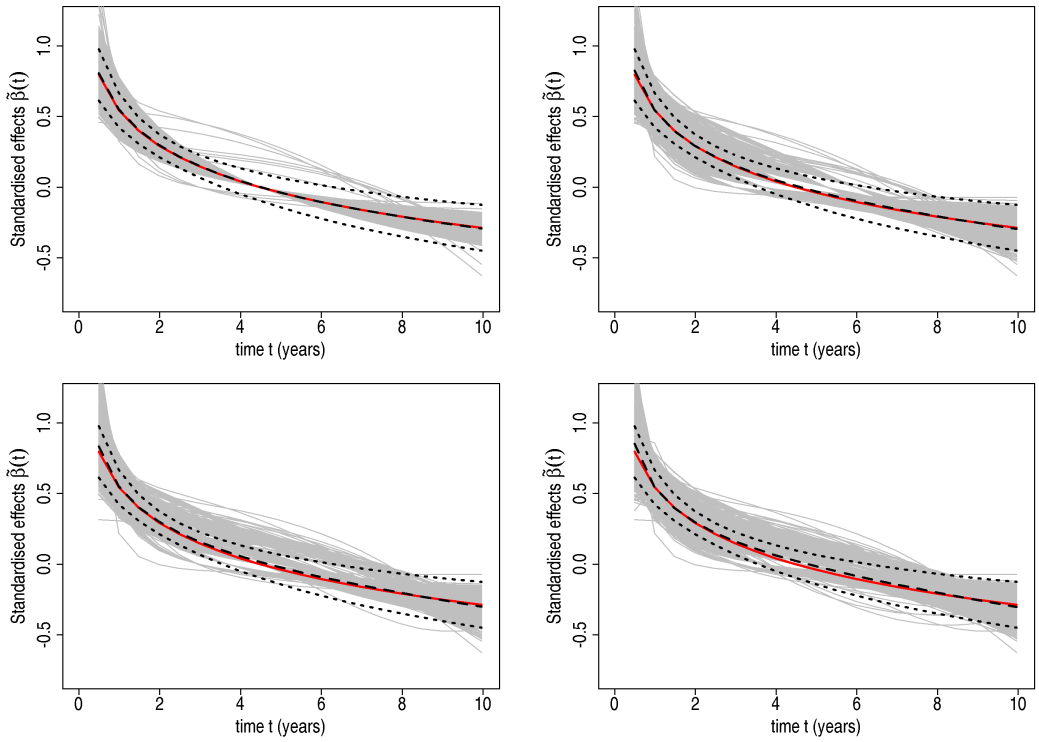


Figure 5.1: Variation of estimated effects for the prognostic index in the Rotterdam breast cancer series in bootstrap replications (—) and their bagging estimate (- -) in comparison to the reference model (—) and its analytical 95% pointwise confidence intervals (· · ·) for FPT with (left) and without (right) default transformation $\log(t)$ and significance levels $\alpha = 0.01$ (top) and $\alpha = 0.157$ (bottom).

	significance level			
	$\alpha = 0.01$		$\alpha = 0.157$	
FP type	with default	w/o default	with default	w/o default
log -	99.0	49.5	85.5	43.3
lin -	0.0	2.4	1.4	1.8
other FP1	0.5	47.2	4.1	40.6
FP2 monotonic	0.5	0.9	8.8	13.2
FP2 unimodal	0.0	0.0	0.2	1.1

Table 5.2: Relative frequencies of FP types (in %) in bootstrap samples of the prognostic index in the Rotterdam breast cancer series for FPT with and without default time transformation $\log(t)$ and different significance levels. FP types PH, log + and lin + have not been selected at all.

FP combinations		relative frequency (in %)	
with default	w/o default	$\alpha = 0.01$	$\alpha = 0.157$
-2	-2		0.1
-2 1	-2 1		2.4
-2 2	-2 2	0.2	2.7
-1	-1	0.3	1.3
-1 3	-1 3		2.0
-0.5	-0.5		1.6
0	**-2 1**		**2.1**
0	**-1**	**2.0**	**0.2**
0	**-1 3**	**0.3**	**1.0**
0	**-0.5**	**23.6**	**19.2**
0	0	49.5	43.3
0	**0.5**	**21.0**	**17.1**
0	**1**	**2.4**	**0.4**
0.5	0.5		1.0
1	1		1.4
		and 4 further combinations with frequency $<$1%	and 13 further combinations with frequency $<$1%

Table 5.3: FP powers selected by FPT with and without default time transformation in the Rotterdam breast cancer series. Relative frequencies of FP combinations (in %) for different significance levels. Bold rows indicate changes in selected FP transformations between the two versions of the FPT algorithm.

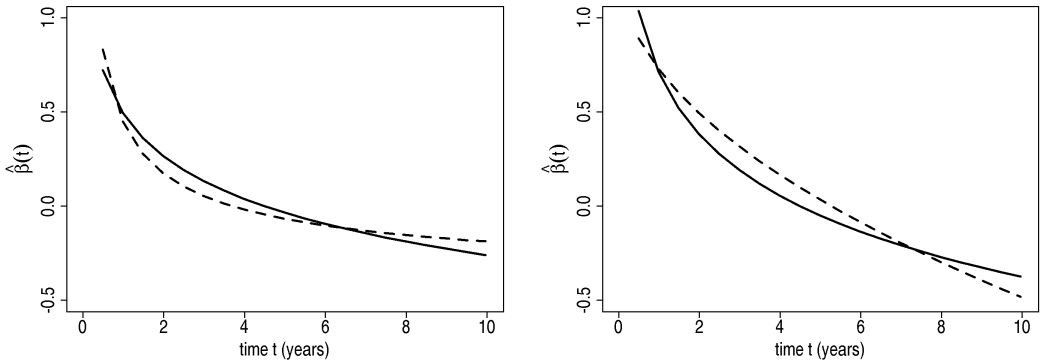


Figure 5.2: Single time-varying effects selected by FPT with (—) and without (- -) default time transformation $\log(t)$. Two exemplary bootstrap samples for which FPT without default selects powers -0.5 (left) or 0.5 (right) instead of the default for significance level $\alpha = 0.01$.

-0.5 (left panel) and 0.5 (right panel) without default transformation, respectively. The selected time-varying effects are similar in shape, but differ slightly in their curvature.

For the larger significance level $\alpha = 0.157$, Figure 5.1 shows an increased variability of

	$\mathcal{T}$ (as factors)		$\mathcal{V}$ (% of $\mathcal{T}$)		$\mathcal{D}^2$ (% of $\mathcal{T}$)	
Significance level	with default	w/o default	with default	w/o default	with default	w/o default
0.01	1.00	2.79	99.26	98.19	0.74	1.81
0.157	3.15	4.77	94.29	92.44	5.71	7.56

Table 5.4: Stability measures for the variation of estimated time-varying effects of the prognostic index in bootstrap samples of the Rotterdam breast cancer series. $\mathcal{T}$ is the total variation of bootstrap functions around the reference function, $\mathcal{V}$ the within-subset variance and $\mathcal{D}^2$ the squared deviation between bagged and reference curves. Given are the magnitude of $\mathcal{T}$ relative to the scenario $\alpha = 0.01$ with default time transformation ($\mathcal{T} = 0.0027$) and the proportion of $\mathcal{V}$ and $\mathcal{D}^2$ (in %) on the corresponding $\mathcal{T}$.

effects selected in the bootstrap samples. Again a decreasing logarithmic function is still the most frequently selected FP type, but other types, especially monotonic FP functions, also gain importance. Without the default time transformation, about half of the selected $\log$ transformations change to 'other FP1'. Although a visual comparison of effects estimated in the bootstrap samples reveals few differences (bottom row of Figure 5.1), the variety of concrete powers increases (Table 5.3).

The stability measures confirm the previous findings (Table 5.4). The variation measures of selected curves show a considerable increase by factor 3 for the larger significance level $\alpha = 0.157$ compared to $\alpha = 0.01$. While for $\alpha = 0.01$ the difference in total variation between effects estimated by FPT with and without default time-transformation is also nearly of factor 3, it reduces to 1.5 for $\alpha = 0.157$. In all four settings, $\mathcal{V}$ explains the major proportion of total variance with values of over 90%. For $\alpha = 0.157$, the proportion of $\mathcal{D}^2$ on $\mathcal{T}$ increases slightly, but even then does not exceed 8%. Hence, the total variation is due to the large variation among the bootstrap effects ($\mathcal{V}$) rather than the difference between the bagging estimate and reference function ($\mathcal{D}^2$).

Summary

These bootstrap investigations indicate reasonable stability of time-varying effects estimated by FPT. The variability increases with increasing significance level. However, especially with the default time transformation $\log(t)$, the shape of time-varying effects is similar in most replications. Even without the default, the shape of selected curves often remains similar, as most of the alternatively selected FP transformations lead to similar curve shapes.

5.2 BootstrapFPT

Model selection uncertainty

Although the bootstrap stability investigations in a univariate setting show promising results, multivariable model building is more complex and may introduce larger variability of effects. In many applications, the number of potential explanatory variables is large and so is the set of candidate models. This model space extensively enlarges if several modelling alternatives such as in-/exclusion of covariates, non-linear effects and time-varying effects are considered. Usually, variable selection methods are applied to select one final model. This neglects that several models may fit the data equally well, which may differ in terms of selected components. To overcome this problem and the model selection uncertainty due to (potential) instability of the model building process, bootstrap based strategies such as bootstrap model averaging (Buckland et al., 1997) have been proposed. Augustin et al. (2005) extended this approach by adding a bootstrap based variable screening step based on parts of a strategy suggested by Sauerbrei and Schumacher (1992).

The bootstrap variable screening step aims to eliminate variables with no or a negligible effect. Hence, it repeats the variable selection procedure in a set of bootstrap samples and eliminates all variables with a bootstrap selection frequency below a specified (small) threshold. This variable screening step is followed by a second model building step, which usually again includes variable selection (e.g. model averaging). Investigation of this approach shows that the bootstrap screening distinguishes between variables with and without an effect, without harming statistical criteria of the final model averaging predictors (Holländer et al., 2006; Buchholz et al., 2008; Sauerbrei et al., 2008).

Shepherd (2008) examines the application of bootstrap resampling when checking and correcting for violations of the PH assumption. He shows that it is most important to apply bootstrap based checking of the PH assumption, if the time-varying effect is borderline significant. For highly significant or hardly significant time-varying effects, results are similar to a single test on PH in the original data.

Bootstrap based selection of time-varying effects

Combining these concepts leads to a bootstrap based selection of time-varying effects which accounts for model selection uncertainty by bootstrapping the selection of time-varying effects to evaluate their importance. Since we do not want to eliminate components with negligible effect from our model as the variable screening procedures mentioned above, but aim to select time-varying effects with a strong influence on survival, we adapt strategy B of Sauerbrei and Schumacher (1992).

A 'strong influence' may be defined by a high selection frequency in bootstrap samples, or alternatively by a large frequency of significant time-varying effects in the out-of-bag (oob) samples. We decide in favour of the latter, as significance in the oob samples is a further evidence for stability and reliability of the selected time-varying effect.

This selection procedure, termed BootstrapFPT, contains the following steps:

1. Sample n_B bootstrap replications with replacement from the original data. Apply the FPT algorithm for selection of time-varying effects to each bootstrap sample.

2. Refit the selected model in the oob observations with FP powers of time-varying as selected in the bootstrap sample. The significance of each time-varying effect in the oob sample is tested to evaluate its importance. The test is identical to the test on significance of time-varying effects in the FPT algorithm, i.e. a likelihood ratio test on the effect of this covariate being time-constant, leaving all other covariate effects unchanged.

3. Time-varying effects are included in the final model, if they are significant in at least k% of the oob samples.

Modelling of the final effects

Another important task is to determine the final form of a time-varying effect. Model averaging techniques, which average over the complete risk score, are out of the question if the time-varying effects are to be interpreted by their own. Instead, variablewise strategies are required. We consider two possible strategies for final effect estimates:

(a) Determine the functional form of the final effect based on the most frequently selected FP transformation of those effects that are significant in the oob sample, preferring FP1 transformations if several transformations have equal frequencies. The effects of all covariates are re-estimated with these powers in the original data.

(b) Use the pointwise mean over all effects that are significantly time-varying in the oob sample, i.e. the bagged estimator (Breiman, 1996). Final time-constant effects are defined as the mean over all time-constant effects.

Application to the Rotterdam breast cancer series

To assess the potential benefits from such a bootstrap based selection of time-varying effects, we apply the BootstrapFPT approach to the Rotterdam breast cancer series (Section 4.1.1) and compare it to the standard FPT approach (both without default time transformation). The Rotterdam data is randomly split into a training set of size 1000 (498 events)

	Standard FPT	BootstrapFPT	
		Selection frequency of time-varying effects in bootstrap samples	Frequency of effects significant in oob samples
X_1	time-constant	0.22	0.00
X_{3a}	time-varying	0.68	0.10
X_{3b}	time-constant	0.32	0.00
X_{4b}	time-constant	0.18	0.00
X_{5e}^2	time-constant	0.06	0.00
$\mathbf{\log(X_6)}$	time-varying	1.00	**0.56**
X_8	time-constant	0.32	0.00
X_9	time-constant	0.24	0.00

Table 5.5: Selection frequency of time-varying effects within the BootstrapFPT approach in a subsample of the Rotterdam breast cancer series at significance level $\alpha = 0.01$.

on which the BootstrapFPT and FPT models are fitted and a test set containing 1982 observations (1020 events), which is used for evaluation of the prediction error.

The significance level for tests on time-varying effects is chosen to be $\alpha = 0.01$ for both BootstrapFPT and FPT. Significance of effects in the oob samples within the BootstrapFPT procedure is evaluated at the same significance level. A time-varying effect for a covariate is included in the final model, if it is significant in at least $k = 50\%$ of $n_B = 50$ oob samples.

The standard FPT approach selects time-varying effects for X_{3a} and $\log(X_6)$, both with power -0.5 (Table 5.5). This model differs from the model presented in Section 4.1.3, where time-varying effects for $X_{5e}^2, \log(X_6), X_8$ and X_9 are selected. Yet, as already mentioned there, the decision between X_{3a} and X_{5e}^2 in the second iteration of the forward selection procedure is very close and influences the subsequent iterations. In the subsample used for this analysis, the time-varying effect for X_{3a} is more significant. This difference also clarifies the potential benefit of bootstrap based selection of time-varying effects, which may help in case of such close decisions.

The BootstrapFPT approach following the aforementioned strategy selects a time-varying effect for $\log(X_6)$ only. Although for each of the eight covariates a time-varying effect is selected in the one or other bootstrap sample (Table 5.5), most of them are not significant in the corresponding oob sample. Only X_{3a} and $\log(X_6)$ are significant in several oob samples, but only the latter exceeds the selection level of $k = 50\%$. Hence, the time-varying effect for X_{3a} seems to be present (as it is selected in the bootstrap samples), but not important enough to be significant also in the oob samples. Consequently, evaluation of the significance of time-varying effects in the oob samples may help to verify that the time-varying effect is strong enough to be important.

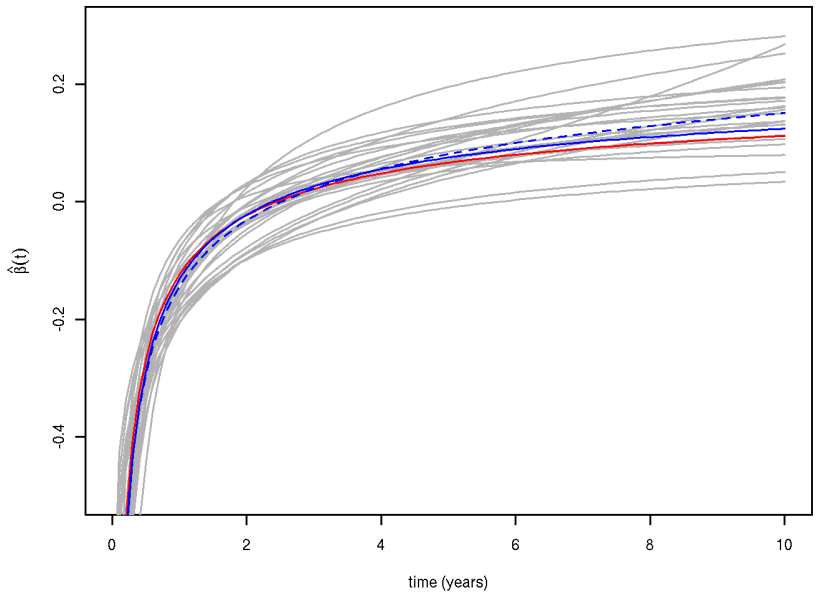


Figure 5.3: Significant time-varying effects in the oob sample (—), and effect estimated by the BootstrapFPT procedure using either the most frequent FP function (—) or the bagging estimate (- -) as well as the FPT estimate (—) for the effect of $\log(X_6)$ in a subsample of the Rotterdam breast cancer series.

The final time-varying effect for $\log(X_6)$ is determined according to both above definitions. The frequency of FP transformations that are significant in the oob sample are

FP power(s)	-0.5	0	-1	0 3
frequency of significant time-varying effects in oob samples	0.38	0.14	0.02	0.02

Hence, the most frequent FP power is -0.5. The estimated effect based on this power, as well as the bagging estimate and all significant effects of the oob samples are given in Figure 5.3. The most frequent effect based on strategy (a) is very similar to the estimate of the standard FPT approach, while the bagging estimate of strategy (b) differs slightly.

To assess the performance of the BootstrapFPT procedure, the prediction error is calculated in the test data. However, the gain in prediction performance is minor. Both the standard FPT as well as the BootstrapFPT show only a slight improvement in prediction error relative to the CoxPH model, with a marginal superiority of BootstrapFPT over FPT (Figure 5.4). The choice of the final effect in the BootstrapFPT procedure, though, has a negligible impact on the prediction performance in this data set. These results are confirmed by the integrated prediction error (IPEC) over the first 10 years (Table 5.6) which is very similar for all methods.

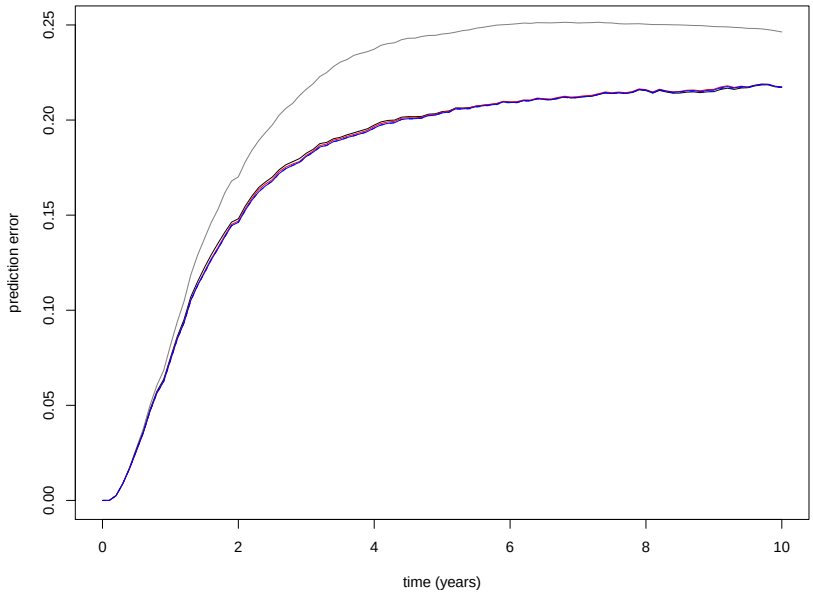


Figure 5.4: Prediction error of the BootstrapFPT procedure using either the most frequent FP function (—) or the bagging estimate (- -) and the standard FPT procedure (—) compared to the CoxPH model (—) and the Kaplan-Meier estimate (—) in a subsample of the Rotterdam breast cancer series.

			BootstrapFPT with strategy	
Kaplan-Meier estimate	CoxPH model	FPT	(a)	(b)
2.069	1.763	1.760	1.756	1.755

Table 5.6: IPEC of BootstrapFPT, standard FPT, CoxPH and the Kaplan-Meier estimate in a subsample of the Rotterdam breast cancer series.

Concluding remarks

Of course, several alterations of the proposed strategy are possible. One alternative would be to base the decision on inclusion of time-varying effects on the selection frequency in the bootstrap samples rather than on the frequency of significance in the oob samples. In that case a larger selection level, e.g. $k = 75\%$, is suitable, as the selection frequency in the bootstrap samples is far more optimistic than in the oob samples, which can easily be seen from Table 5.5. In this example, though, a selection based on bootstrap samples with $k = 75\%$ would yield virtually identical results to those based on the oob sample presented above.

The bootstrap investigation in the previous section shows that selected effects are in general rather stable, but estimates in some individual bootstrap samples may show deviant time-

varying behaviour, caused by special patterns in the data such as a few extreme values of event time. The FPT approach is, like several other methods, sensible to such extreme values. Bootstrap based approaches may reduce the influence of extreme event times on the selection and modelling of time-varying effects. A selection and modelling strategy as described above seems to be a promising alternative to a test procedure based on a single data set, but at the cost of increased computational effort. However, as FPT seems to give rather stable results in large data sets, the benefits of bootstrap based model selection are expected to be most pronounced in smaller samples, where computational complexity is less severe. Yet, for reliable statements about the performance of this approach, more elaborate investigations are required.

Chapter 6

Simulation study

The main aim of this simulation study is the assessment of properties of the FPT approach. The selection procedure is evaluated in terms of type I and II error. However, the detection of time-varying effects is only a small part of the story. Appropriate modelling of the time-varying effect is a more important task. Hence, a special emphasis is placed on evaluating the shape of selected effects. This includes extending the definition of a type II error according to certain qualitative criteria. Estimated effects are further compared graphically to the true effects and their similarity is quantified by ABCtime, a new measure developed to quantify the distance between the two curves. The prediction performance of fitted models is assessed by prediction error curves.

To put the results into context, the performance of FPT is compared to the standard CoxPH model. For most of the four further approaches applied in Section 4, though, simulation studies are beyond the scope of this thesis, as the approaches or the corresponding programs require substantial modifications and improvements before simulation studies are feasible. The Empirical Bayes approach and the Dynamic Cox model are disregarded due to technical reasons. While the former needs manual selection of time-varying effects and is extremely time-consuming, the latter is implemented in an old `S-Plus` version and is not applicable to larger data sets unless the survival times are categorised. The Reduced Rank approach, on the contrary, is omitted for theoretical reasons. The missing selection of time-varying effects results in far too complex models and limits practical applicability. Hence, the Semiparametric Extended Cox model is the only approach, which seems to be suitable for our simulation study. However, application of this approach results in considerable problems, which lead to its exclusion from the simulation study. Consequently, the simulation study focuses on investigation of the properties of the FPT approach with the CoxPH model for comparison.

The structure of this chapter is as follows. The simulation design is introduced in Section 6.1, followed by a description of the data generation algorithm (Section 6.2), a brief discussion

of potential problems with extreme survival times (Section 6.3) and a summary of the assessment criteria for estimated effects (Section 6.4). The results of the simulation study for univariate and multivariable settings are presented in Sections 6.5 and 6.6, respectively. The simulation study reveals potential convergence problems for some FP combinations, which are briefly discussed in Section 6.7. Thereafter, we give details on the difficulties with the Semiparametric Extended Cox model (Section 6.8) and end the chapter with a short summary of all findings (Section 6.9).

6.1 Simulation design

6.1.1 Univariate settings

As in many other simulation studies, investigated effects and simulation parameters are rather arbitrary. However, our aim is to design settings that are plausible in medical applications. More technically, effect functions should be chosen, which do not favour the FPT approach over other approaches that are not based on FPs. Therefore, we use functions that are included in the class of FPs as well as potentially realistic functions whose shape is not exactly included in the FP class, i.e. FPs with slightly modified powers, and functional forms not included in the class of FPs, which thus cannot be fully described by FPT. By using such a mixture of functions we intend to have a good balance between enabling FPT to find an effect that is close to the correct function, without giving it too much advantage over other approaches. Furthermore, underlying functions will in reality rarely be an exact member of the class of FPs and this simulation study can therefore investigate performance in real life settings.
The development of an informative simulation design, though, is a difficult task. Especially with time-varying effects, inter-dependencies between parameters can be large. For example, the time scale and censoring patterns have a strong influence on whether effects of a certain size and shape are detectable. Vice versa, the chosen effects influence the distribution of generated survival times. In addition, generated times are influenced by the baseline hazard and the values of variables. Hence, parameter combinations need to be developed that ensure sufficiently large event times to enable detection of time-varying patterns, in combination with effects that are large enough to be detectable for the simulated survival times.

(Time-varying) effects

The aforementioned considerations result in the following ten effect functions, which are illustrated in Figure 6.1.

(C) Constant effect: $\beta(t) = 0.8$

(Ls) Strong linearly decreasing effect: $\beta(t) = 2 - 0.18t$

(Lw) Weak linearly decreasing effect: $\beta(t) = 1 - 0.09t$

(Ds) Strong non-linearly decreasing effect: $\beta(t) = 0.32 + \frac{1.42}{\exp(t)} - 0.02t^{0.7}$

(Dw) Weak non-linearly decreasing effect: $\beta(t) = 0.2 + \frac{0.8}{\exp(0.6t)}$

(Is) Early strong increasing effect: $\beta(t) = 0.1 + 0.8t^{0.3}$

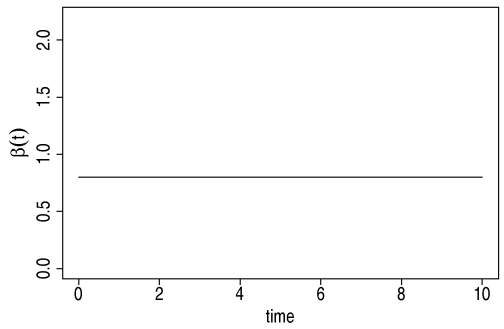
(Iw) Early weak increasing effect: $\beta(t) = \frac{6}{1+\exp(3-0.2t)} - 0.2$

(S) Step-like (sigmoid) effect: $\beta(t) = \frac{2}{1+\exp(6-2t)}$

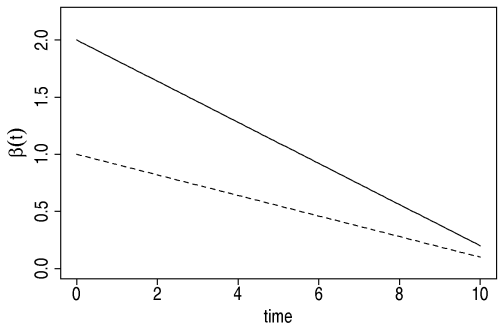
(Bs) Bathtub effect with strong end: $\beta(t) = \frac{4}{1+\exp(1.2(t+0.5))} + \frac{4}{3(1.1+\exp(10-t))} + 0.02$

(Bw) Bathtub effect with weak end: $\beta(t) = \frac{1.5}{1+\exp(2.5t)} + \frac{4}{1+\exp(1.5(10-t))} + 0.02$

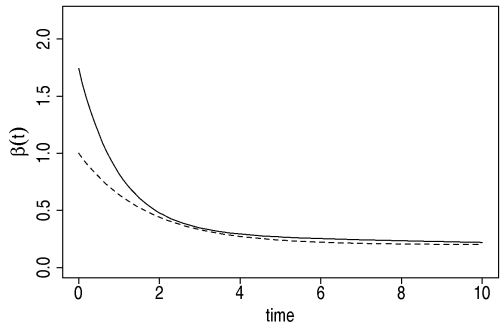
These functions represent effects of different shape which are feasible in medical applications. The motivation for the linear decreasing effects (Ls) and (Lw) are prognostic factors that have a strong effect initially which diminishes steadily over time. Such an effect has, for example, been detected for the period of diagnosis in gastric cancer studies (Binquet et al., 2009). (Ds) and (Dw) reflect effects that diminish strongly within a short time, as the effects of tumour size in breast cancer, which is known to have a strong prognostic effect shortly after surgery but after some years hardly influences prognosis anymore (Sauerbrei et al., 2007). Other examples include the effects of prothrombin time in primary biliary cirrhosis (Abrahamowicz et al., 1996) and the Karnofsky performance status in ovarian cancer studies (Verweij and van Houwelingen, 1995). Both classes, linear and non-linear decreasing effects, are examined in a strong and weak version.
Increasing effects such as (Is) and (Iw) are imaginable, for example, for two therapies, where therapy A looses its effect while therapy B gains in importance. (Is) and (Iw) are of completely different shape. While effect (Iw) increases slowly over time, (Is) shows a strong increase early in time, as also detected for the effect of diabetes on mortality after coronary artery bypass graft surgery (Gao et al., 2006).
Prognostic factors or therapies that need a certain time to show a visible effect are accounted for by effect (S) which has a sigmoid shape, increasing with some delay. This effect changes rapidly over a limited time interval and is useful for assessing the performance of FPT in a situation, where the true effect is too complex to be fully recovered by FPs of maximum degree 2. The two bathtub-shaped effects (Bs) and (Bw) describe another complex pattern. Such effects are found, for example, when following a population from birth. After an initial medium sized effect due to infant diseases, the death rate stabilises and increases again



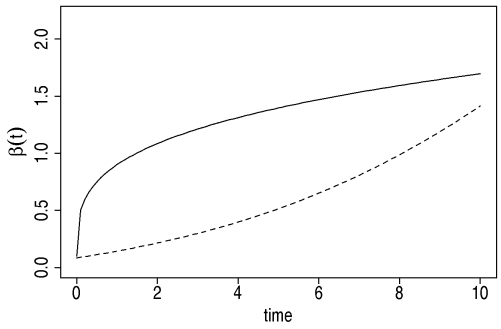
(C) Constant effect



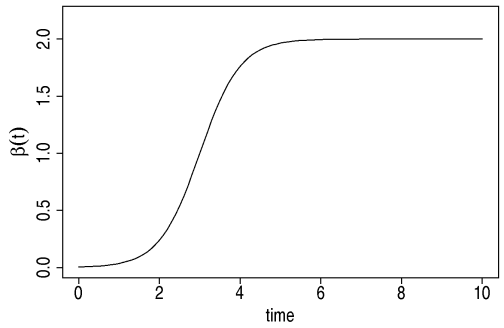
(Ls, Lw) Strong (—) and weak (**- -**) linear effects



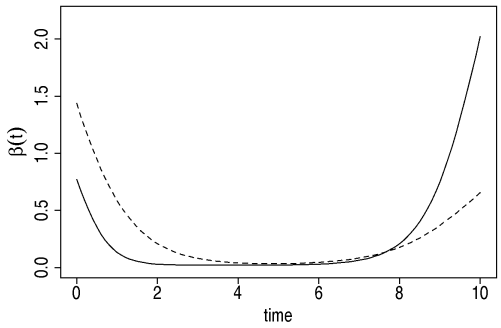
(Ds, Dw) Strong (—) and weak (**- -**) decreasing effects



(Is, Iw) Early strong (—) and weak (**- -**) increasing effect



(S) Step-like (sigmoid) effect



(Bs, Bw) Bathtub-shaped effect with strong (—) and weak (**- -**) end

Figure 6.1: Time-varying effects for the simulation study. For some function types a weak and a strong effect is considered.

due to natural ageing. Another possible explanation is a setting with composed endpoint, where overall survival (or natural death) begins to dominate in the course of time. An effect of this type has, for example, been detected for age on overall survival in colon cancer patients (Quantin et al., 1999).

Variables, sample sizes and simulation parameters

For the sake of completeness, all ten effects are investigated in combination with a binary ($P(X = 1) = 0.5$) and a standard normal distributed variable. Furthermore, random censoring is applied. We consider a moderate sample size of $n = 1000$ with light censoring to be an important realistic setting. Therefore, all ten effects for both binary and normal variable are investigated in this setting. Depending on the type I and II error in this setting, investigations are extended to high censoring or to no censoring and larger sample size ($n = 3000$). The exact censoring rate depends on the baseline hazard and the specific effects. Run time considerations lead to the decision that all scenarios with sample size $n = 1000$ are run in 1000 replications, while sample size $n = 3000$ is repeated only 100 times.

To enable the detection of time-varying effects, a sufficiently long follow-up is required. In breast-cancer applications like our main example, the Rotterdam breast cancer series, 10 to 15 years seem to be a reasonable follow-up time. In combination with the size of time-varying effects used for this simulation, we deem a similar time span to be quite reasonable. The event and censoring times T and C as well as the censoring rate are controlled by the baseline hazards λ_E and λ_C, respectively. In detail we choose

- $\lambda_E = 0.2$ and $\lambda_C = 0.1$ for light censoring,
- $\lambda_E = 0.1$ and $\lambda_C = 0.2$ for heavy censoring and
- $\lambda_E = 0.3$ for no censoring,

which give survival times mostly lying in [0, 10]. The median survival times, i.e. the time at which half of the subjects have experienced an event, range between 0.8 and 6.7, depending on the effect, censoring pattern and distribution of the variable. Similarly, the median follow-up times according to the reverse Kaplan-Meier method (Altman et al., 1995) vary between approximately 3.5 and 7.

For most scenarios, the simulated uncensored event times tend to be larger for normal X than for binary variable. In addition, an increased number of extremely small times is observed, which coincide with large values of X. The censoring rates differ between 20% and 35% for light censoring and about 60% for heavy censoring, depending on the specific scenario. Details on generated survival times, event and censoring distributions can be found in Appendix C.1.

X	Sample size	Censoring	Average population effects (Ls)	(Lw)	(Ds)	(Dw)	(Is)	(Iw)	(S)	(Bs)	(Bw)
binary	1000	heavy	-	-	0.76	-	0.95	-	1.24	0.68	0.54
	1000	light	1.73	0.70	0.62	0.28	1.04	0.40	0.82	0.28	0.46
	1000	no	1.84	0.80	0.43	0.37	1.22	0.28	-	-	-
normal	1000	heavy	0.91	0.44	0.70	0.19	0.79	-	0.48	0.01	0.44
	1000	light	1.18	0.61	0.59	0.29	0.79	0.07	0.67	0.06	0.45
	1000	no	-	-	0.43	-	0.77	0.11	-	-	-

Table 6.1: Average population effects for scenarios with time-varying effects in the simulation study

The effect estimates presented in this chapter are not standardised. Since the Cox model has no intercept term, effect estimates may possibly be subject to a shift on the y-axis. However, for larger sample sizes as used in our simulation study this should have hardly any influence on estimates. The simulation results show no peculiarities in this respect. Furthermore, our investigation reflects the analysis of real studies.

Effect sizes

As the effect size of time-varying effects is difficult to judge, we calculate the average population effects (Xu and O'Quigley, 2000), which coincide with the CoxPH estimates in a large data set without censoring. These average effects (Table 6.1) give an impression of the strength of the main effects for the different parameter settings. The variation of average population effects is broad. Not surprising, the average population effects in scenarios with decreasing effects are relatively large, but considerably smaller for the increasing effects. For the binary variable, effect (Ls) is the by far strongest effect with an average population effect of about 1.8, while the others range between 0.28 and 1.24, depending on the effect and parameter settings. With standard normal variable, the average population effects for the same parameter settings are in general considerably smaller. For scenarios (Iw) and (Bs) they are even close to zero.
Due to this systematically different effect sizes, a joined consideration of settings with binary and standard normal variable is not sensible. Hence, they are considered separately in the sequel.

6.1.2 Multivariable settings

To evaluate the multivariable model building strategy of FPT, data are generated from a multivariable setting with five binary variables with $P(X = 1) = 0.5$ and sample size 1000

according to

$$\lambda(t|X) = \lambda_0 \exp\left(\beta_1(t)X_1 + \beta_2(t)X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5\right)$$

where $\beta_1(t)$ and $\beta_2(t)$ are identical to effects (Ds) and (Is) in the univariate setting, which are deemed most relevant in practical applications, and β_3 to β_5 are constant:

$$\beta_1(t) = 0.32 + \frac{1.42}{\exp(t)} - 0.02t^{0.7} \qquad \beta_2(t) = 0.1 + 0.8t^{0.3}$$

$$\beta_3 = 0.3 \qquad \beta_4 = 0.5 \qquad \beta_5 = 0.7$$

The baseline hazards are chosen to be $\lambda_E = 0.1$ without censoring, $\lambda_E = 0.06$ and $\lambda_C = 0.1$ for light censoring and $\lambda_E = 0.03$ and $\lambda_C = 0.15$ for heavy censoring. This corresponds to on average 0%, 27% and 50% of censored observations, respectively. The median survival times of these settings vary between 1 and 5 years, depending on the setting. Details on median survival and follow-up times, as well as survival and censoring distributions can be found in Appendix C.2.
Furthermore, we investigate these settings with uncorrelated variables as well as a correlation between X_1 and X_4 of $\rho_{X_1,X_4} = 0.5$. The correlation is introduced following Lunn and Davies (1998).
Due to the increased runtime compared to the univariate settings, investigations are limited to 100 simulation runs for each of the six settings.

6.2 Simulating survival data with time-varying effects

The literature on generation of survival data with time-varying effects is scarce. MacKenzie and Abrahamowicz (2002) propose a permutational algorithm for generating survival data with time-varying effects. The algorithm generates survival times, including censoring, simultaneously controlling for the marginal distributions of covariates X_i, censoring and event times T as well as the (time-varying) hazard ratio. The main idea of this approach is to match random pairs of T and independently generated X_i according to a permutational probability law. The corresponding permutational probability is identical to the partial likelihood formula. The method is shown to be asymptotically exact. It is very general and also applicable to time-dependent covariates (Sylvestre and Abrahamowicz, 2008), but the permutational algorithm is computationally expensive and also unnecessarily complex, if the interest lies in time-varying effects only.
Berger et al. (2003) generate data from a logistic setting with specified probabilities of failure and censoring at each time point. This method, though, gives more or less discrete survival times. Recently, He et al. (2010) proposed another approach, which simulates data using

a specified time grid, drawing survival times from the exponential distribution in each interval. Their approach, though, requires approximation of continuous time-varying effects by piecewise constant functions.

We use a simpler and more intuitive approach based on inversion of the cumulative hazard function. Bender et al. (2005) describe this method for simulating survival data under the PH assumption. In general, with time-varying effects the integral and the inverse distribution function cannot be solved analytically any more. Hence, they are substituted by numeric integration and inversion. The same concept is applied by Leemis et al. (1990) for generation of survival data with time-dependent covariates and by Beyersmann et al. (2009) for competing risks data. Hofner et al. (2010) use a similar procedure for generating data with time-varying effects.
The extended Cox model for non-linear and time-varying effects is

$$\lambda(t|X) = \lambda_0(t) \exp\left(\sum_{i=1}^{q} f_i(X_i)\beta_i(t)\right),$$

where $\lambda_0(t)$ is the baseline hazard, X_i are the covariates which may be transformed by functions $f_i(.)$ and $\beta_i(t)$ are the time-varying (or time-constant) effects.
The survival function of this model is

$$S(t|X) = \exp(-\int_0^t \lambda(u|X)du) = \exp(-\int_0^t \lambda_0(u)\exp(\beta(u)f(X))du) = 1 - F(t|X)$$

and the distribution function

$$F(t|X) = 1 - \exp(-\int_0^t \lambda_0(u)\exp(\beta(u)f(X))du) = 1 - \exp(-\Lambda(t)).$$

Let $T \sim F$, then $U = F(T) \sim U(0,1)$. Therewith, survival times for the above hazard can be generated as $T = F^{-1}(U)$:

$$\begin{aligned}
& U = F(T) \\
\Leftrightarrow \quad & U = 1 - \exp(-\Lambda(T)) \\
\Leftrightarrow \quad & \exp(-\Lambda(T)) = 1 - U \\
\Leftrightarrow \quad & -\Lambda(T) = \ln(1-U) \\
\Leftrightarrow \quad & T = \Lambda^{-1}(-\ln(1-U)) \\
\Leftrightarrow \quad & T = F^{-1}(U)
\end{aligned}$$

where $F^{-1}(U)$ is calculated via numerical inversion, i.e. by solving $U - F(T) = 0$, i.e. $\Lambda(T) + \ln(1-U) = 0$. This algorithm allows exact generation of survival times with time-varying

effects. The sole condition on the $\beta_i(t)$ is that the integral of the resulting hazard must be finite.

6.3 Problems with extreme survival times

Complex approaches for modelling time-varying effects such as FP or spline based approaches may be sensible to extreme survival times. Extremely small or large survival times may cause erroneous selection of a time-varying effect (type I error). For both FP and spline transformations, this problem is already known from modelling functional forms of covariates. Non-linear transformations $f_i(.)$ of extreme values of covariate X_i may magnify these values, and thus their leverage, considerably. This can have a strong influence on the partial likelihood

$$PL(\beta) = \prod_{j=1}^{n_e} \frac{\prod_{k=1}^{d_j} \exp\left(\sum_{i=1}^{q} \boldsymbol{f_i(X_{ji})}\beta_i\right)}{\{\sum_{l \in R(t_{(j)})} \exp\left(\sum_{i=1}^{q} \boldsymbol{f_i(X_{li})}\beta_i\right)\}^{d_j}},$$

but directly affects this specific observation only, i.e. $l = j$. For time-varying effects, extreme survival times have a direct influence on all observations l at risk at $t_{(j)}$, the event time of the jth individual:

$$PL(\beta) = \prod_{j=1}^{n_e} \frac{\prod_{k=1}^{d_j} \exp\left(\sum_{i=1}^{q} X_{ji}\boldsymbol{\beta_i(t_{(j)})}\right)}{\{\sum_{l \in R(t_{(j)})} \exp\left(\sum_{i=1}^{q} X_{li}\boldsymbol{\beta_i(t_{(j)})}\right)\}^{d_j}},$$

For this reason, the influence of extreme values is even larger in modelling time-varying effects than in modelling non-linear effects. Yet, while with non-linear transformations of covariates, problems due to extreme values likewise arise on both edges, with time-varying effects problems are mainly restricted to extremely small times, i.e. to the left edge.

Such extreme values may also cause convergence problems. For modelling of functional forms of covariates, Royston and Sauerbrei (2007) propose a robustness transformation to reduce the leverage of extreme values. This transformation maps extreme values smoothly to asymptotes, while transforming the bulk of points linearly. This concept can also be transferred to survival times. The disadvantage of this transformation is that time-varying effects are estimated on a transformed time scale. Although re-transformation of effects is possible, we believe that convergence problems may be informative. The set of fractional polynomial powers is somewhat arbitrary. Although it turned out to be useful for the investigation of non-linearity in FP transformations of covariates, some transformations may be less appropriate for transformation of survival times. Consequently, an investigation of convergence problems of FPs may give further insight into the practical applicability of the standard set of powers.

Since the simulation design assumes an exponentially distributed baseline hazard, a large

number of extremely small survival times may be generated which can possibly cause convergence problems. However, in most practical applications a very large amount of extremely small times is unlikely to occur because of in- and exclusion criteria of a study. This is mimicked in our simulation study by the exclusion of survival times below a certain threshold. We decide to drop all times smaller than $\frac{1}{52}$ and re-simulate the times for these subjects. The corresponding X values remain unchanged. This approach reduces convergence problems that are typical for simulation studies, i.e. introduced by a disproportionately large number of very small survival times, to a tolerable amount without suppressing them completely. Thus, it enables conclusions about possibly inappropriate FP powers without distorting simulation results.
Furthermore, as no realistic effect would increase to infinity and to avoid artefacts for extremely large times, artificial asymptotes are introduced to effects (Iw), (Bs) and (Bw) (Figure D.1).

6.4 Assessment criteria

To investigate the properties of the FPT approach, we consider the test procedure for selection of time-varying effects as well as the estimated effects themselves and the prediction performance of the complete models. The test procedure is evaluated by means of type I and II error. For tests on time-varying effects, a type I error is defined as rejection of the null hypothesis of a time-constant effect, when it is actually true. Defining a type II error, though, is more complex. The simplest definition is failure of rejection of the null hypothesis (PH), when in fact a time-varying effect is present. This definition is a mere evaluation of the test procedure on the presence of a time-varying effect and is very rough. Whenever a time-varying effect is identified, even if the shape is severely different from the true shape, it would be considered to be correct. However, recovery of the nature of the time-varying effect is important as well, as a wrong shape of the selected time-varying effect would lead to incorrect conclusions in practical applications. Hence, a more suitable definition of type II error is required, which considers the shape of selected effects as well. To account for this task, we additionally define a qualitative type II error based on certain qualitative criteria concerning monotonicity, slope, extrema and/or inflection points. The basic idea is in analogy to Binder et al. (2010), who propose such criteria to compare functions in the context of regression models. The qualitative criteria for the nine time-varying effects are summarised in Table 6.2. This qualitative type II error simultaneously assesses the estimated effects, as it specifies whether the estimated effect is of similar shape than the true effect. Furthermore, shadow plots are used to represent the variation of estimated effects. These plots graphically contrast the estimated (time-varying) effects of all simulation runs to the true effect, together with their pointwise mean and 95% pointwise empirical confidence intervals. The

Effect(s)	Slope	Extrema	Inflection point
Ls	negative, constant (SD $<$ 0.0018)	-	-
Lw	negative, constant (SD $<$ 0.0009)	-	-
Ds	negative, strictly increasing	-	-
Dw	negative, strictly increasing	-	-
Is	positive, strictly decreasing	-	-
Iw	positive, increasing	-	-
S	positive, increasing for $<$ 2.5 and decreasing for $>$ 3.5	-	exactly one between 2.5 and 3.5
Bs	increasing, negative for $<$ 3, positive for $>$ 7	exactly one between 3 and 7	-
Bw	increasing, negative for $<$ 3, positive for $>$ 7	exactly one between 3 and 7	-

Table 6.2: Criteria for a qualitative type II error in the simulation study.

pointwise mean is calculated on truncated effects, i.e. the upper and lower extremes at each time point are set to the 1% and 99% quantile. In settings with $n = 3000$, where only 100 simulation runs are executed, these quantiles are equivalent to the minimum and maximum. Hence, in these settings, the 2% and 98% quantiles are used, which corresponds to truncating the minimum and maximum value at each time point. This helps to avoid a domination of the mean by artefacts that occur in single simulation runs only. Yet, if artefacts occur more frequently, i.e. in more than 1% of simulation runs, they are still reflected by the pointwise mean.

The distance of estimated effects to the true effect is evaluated using pABCtime as introduced in Section 3.2. pABCtime quantifies the weighted distance of estimated effects to the true effect and is given in percent relative to the weighted area under the true effect. Consequently, when comparing several approaches, the approach with the smallest value of pABCtime is closest to the true effect and hence performs best. pABCtime is calculated over the time interval $[\frac{1}{52}, 10]$. The lower bound of $\frac{1}{52}$ avoids the distortion of pABCtime by artefacts at extremely small times and, since the same value has been used as the minimum acceptable event time in the data generation algorithm, restricts comparison of effects to the range of times on which they are estimated. The variation of pABCtime over simulation runs is presented by boxplots, where boxes range from the 25% to the 75% quantile and the whiskers extend to 1.5 times the interquartile range. To assess the improvement due to incorporation of time-varying effects, we compare the (possibly time-varying) effects estimated by FPT to estimates of a CoxPH model (mean CoxPH estimates per scenario are given in Table D.1 in the Appendix).

Measure	Description
Type I error	The true effect is constant, but a time-varying effect is selected.
Crude type II error	The true effect is time-varying, but a time constant effect is selected. Each time-varying effect, irrespective of its shape, is considered to be 'correct'.
Qualitative type II error	The true effect is time-varying, the selected effect is either time-constant or time-varying, but of a different shape than the true effect (according to the qualitative criteria in Table 6.2).
Shadow plots	Graphical representation of all estimated effects, their pointwise mean and 95% empirical confidence intervals.
pABCtime	Quantitative measure to assess the difference to the true effect as introduced in Chapter 3.2.
PEC / mPEC	Prediction error curves (as pointwise mean over all simulation runs) as introduced in Chapter 3.1.
IPEC / dIPEC	Integrated prediction error (as difference to the Kaplan-Meier estimate) as introduced in Chapter 3.1.

Table 6.3: Criteria for assessment of FPT in the simulation study.

Furthermore, the predictive ability of estimated models is assessed in terms of the 'true' prediction error in a reference data set containing 5000 uncensored observations. We evaluate the mean prediction error curves (mPEC) as the pointwise mean over all simulation runs, as well as the integrated prediction error (IPEC) based on the Riemann integral over the interval [0, 10]. The IPEC is presented as the percentaged difference to the Kaplan-Meier estimate, denoted by dIPEC, and the variation among simulation runs is visualised by boxplots. Negative values of dIPEC represent an improvement relative to the Kaplan-Meier estimate and positive values a decline. The difference to the Kaplan-Meier estimate is chosen in order to evaluate the gain in prediction performance caused by inclusion of covariate information, taking the data at hand into account. To evaluate the additional gain due to incorporating time-varying effects, the dIPEC of FPT is contrasted with that of the CoxPH model. As the computation of prediction error curves for the FPT approach is very time consuming, we reduce computational efforts by coarsening the estimate of the cumulative baseline hazard $\hat{H}(t)$ to 5% quantiles of distinct event times. This leads to a considerable reduction in the dimension of $\hat{S}(t|X)$ and speeds up computations.
An overview of all of these assessment criteria is given in Table 6.3.

6.5 Properties of FPT in univariate settings

Estimation of FPT models is accomplished in `Stata 10` (StataCorp, 2007). Time-varying effects are selected at a significance level of $\alpha = 0.01$ with default time transformation $\log(t)$. We start with examining the main scenario with sample size 1000 and light censoring. If the crude type II error is worse than 20%, the more advantageous settings with larger effective sample size ($n = 1000$ without censoring and $n = 3000$ with light censoring) are additionally examined. For better type II error ($< 20\%$), the investigation is extended to the setting with heavy censoring. Beyond this, the influence of the sample size is investigated in more detail for effects (C), (Ds) and (Is), which are deemed most important in practical applications. For these effects, all four aforementioned settings are examined, as well as the additional sample sizes 500 and 250 with light censoring, in order to assess small sample properties. As mentioned before, the main effects are of systematically different sizes in settings with binary and standard normal variable. Hence, the simulation study is split in two parts and results for the binary and the standard normal variable are presented separately.

6.5.1 Part 1: Binary variable

In the sequel, we describe the results for binary covariate, subdivided into type I and II error, estimated effects, their difference to the true effect and the prediction performance.

Type I error

The type I error for the constant effect (C) is marginally inflated with up to 1.4% for not too small sample size (≥ 500). Only for $n = 250$ it increases to 3.7% (Table 6.4). Hence, for not too small sample size, FPT holds the type I error well. Most of the erroneously selected time-varying effects show a strong departure from PH mainly near zero, as can be seen in Figure 6.2. For very small samples, this is also reflected by the pointwise mean over all estimated effects. With increasing sample size this pattern vanishes soon and the mean effect is virtually identical to the true effect. If constant effects are selected, estimates are unbiased.

Type II error and modelling of time-varying effects

In the analysis of the different time-varying effects, the performance of FPT varies broadly among different scenarios and parameter settings. For the linear decreasing effects (Ls) and (Lw), for example, the power is extremely poor. The already large (crude) type II error for moderate sample size 1000 even increases when considering the shape of selected effects, indicating that incorrect functional forms are selected in more than 95% of simulation runs.

Sample size	Censoring	Effects									
		(C)	(Ls)	(Lw)	(Ds)	(Dw)	(Is)	(Iw)	(S)	(Bs)	(Bw)
		Type I error	Crude type II error								
250	light	3.7	-	-	58.2	-	89.1	-	-	-	-
500	light	1.0	-	-	18.6	-	82.5	-	-	-	-
1000	heavy	1.3	-	-	13.1	-	62.5	-	0.0	0.5	15.8
1000	**light**	**1.4**	**74.5**	**73.5**	**0.8**	**54.1**	**48.7**	**27.9**	**0.0**	**1.3**	**1.1**
1000	no	1.4	87.0	67.3	0.0	8.1	41.4	27.7	-	-	-
3000	light	0.0	20.0	21.0	0.0	0.0	1.0	0.0	-	-	-
			Qualitative type II error								
250	light		-	-	60.2	-	94.1	-	-	-	-
500	light		-	-	19.6	-	84.9	-	-	-	-
1000	heavy		-	-	15.4	-	65.1	-	100.0	74.7	74.5
1000	light		95.8	97.1	3.0	55.4	50.3	71.8	100.0	56.7	56.5
1000	no		97.6	95.3	3.8	8.6	42.7	71.5	-	-	-
3000	light		87.0	81.0	13.0	1.0	2.0	13.0	-	-	-

Table 6.4: Type I error of scenario (C) and crude and qualitative type II error of scenarios (Ls) - (Bw) (in %) for significance level $\alpha = 0.01$ with binary variable in a simulation study of univariate scenarios.

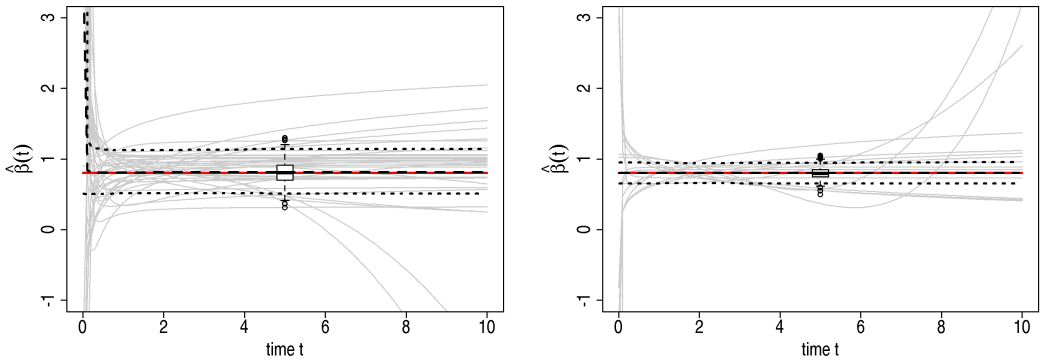


Figure 6.2: True (—) effect, estimated time-varying (—) and time-constant (boxplot) effects, pointwise mean (- -) and 95% empirical confidence intervals (· · ·) for sample sizes 250 (left) and 1000 (right) with light censoring and binary X with constant effect (C).

Although the crude type II error for the larger sample size 3000 is acceptable, the qualitative type II error deteriorates considerably. A closer look at the shape of selected effects shows that most time-varying effects are $\log$-like (FPs with powers -0.5, 0 or 0.5) rather than linear, which explains the large qualitative type II error. These functions, as well as the many time-constant effects dominate the pointwise mean, which differs largely from the true effect (Figure 6.3). Striking, however, is the increased crude type II error for (Ls) without censoring compared to light censoring. In general, the power is expected to improve with increasing effective sample size. In this setting, the survival probability is smaller for large times, com-

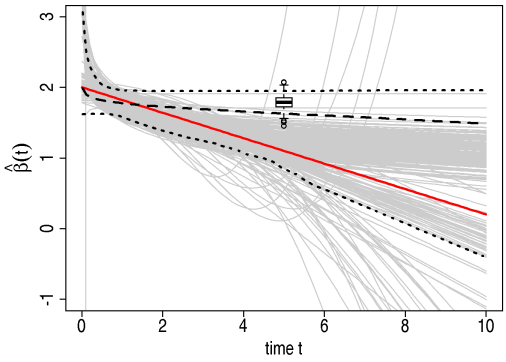


Figure 6.3: True (—) effect, estimated time-varying (—) and time-constant (boxplot) effects, pointwise mean over all estimated effects (- -) and 95% pointwise empirical confidence intervals (· · ·) for binary variable with sample size 1000 and light censoring for effect (Ls).

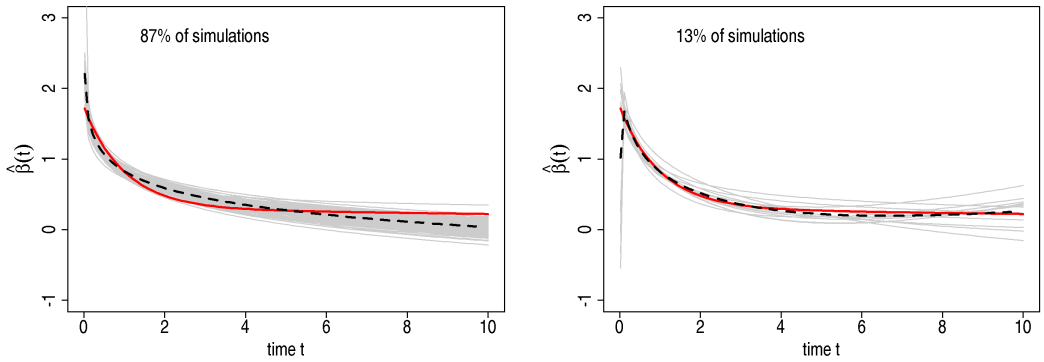
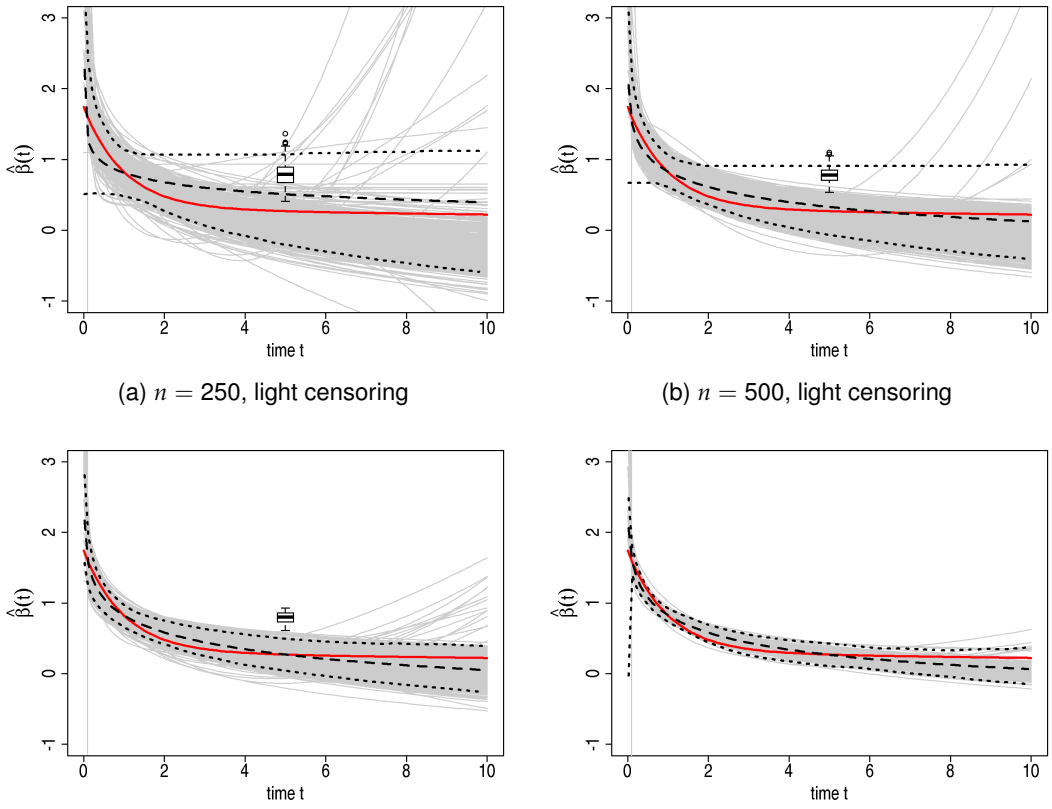


Figure 6.4: Correct (left) and incorrect (right) functional forms according to the qualitative type II error for effect (Ds) with binary variable, sample size 3000 and light censoring. Shown are the true effect (—), estimated effects (—) and pointwise mean (- -).

---

pared to light censoring, caused by different baseline hazards. Although this does not seem to have much impact in other scenarios, it may influence the detection of time-varying effects in this scenario, where the power is already extremely poor in the main setting. Indeed, in a small investigation of 100 simulation runs with baseline hazard λ_E identical to the setting with light censoring, the crude type II error drops to 62%.

The strong non-linear decreasing effect (Ds) is considered as practically very important and is therefore investigated in all parameter settings, including small sample sizes. The crude type II error is extremely good to moderate for not too small sample size ($\geq$ 500) and the qualitative type II error is only slightly larger. The only exception is observed for the larger sample size 3000, where the crude type II error of 0% increases to moderate 13%, when considering the qualitative criteria. Most of the effects that do not meet the qualitative criteria in this setting show either artefacts for extremely small times or a unimodal function



(a) $n = 250$, light censoring

(b) $n = 500$, light censoring

(c) $n = 1000$, light censoring

(d) $n = 3000$, light censoring

Figure 6.5: True (—) effect, estimated time-varying (—) and time-constant (boxplot) effects, pointwise mean (- -) and 95% pointwise empirical confidence intervals (· · ·) for effect (Ds) and binary variable.

---

with a slight increase for large times (Figure 6.4) which may be caused by some extremely large event times. In general, the variation of selected effects is relatively broad for smaller effective sample size but decreases for moderate and large sample sizes (Figure 6.5). The pointwise mean is reasonably close to the true effect in all parameter settings. For the weaker effect (Dw), the power in the main parameter setting n =1000 with light censoring is poor. For larger effective sample size, though, the power considerably improves and results are analogue to the stronger pendant (Ds).

The performance of the FPT algorithm for effect (Is) strongly depends on the sample size. It decreases from over 80% to merely 1% for increasing sample size. The qualitative type II error is only marginally larger in all settings. Similarly, the variability of estimated effects considerably decreases with increasing sample size (see Figure D.2 in the Appendix) and the pointwise mean approaches the true effect. The power for the gently inclined effect (Iw) is moderately poor (30%) for sample size 1000, but the qualitative type II error is with about

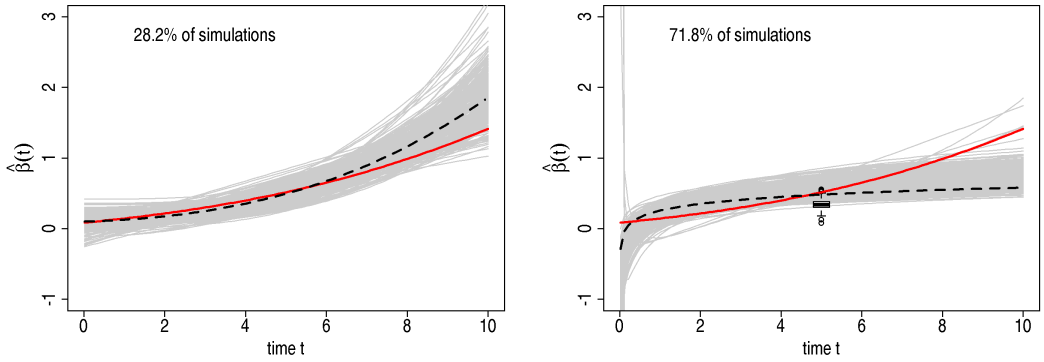
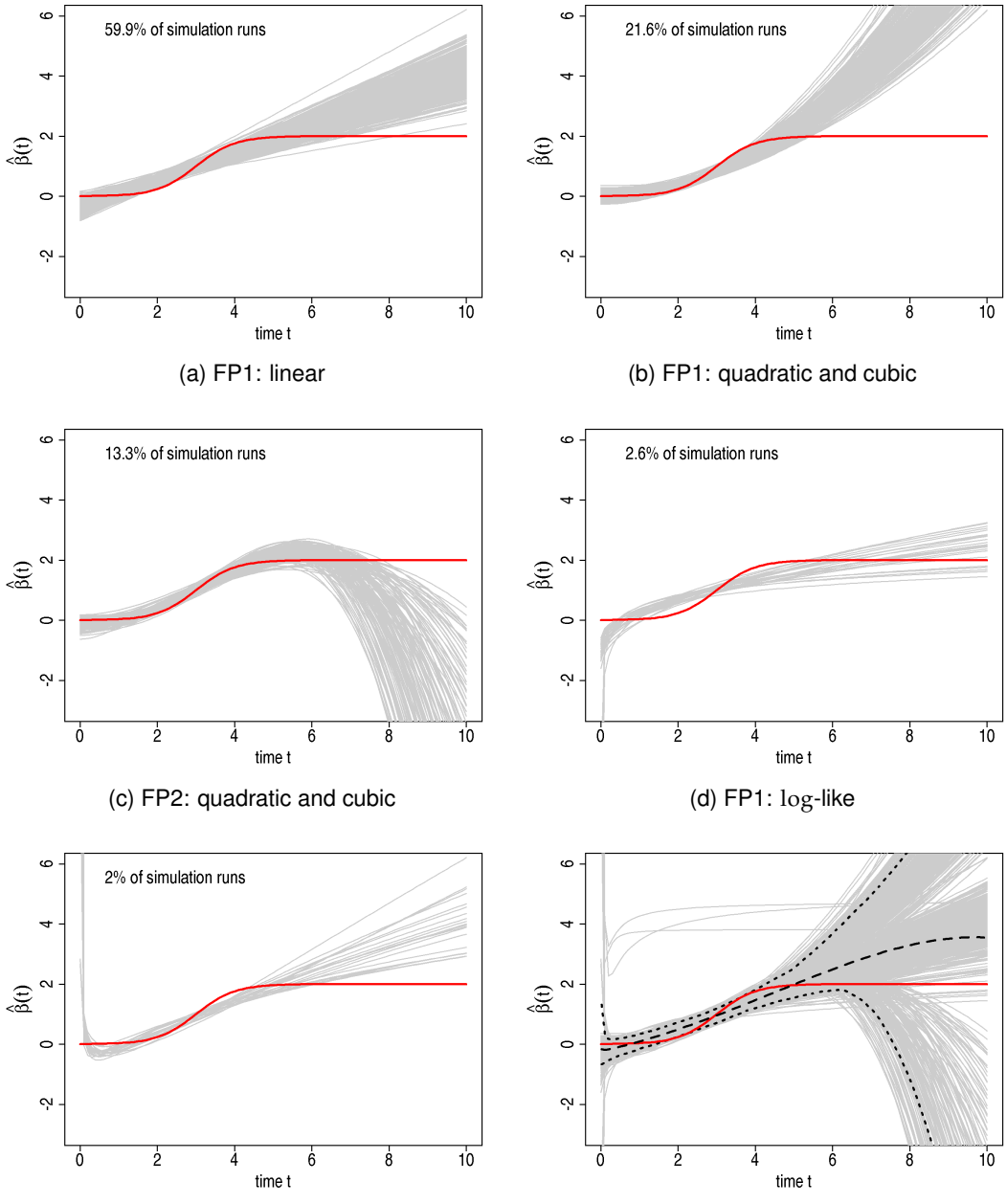


Figure 6.6: Correct (left) and incorrect (right) functional forms according to the qualitative type II error for effect (lw) with binary variable, sample size 1000 and light censoring. Shown are the true effect (—), estimated time-varying (—) and time-constant (boxplot) effects and their pointwise mean (**- -**).

72% considerably larger. Most of these wrong shapes are $\log$-like effects with a strong initial increase as shown in Figure 6.6. For larger sample size, both the crude and qualitative type II error considerably improve and the pointwise mean is close to the true effect.

The sigmoid effect (S) explores the limits of the FPT approach. Although a time-varying effect is detected in each simulation run, none of the selected effects fulfils the qualitative criteria. FPs of degree 2 are, although flexible per se, not flexible enough to describe this complex pattern. The variety of shapes selected as a substitute for the true sigmoid effect is manifold. Figure 6.7 gives an overview of the different types of selected FP functions. About half of the effects are linear. Other frequent alternatives are $\log$-like and quadratic functions, or more complex FP2 functions based on powers 0, 2 or 3. Detailed information on selected powers is given in Table D.2 in the Appendix.

Although the bathtub effects (Bs) and (Bw) are similarly complex, FPT shows better power in these scenarios. The crude type II error is very small, except for (Bw) with heavy censoring, where it is moderately large (15.8%). The qualitative type II error, on the contrary, increases. This is not surprising, as bathtub shapes are not included in the current class of FPs. As a consequence, the functions that are selected as substitutes is expected to vary broadly. Functions satisfying the qualitative criteria are unimodal with a minimum between 3 and 7. Most functions that do not meet these criteria are either monotonic or have an additional plateau near zero before they begin to decrease (Figure 6.8). However, the pointwise mean shows roughly the same shape than the true effect (Figure D.3).



(a) FP1: linear

(b) FP1: quadratic and cubic

(c) FP2: quadratic and cubic

(d) FP1: log-like

(e) FP2: log-like

(f) Shadow plot

Figure 6.7: Different shapes of estimated time-varying effects (—) for sigmoid effect (S) (—) with binary variable, sample size 1000 and light censoring. Shown are the most frequent FP types (Subfigures (a)-(e)) and the shadow plot of all selected effects (Subfigure (f)), including the remaining 0.6% of different FP types, together with the pointwise mean (- -) and 95% pointwise empirical confidence intervals (· · ·).

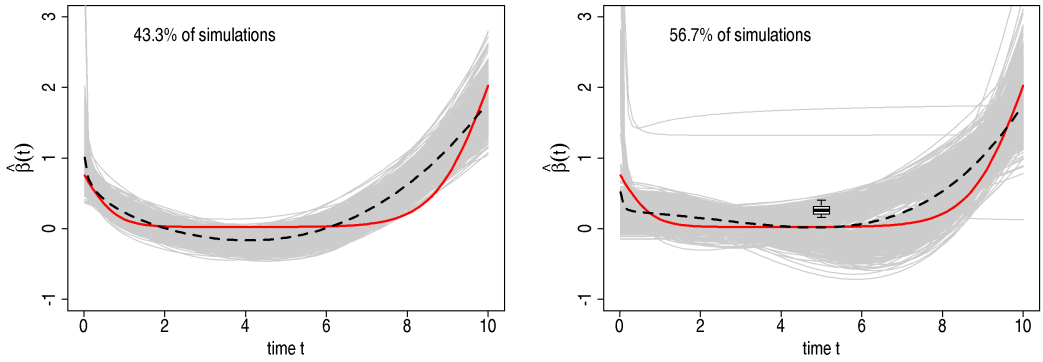


Figure 6.8: Correct (left) and incorrect (right) functional forms according to the qualitative type II error for effect (Bs) with binary variable, sample size 1000 and light censoring. Shown are the true effect (—), estimated time-varying (—) and time-constant (boxplot) effects and their pointwise mean (**- -**).

Difference of effects estimated by FPT to the true effects and comparison to CoxPH estimates

For the constant effect (C), pABCtime of FPT and CoxPH is similar. This is self-evident, as FPT is identical to a CoxPH model except for simulation runs with a type I error. The median pABCtime of both approaches decreases from about 20% for the smallest sample size to 5% for the largest (see Table 6.5). Hence, estimated effects are very close to the true effect. Analogously, the complete range of values decreases considerably with increasing sample size. While for smaller sample size some very large values of pABCtime are observed, which correspond to single time-varying effects (i.e. type I error), this peculiarity vanishes for larger sample sizes.

For the linear decreasing effects (Ls) and (Lw), the crude type II error with moderate sample size is large, hence differences in pABCtime between FPT and the CoxPH model are negligible. For larger sample size, where the power increases, FPT improves to 12%, while the median pABCtime for the CoxPH model remains at the same level of about 20% (Table 6.5).

For the non-linear decreasing effects (Ds) and (Dw), pABCtime clarifies the superiority of FPT estimates over the time-constant CoxPH effects. The latter show pABCtime values varying around 60% throughout (Table 6.5), while the FPT estimates are considerably better for not too small sample size and improve further with increasing sample size. Figure 6.9 exemplarily shows the pABCtime with moderate sample size 1000 and light censoring. The range of pABCtime of both approaches is nearly non-overlapping. The outliers of FPT, which extend into the range of CoxPH values correspond mostly to simulation runs with crude type II error, i.e. to time-constant effects.

For the strong increasing effect (Is), the crude type II error is quite large for small to moderate sample size. Hence, both CoxPH and FPT yield a similar median pABCtime of about 24%,

Model	Sample size	Cen-soring	Effects (C)	(Ls)	(Lw)	(Ds)	(Dw)	(ls)	(lw)	(S)	(Bs)	(Bw)
FPT	250	light	13.88	-	-	54.27	-	24.96	-	-	-	-
	500	light	9.33	-	-	20.39	-	23.99	-	-	-	-
	1000	heavy	8.12	-	-	18.78	-	23.14	-	43.90	134.01	44.92
	1000	light	6.25	20.44	20.37	15.22	62.88	23.04	49.80	45.45	96.96	34.15
	1000	no	5.89	22.47	19.77	14.23	18.34	12.36	44.57	-	-	-
	3000	light	4.27	11.65	11.30	12.59	12.97	6.03	21.50	-	-	-
CoxPH	250	light	13.53	-	-	59.68	-	24.80	-	-	-	-
	500	light	9.30	-	-	59.28	-	24.15	-	-	-	-
	1000	heavy	8.02	-	-	58.05	-	23.55	-	118.16	149.44	93.16
	1000	light	6.20	20.93	20.86	58.72	72.23	23.87	62.70	112.28	125.50	91.41
	1000	no	5.79	22.69	20.53	60.02	58.31	24.57	59.73	-	-	-
	3000	light	4.27	20.77	20.29	59.21	57.48	23.63	63.13	-	-	-

Table 6.5: Median pABCtime of effects estimated by FPT and the CoxPH model in univariate settings with binary variable.

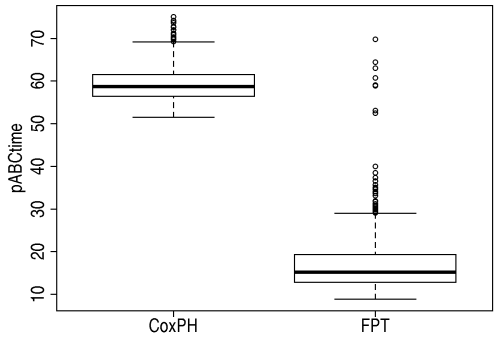


Figure 6.9: pABCtime of effects estimated by FPT and the CoxPH model for effect (Ds) with binary variable, sample size 1000 and light censoring.

with some larger outliers for exceptional effects estimated by FPT. For increasing sample size, though, FPT improves and reduces to about 7%. For the weak increasing effect (lw), differences to the true effect are larger. The median pABCtime of the CoxPH model is about 65% in all settings, while FPT decreases from 50% to 24% with increasing sample size.
Although the sigmoid effect (S) is not a member of the FP class and thus FPT is not capable of describing the true shape correctly, it is clearly superior to the CoxPH model. The latter either estimates an effect close to the lower plateau, differing considerably from the true effect from about year 3 on, or vice versa an effect similar to the upper plateau, which is different from the truth in the first three years. Alternatively, effect estimates may lie somewhere in between the two plateaus, resulting in a large difference to the true effect over the complete time period. None of these possibilities would be optimal. Hence, the sheer detection of the increasing nature of the effect by FPT considerably improves the pABCtime to about 30-50% compared to about 100% for the CoxPH model.

For the bathtub effects (Bs) and (Bw), the time-varying effects selected by FPT are again closer to the true effects in terms of pABCtime than the time-constant CoxPH effects. For effect (Bs) with a strong late increase, distances to the true effect are in general relatively large and so is the range of pABCtime (see Figure D.8 in the Appendix). For the effect (Bw) with a stronger initial effect, estimates of both approaches are more similar to the true effect. Simultaneously, the benefit of FPT compared to CoxPH is more pronounced for effect (Bw) and even intensifies with increased sample size.

Prediction error

The mPEC and dIPEC of CoxPH and FPT for the constant effect (C) are virtually identical, except for some outlying values of dIPEC for FPT, which correspond to type I errors, i.e. time-varying effects. The extent of this improvement in prediction error of FPT relative to CoxPH depends on the specific scenarios and parameter settings. The median dIPEC over the interval [0, 10] is given in Table 6.6. Additional information on the median dIPEC over different (shorter) intervals for all scenarios is given in Table D.3 in the Appendix.
For the linear decreasing effects (Ls) and (Lw) and the weak non-linear decreasing effect (Dw), the differences in dIPEC and mPEC between FPT and the CoxPH model are minor, due to the relatively large crude type II error. For larger sample sizes, where the type II error decreases, FPT improves slightly in terms of dIPEC, i.e. the amount of small dIPEC values increases. For the strong non-linear decreasing effect (Ds), the improvement of FPT compared to CoxPH increases. Differences between both approaches increase with increasing sample size. With large sample size, the ranges of dIPEC values of both approaches are non-overlapping.
For the strong increasing effect (Is), on the contrary, FPT is not able to improve the prediction

	Sample	Cen-	Effects									
Model	size	soring	(C)	(Ls)	(Lw)	(Ds)	(Dw)	(Is)	(Iw)	(S)	(Bs)	(Bw)
FPT	250	light	-7.98	-	-	-7.52	-	-11.99	-	-	-	-
	500	light	-8.13	-	-	-7.98	-	-12.18	-	-	-	-
	1000	heavy	-8.20	-	-	-5.34	-	-16.71	-	-15.94	-0.07	-1.89
	1000	light	-8.19	-32.66	-8.93	-8.10	-1.76	-12.31	-1.25	-7.02	-0.30	-3.59
	1000	no	-7.78	-29.91	-9.09	-10.01	-1.97	-9.79	-0.79	-	-	-
	3000	light	-8.22	-32.75	-9.01	-8.14	-2.17	-12.43	-1.35	-	-	-
CoxPH	250	light	-8.02	-	-	-7.31		-12.14	-	-	-	-
	500	light	-8.15	-	-	-7.44		-12.27	-	-	-	-
	1000	heavy	-8.21	-	-	-5.01		-16.76	-	-14.06	0.04	-1.61
	1000	light	-8.20	-32.63	-8.89	-7.48	-1.82	-12.33	-1.16	-5.42	-0.25	-3.06
	1000	no	-7.79	-29.92	-9.07	-9.28	-1.90	-9.78	-0.74	-	-	-
	3000	light	-8.24	-32.71	-8.93	-7.52	-2.00	-12.38	-1.18	-	-	-

Table 6.6: Median differences in integrated prediction error (dIPEC) to the Kaplan-Meier estimate (in %) over the interval [0,10] for FPT and CoxPH models in univariate settings with binary variable.

performance. For most settings, dIPEC and mPEC of FPT are similar to those of the CoxPH model. Several largely positive outliers are observed for the dIPEC of FPT for small sample sizes, which indicate an impairment compared to the Kaplan-Meier estimate in these simulation runs. This extremely bad prediction performance is caused by extraordinary effects selected in the corresponding simulation runs. In settings with many such extreme dIPEC values, simultaneously a marginally increased mPEC is observed for early times (see Figure 6.10). The proportion of such outliers, though, strongly reduces with increasing sample size.

For the weak increasing effect (Iw), differences in dIPEC and mPEC between FPT, CoxPH and the Kaplan-Meier estimate are negligible. In this scenario, a large proportion of events occurs in the first few years, where the true effect is close to zero. Hence, an effect estimate largely different from zero would cause an increased prediction error compared to the Kaplan-Meier method. This is not the case here, as both CoxPH and FPT correctly model the nearly absent effect in the relevant period where many individuals are at risk. Estimated effects are either very small or, in the case of FPT, increase over time, which mimics the shape of the true effect.

Very similar results are observed for the bathtub effect with weak initial effect and strong late increase (Bs). As the relatively weak initial effect vanishes soon, the variable has hardly any effect over a long time period, which results in a prediction performance that is only marginally better than the Kaplan-Meier estimate. The second bathtub effect (Bw), on the contrary, which has a strong initial effect and a weaker effect towards the end, shows a clear improvement of the two regression models compared to the Kaplan-Meier estimate in the relevant time period.

For the sigmoid effect (S), FPT is clearly superior to CoxPH in terms of dIPEC. For heavy censoring, several largely positive outliers are observed for FPT (Figure 6.11, left), which diminish for a smaller proportion of censoring. These bad predictions are again caused by some exceptional effects selected in these simulation runs. They also explain the discrepancy between the dIPEC, which is better for FPT than for CoxPH, and the mPEC which shows a minor impairment of FPT over the first 4 years (Figure 6.11, right). Estimated effects with artefacts, such as the $\log$-like effects shown in Figure 6.7, may dominate the prediction error in this region. Furthermore, the mean prediction error is more affected by these extreme values than the quantiles shown for the dIPEC. The mean dIPEC of FPT (-14.33%) also differs considerably from the median (-15.94%).

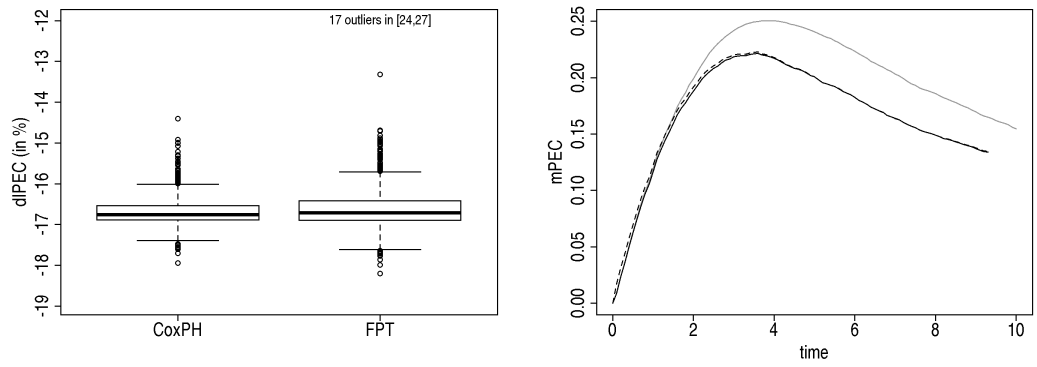


Figure 6.10: Differences in integrated prediction error (dIPEC, left) to the Kaplan-Meier estimate (in %) over the interval [0,10] and mean prediction error curves (mPEC, right) as pointwise mean over all 1000 simulation runs for the Kaplan-Meier estimate (—), CoxPH model (—) and FPT (- -) model in the setting with $n = 1000$, heavy censoring and binary variable for effect (ls).

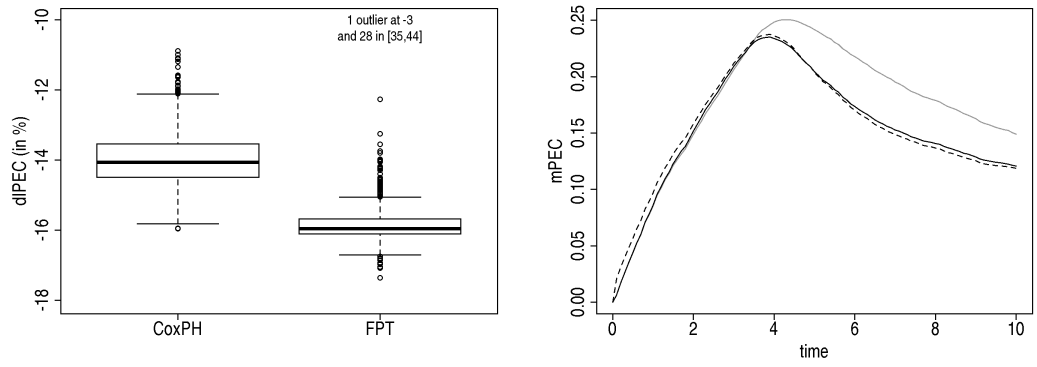


Figure 6.11: Differences in integrated prediction error (dIPEC, left) to the Kaplan-Meier estimate (in %) over the interval [0,10] and mean prediction error curves (mPEC, right) as pointwise mean over all 1000 simulation runs for the Kaplan-Meier estimate (—), CoxPH model (—) and FPT (- -) model in the setting with $n = 1000$, heavy censoring and binary variable for effect (S).

6.5.2 Part 2: Standard normal variable

The second part of the univariate simulation study presents the results for standard normal distributed variable.

Type I error

For scenario (C) with standard normal variable the type I error of the FPT algorithm is even slightly too conservative with values of at most 1% (Table 6.7).

Type II error and modelling of time-varying effects

The type II error depends on the specific parameter settings, but is in general very good. For the two linear decreasing effects (Ls) and (Lw), for example, the crude type II error is extremely good, apart from effect (Lw) with heavy censoring. Due to a relatively large number of selected $\log$-like and quadratic effects, the qualitative type II error is only moderate to poor for the strong and weak effect, respectively. Yet, visualisation of selected effects (Figure 6.12) reveals that most effects are at least close to linear. This is also reflected by the pointwise mean, which is virtually identical to the true effect for the stronger effect (Ls) and only slightly different for the weaker version (Lw).
The power for non-linear decreasing effects (Ds) and (Dw) is also extremely good. Crude type II error rates hardly differ from zero and even the qualitative type II error is very good in

Sample size	Censoring	Effects									
		(C)	(Ls)	(Lw)	(Ds)	(Dw)	(Is)	(Iw)	(S)	(Bs)	(Bw)
		Type I error	Crude type II error								
250	light	0.8	-	-	2.3	-	74.8	-	-	-	-
500	light	0.5	-	-	0.0	-	41.3	-	-	-	-
1000	heavy	0.7	0.5	25.8	0.0	2.4	12.1	-	0.0	0.0	0.0
1000	**light**	**0.7**	**0.0**	**0.5**	**0.0**	**0.3**	**6.9**	**56.6**	**0.0**	**0.1**	**0.0**
1000	no	1.0	-	-	0.0	-	2.5	57.7	-	-	-
3000	light	0.0	-	-	0.0	-	0.0	1.0	-	-	-
			Qualitative type II error								
250	light		-	-	5.1	-	75.6	-	-	-	-
500	light		-	-	4.6	-	43.4	-	-	-	-
1000	heavy		28.7	84.1	5.2	3.1	16.3	-	100.0	23.4	14.0
1000	light		13.7	45.4	13.3	2.5	9.9	82.7	100.0	9.3	2.1
1000	no		-	-	23.9	-	32.7	92.6	-	-	-
3000	light		-	-	68.0	-	15.0	32.0	-	-	-

Table 6.7: Type I error of scenario (C) and crude and qualitative type II error of scenarios (Ls) - (Bw) (in %) for significance level $\alpha = 0.01$ with standard normal variable in a simulation study of univariate scenarios.

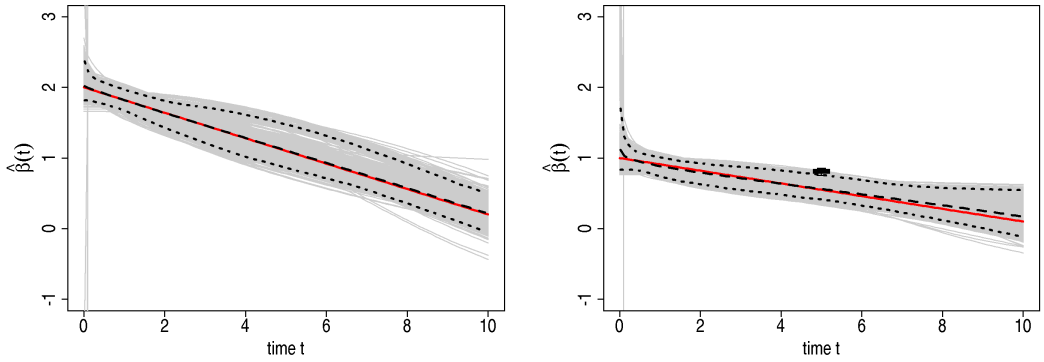


Figure 6.12: True (—) effect, estimated time-varying (—) and time-constant (boxplot) effects, pointwise mean over all estimated effects (- -) and 95% pointwise empirical confidence intervals (· · ·) for standard normal variable with sample size 1000 and light censoring for effects (Ls) (left) and (Lw) (right).

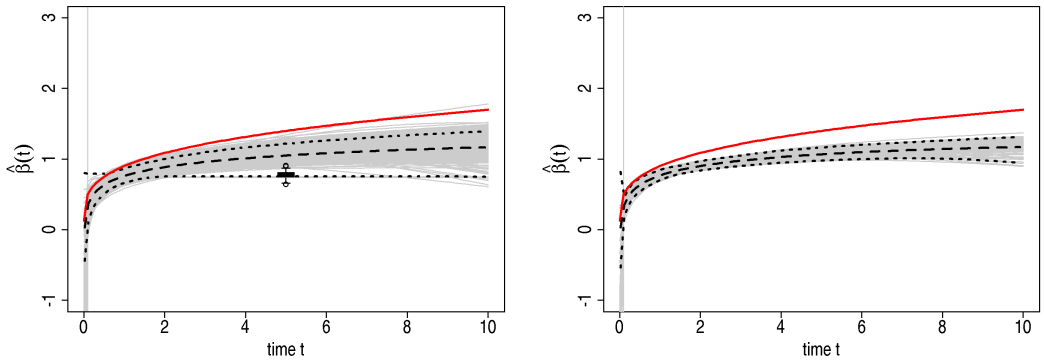


Figure 6.13: True (—) effect, estimated time-varying (—) and time-constant (boxplot) effects, pointwise mean (- -) and 95% empirical confidence intervals (· · ·) for effect (Is) with standard normal variable, sample sizes 1000 (left) and 3000 (right) and light censoring.

---

most settings. Surprisingly, it seems to be adversely affected by sample size. This is due to the fact that for larger sample sizes an increased number of unimodal functions is selected. However, these estimates are very similar to the true effect in the first ten years and the pointwise mean is very similar to the true effect in all parameter settings.

For the early increasing effect (Is), the crude type II error is very good for moderate to large sample size but increases considerably for small sample sizes. The qualitative type II error is only marginally larger, except for the two largest effective sample sizes. In these settings, several unimodal functions with minor decrease later in time are selected, which hardly differ from the 'correct' shape in the first 10 years. However, it is striking that nearly all estimated effects underestimate the true effect (Figure 6.13). The underestimation of the time-varying effect is paired with an overestimation of the baseline hazard. Hence, the resulting cumu-

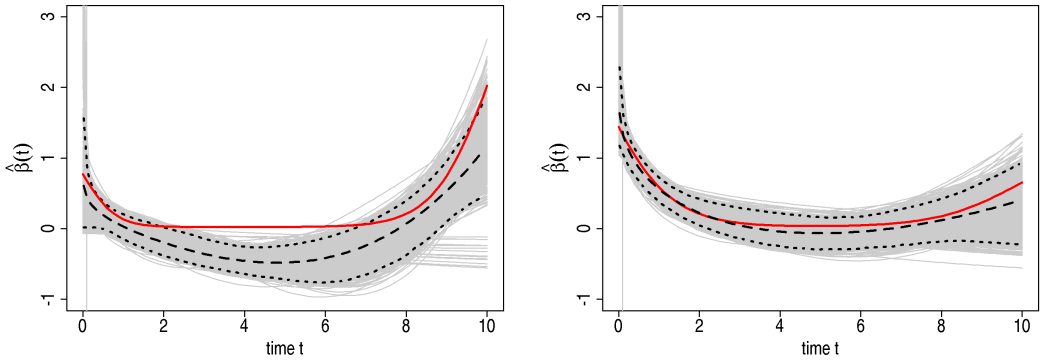


Figure 6.14: True (—) effect, estimated time-varying (—), pointwise mean (- -) and 95% empirical confidence intervals (· · ·) for effects (Bs) (left) and (Bw) (right) with standard normal variable, sample size 1000 and heavy censoring.

---

lative hazard functions are less affected by this phenomenon but still tend to underestimate the true cumulative hazard (see Figure D.5 in the Appendix). This can be explained by a frailty effect. Observations with large absolute X values relate to early events, i.e. the subgroup of individuals at risk changes systematically over time. The upper edge with large covariate values thins out faster than the lower edge (see Figure D.4 in the Appendix). After these subjects with large X value experienced the event, the values of observations still at risk are smaller and thus the observed effect at these time points may be reduced. A small investigation on the impact of the variance of X ($\sigma^2 = 0.25$, 0.5, 1, 2 and 4) based on one data set each with $n = 1000$ reveals that the amount of underestimation increases with increasing variance, caused by the frailty effect (Figures D.6 and D.7). The estimate of the cumulative baseline hazard shows a trend in the opposite direction. The resulting cumulative hazard functions show a slight tendency towards underestimation of the true hazard, but to a lesser extent than the effect estimates suggest. Such deviations from the true effect are not detected by the qualitative type II error, but ABCtime, which is applied in Section 6.5.2 is designed to measure deviations of such nature.

For the gently inclined effect (lw), the time-varying effect is overlooked in about half of the simulation runs with sample size 1000 and only 10-20% of selected effects comply with the qualitative criteria. For sample size 3000, error rates considerably improve. Again, a systematic underestimation of the true effect is observed. In this scenario, though, both the poor power and the underestimation may also be explained by the extremely small average population effect. The true effect is close to zero in the first years and begins to rise later in time, where the risk set is considerably decreased. This naturally results in a decreased power for detecting the time-varying effect.

For the sigmoid effect (S), results are analogue to those with binary variable. FPT detects the time-varying effect in all simulation runs, but completely fails to describe the true functional

form, as it is not a member of the FP class. The most frequent substitutes in this setting are FP2 functions composed of powers 1, 2 and/or 3 (71.9%) or powers 0, 0.5 and/or -0.5 (25.4%). More details on selected FP combinations are given in Table D.2 in the Appendix.
For the bathtub effects (Bs) and (Bw), the crude and qualitative type II error of FPT are very good. Only for the combination of effect (Bs) with heavy censoring it increases to moderate 23%, which is mainly due to an increased number of effects with an initial plateau (Figure 6.14, left). The pointwise mean for scenario (Bw) is even quite close to the true effect (Figure 6.14, right).

Difference of effects estimated by FPT to the true effects and comparison to CoxPH estimates

Analogue to the settings with binary variable, pABCtime for the constant effect (C) is virtually identical for FPT and CoxPH effects, due to the very low amount of type I errors. For the time-varying effects, FPT effects are in general closer to the true effect than the time-constant CoxPH effects (Table 6.8). For example, the pABCtime of the CoxPH effects remains relatively constant over all settings with median values around 25% for both linear decreasing effects (Ls) and (Lw). Results for FPT in the same settings are considerably better, with improved pABCtime for larger sample size. The same tendency is observed for the non-linear decreasing effects (Ds) and (Dw). The difference between the CoxPH effects and the true effects is in general rather large and remains relatively constant over all parameter settings (see Table 6.8), while the range and median of pABCtime tends to decrease with increasing sample size.
For the strong increasing effect (Is), the median pABCtime for the CoxPH model is again relatively constant over all settings. Although the effects estimated by FPT tend to under-

	Sample	Cen-	Effects									
Model	size	soring	(C)	(Ls)	(Lw)	(Ds)	(Dw)	(Is)	(Iw)	(S)	(Bs)	(Bw)
FPT	250	light	7.96	-	-	18.09	-	32.78	-	-	-	-
	500	light	5.50	-	-	15.52	-	27.04	-	-	-	-
	1000	heavy	5.27	6.01	14.14	14.79	16.69	31.06	-	43.05	165.83	25.40
	1000	light	4.05	4.33	9.21	14.32	43.36	22.05	85.77	32.98	86.77	16.66
	1000	no	3.55	-	-	14.55	-	22.90	63.26	-	-	-
	3000	light	2.03	-	-	10.16	-	21.22	83.22	-	-	-
CoxPH	250	light	7.87	-	-	63.23	-	35.45	-	-	-	-
	500	light	5.50	-	-	63.16	-	35.18	-	-	-	-
	1000	heavy	5.26	23.44	22.85	62.93	58.45	40.06	-	93.31	147.55	89.50
	1000	light	4.02	26.68	24.63	62.72	78.51	35.26	87.73	92.63	86.72	92.35
	1000	no	3.49	-	-	62.69	-	35.67	65.24	-	-	-
	3000	light	2.03	-	-	62.47	-	34.65	86.97	-	-	-

Table 6.8: Median pABCtime of effects estimated by FPT and the CoxPH model in univariate settings with standard normal variable.

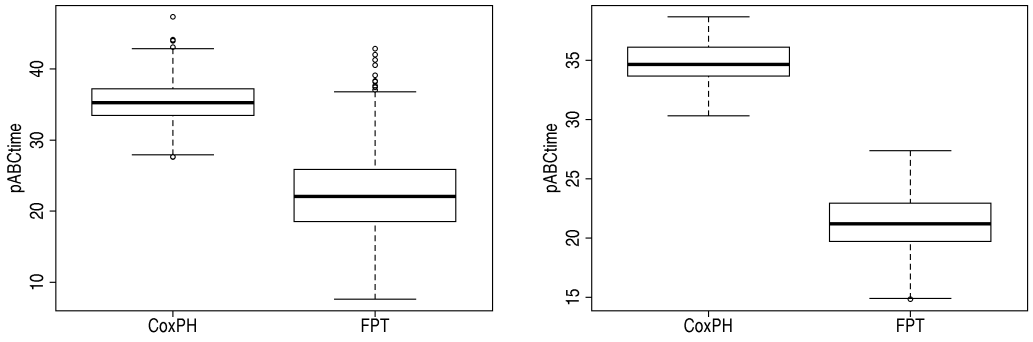


Figure 6.15: pABCtime of effects estimated by FPT and the CoxPH model for effect (Is) with sample sizes 1000 (left) and 3000 (right), light censoring and standard normal variable.

estimate the true effects, their pABCtime is still better than for the CoxPH estimates and decreases further with increasing sample size (Figure 6.15). For the weak increasing effect (Iw), on the contrary, CoxPH and FPT are similar in terms of pABCtime in all settings.
The results for the sigmoid effect (S) show the gain in pABCtime by time-varying effects relative to time-constant effects, even if the true effect cannot be fully described by the estimated effects (Figure D.9). The median pABCtime of 90% for the CoxPH effects is a clear contrast to the 30-40% of FPT effects.
For the bathtub effects both FPT and CoxPH are not able to describe the true underlying effect exactly and thus largely differ from the true effect either in the early or late time period, where the effect differs from zero, or in the time period, where it has a plateau near zero. This results in a large pABCtime for both approaches in scenario (Bs) and similarly applies to the CoxPH model for scenario (Bw) with a stronger initial effect, where it gives a median pABCtime of about 90%. FPT, though, is considerably closer to the true effect in this scenario and further improves with smaller proportion of censoring (Table 6.8).

Prediction error

The prediction performance of FPT and CoxPH models in the settings with constant effect (C) is virtually identical (Table 6.9). Results for scenarios with time-varying effects, though, are intermingled. Detailed information on the median dIPEC over different intervals for all scenarios is given in Table D.4 in the Appendix.
While the dIPEC over [0,10] in scenarios (Ls), (Lw) and (Dw) reveals a slight superiority of FPT relative to CoxPH models, the mPEC is very similar, with only a minor advantage for FPT at later times. For the stronger non-linear decreasing effect differences between both approaches become more apparent. The improvement in mPEC is visible over the complete time span (Figure 6.16) and remains relatively constant over all parameter settings, while

Model	Sample size	Cen-soring	Effects (C)	(Ls)	(Lw)	(Ds)	(Dw)	(Is)	(Iw)	(S)	(Bs)	(Bw)
FPT	250	light	-25.05	-	-	-18.70	-	-20.45	-	-	-	-
	500	light	-25.17	-	-	-18.86	-	-20.40	-	-	-	-
	1000	heavy	-21.74	-70.51	-21.99	-15.46	-5.03	-18.73	-	-16.55	-2.01	-5.48
	1000	light	-25.26	-70.35	-23.78	-18.97	-5.61	-20.36	-0.04	-10.98	-0.56	-8.89
	1000	no	-26.25	-	-	-21.57	-	-21.95	-0.22	-	-	-
	3000	light	-25.28	-	-	-19.04	-	-20.40	-0.12	-	-	-
CoxPH	250	light	-25.05	-	-	-16.26	-	-20.53	-	-	-	-
	500	light	-25.17	-	-	-16.34	-	-20.59	-	-	-	-
	1000	heavy	-21.74	-69.48	-21.78	-13.48	-4.66	-18.56	-	-7.99	-0.93	-2.96
	1000	light	-25.24	-68.66	-23.38	-16.41	-4.48	-20.65	-0.03	-6.37	0.08	-5.84
	1000	no	-26.25	-	-	-18.61	-	-22.12	-0.23	-	-	-
	3000	light	-25.28	-	-	-16.41	-	-20.68	-0.06	-	-	-

Table 6.9: Median differences in integrated prediction error (dIPEC) to the Kaplan-Meier estimate (in %) over the interval [0,10] for FPT and CoxPH models in univariate settings with standard normal variable.

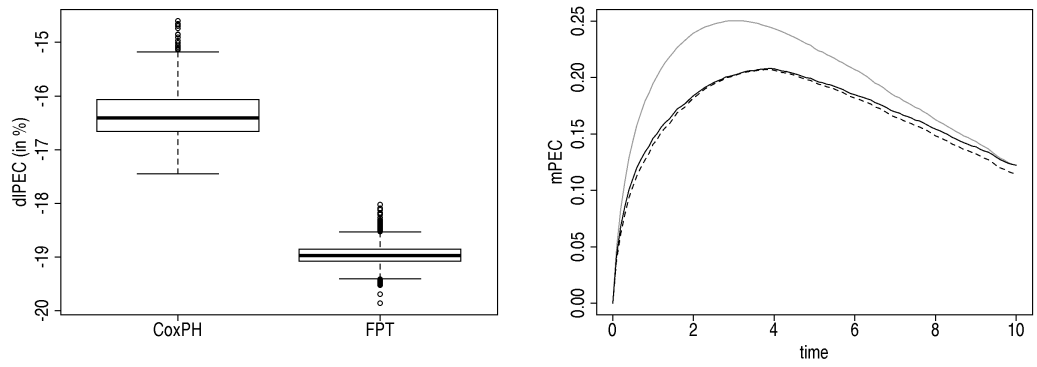


Figure 6.16: Differences in integrated prediction error (dIPEC, left) to the Kaplan-Meier estimate (in %) over the interval [0,10] and mean prediction error curves (mPEC, right) as pointwise mean over all 1000 simulation runs for the Kaplan-Meier estimate (—), CoxPH model (—) and FPT (- -) model in the setting with $n = 1000$, heavy censoring and standard normal variable for effect (Ds).

---

the dIPEC further improves with increasing sample size.

For the increasing effects (Is) and (Iw), on the contrary, differences in prediction error between FPT and CoxPH are negligible. For (Iw) this may be explained by the large variety of time-varying effects estimated by FPT, which are not able to describe the true effect correctly and hence are not beneficial in terms of the prediction performance. Furthermore, both CoxPH and FPT are similar to the Kaplan-Meier estimate. This result is not surprising, as the true effect is extremely small over the first years and increases only later in time, where the risk set is considerably decreased. Hence, as the covariate indeed has hardly any influence in the early years, the regression models cannot be expected to improve prediction performance compared to the Kaplan-Meier estimate in this time period.

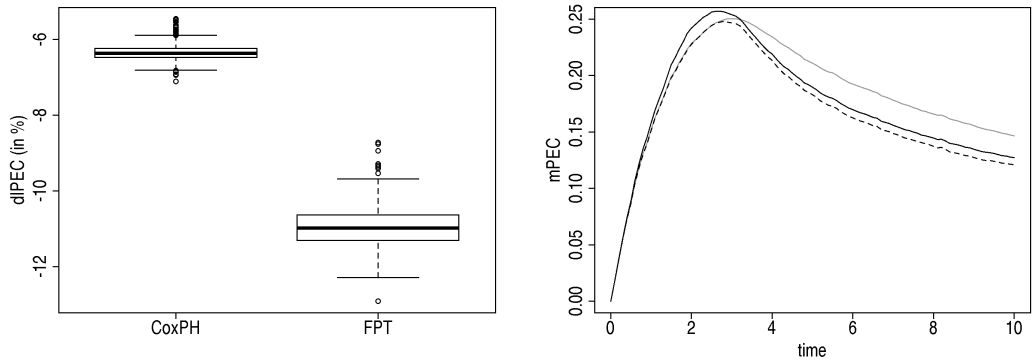


Figure 6.17: Differences in integrated prediction error (dIPEC, left) to the Kaplan-Meier estimate (in %) over the interval [0,10] and mean prediction error curves (mPEC, right) as pointwise mean over all 1000 simulation runs for the Kaplan-Meier estimate (—), CoxPH model (**—**) and FPT (**- -**) model in the setting with $n = 1000$, heavy censoring and standard normal variable for effect (S).

For the sigmoid effect (S), FPT shows an improvement in prediction error compared to CoxPH. Although the median dIPEC of both approaches is rather small, i.e. the integrated prediction error is relatively similar to the Kaplan-Meier estimate, the mPEC shows the differences more explicitly (Figure 6.17). The true effect is nearly zero in the first years with a steep increase around year 3. Hence, the Kaplan-Meier estimate is expected to perform reasonably well early in time, where the covariate information indeed has hardly any influence on survival. The mean CoxPH effect, though, is about 0.3 (see Table D.1 in the Appendix) and hence considerably overestimates the true effect early in time, resulting in an inflated prediction error in this period. From about year 3 on, the mPEC of the CoxPH models considerably improves relative to the Kaplan-Meier estimate. Although the FPT models are not able to describe the true effect correctly, they reproduce the increasing nature well enough to give a similarly good prediction error than the Kaplan-Meier estimate for the early time period with a considerably improved mPEC later on and hence yield the best prediction performance throughout.
For the bathtub effect (Bw) with a strong initial effect, FPT again improves both the dIPEC and mPEC compared to CoxPH and Kaplan-Meier. For the initially weaker effect (Bs) benefits relative to the Kaplan-Meier estimate vanish.

6.5.3 Power of FP analysis

The hierarchical closed test procedure of FPT, which first tests FP2 vs. constant effect, may have reduced power if the true effect is logarithmic or FP1, as the additional degrees of freedom are wasted. Royston and Sauerbrei (2008, chap. 4.16) present a small simulation

		Crude (qualitative) type II error with FPs of maximum degree	
Effect	X	2	1
(Ls)	binary	74.5 (95.8)	72.0 (95.7)
(Is)	binary	48.2 (50.3)	32.9 (36.2)
(lw)	normal	56.6 (82.7)	44.8 (71.9)

Table 6.10: Crude and qualitative type II error (in %) of FPT with maximum degree 1 and 2 in three scenarios with sample size 1000 and light censoring.

study in the framework of non-linear functional forms of covariates using a similar function selection procedure. They show that if the true function is simple, i.e. similar to the default or an FP1, the tests on more complex FPs loose power. Especially with weak effects and/or small sample size, time-varying effects are more likely to be detected when restricting FPs to maximum degree 1. However, this restriction involves the risk of overlooking or mismodelling more complex or non-monotonic effects.

The improvement of power with increasing sample size is also observed for the FPT algorithm (Tables 6.4 and 6.7). To check whether the poor power of some scenarios with moderate sample size ascribes to the test on FP2, we restrict selection of FPs to maximum degree 1 for three settings with a large crude type II error: (Ls) and (Is) with binary variable and (lw) with normal variable in the main setting. Effects (Ls) and (Is) are FPs of degree 1. Although (Is) is not exactly included in the standard class of FPs, it is rather similar to power 0.5 or the default transformation $\log(t)$. The true form of effect (lw) is more complex and not a member of the FP class. Hence, at least the power of scenarios (Ls) and (Is) is expected to improve when omitting the search for a possible improvement in fit by FP2 functions.

Table 6.10 contrasts the crude type II error for the restricted test procedure to the standard FPT algorithm with maximum degree 2. The improvement for the linear effect (Ls), which was supposed to benefit most from the restriction, is with a value of 2.5% surprisingly small. The crude type II error of scenarios (Is) and (lw), on the contrary, considerably improves. As the qualitative type II error improves about the same amount, restriction to FP1 is not at the expense of the functional form but improves the power in both aspects. Hence, consideration of too complex functional forms may explain the poor power of some scenarios to a certain extent, but seems not to be the prime reason.

6.5.4 Time transformation - the default

The FPT algorithm utilises a default time transformation in order to stabilise selected effects. This default transformation is chosen as $\log(t)$, which is a plausible choice for short-term

effects and is widely used in the analysis of time-varying effects. If the true effect is approximately logarithmic, this default should also control the qualitative type II error, as it helps to avoid selection of a larger number of exceptional curve shapes caused by artefacts in the data. However, if the underlying effect differs from this shape, the default transformation may potentially inflate the qualitative type II error. The crude type II error, on the contrary, is not affected by the choice of the default transformation, as it is based only on the test of the best FP2 vs. constant effect.

To investigate the impact of the default transformation on the qualitative type II error, we modify the FPT algorithm by omitting the second test, i.e. the test on FP2 vs. default $\log(t)$. Hence, if the time-varying effect is significant (best FP2 vs. constant), but the indication for the more complex FP2 is not strong enough, i.e. the test on the best FP2 vs. best FP1 is not significant, then the best FP1 function is selected. This modified algorithm without the default transformation is applied to selective scenarios with poor, moderate and good qualitative type II error or large difference between crude and qualitative type II error.
The linear effect (Ls) already has a poor crude type II error in combination with binary variable, but worsens further in terms of qualitative type II error. As the true effect is not conform with the steep initial decrease of the $\log$ function, the qualitative type II error is expected to improve when omitting the default. The effects (Ds) and (Is) with good and moderately worse type II error, on the contrary, are quite similar to the default and thus are expected to hold the qualitative type II error or even to change for the worse. (Iw) and (Bs) have good to moderate crude type II error rates with binary variable but considerably inflated qualitative type II error. As both have little similarity to the $\log$ function, the qualitative type II error is expected to improve without the default transformation.

The observed results do indeed point in these directions. Although the FP powers in general change in a relatively large number of simulation runs when omitting the default transformation (Table 6.11), the gain in qualitative type II error differs widely (Table 6.12). Especially the improvement for the linear effect is less pronounced than expected. Only half of the former default transformations selected with binary variable change to linear FPs (Table 6.11). Hence, the gain in qualitative type II error is moderate (Table 6.12). Results for the setting with standard normal variable even hardly change at all. Hence, the (relatively low) qualitative type II error of FPT with default is not caused by the usage of the default transformation $\log(t)$.
For the $\log$-like effects (Ds) and (Is), the qualitative type II error even worsens. In most of the simulation runs with changes in selected FPs, the alternatives are still $\log$-like functions like $t^{0.5}$ and $t^{-0.5}$. These powers are also very frequent with the bathtub effect (Bw), as well as a larger amount of FP2 functions based on the same powers. The improvement in qualitative type II error, though, is moderate in this scenario. The largest gain is observed for the gently inclined effect (Iw), where the default $\log(t)$ is most frequently replaced by linear, quadratic,

FP powers		X binary					X standard normal		
p_1	p_2	(Ls)	(Ds)	(Is)	(Iw)	(Bw)	(Ls)	(Ds)	(Is)
-1	-0.5		0.3					1.6	
-0.5		0.3	8.1	3.4	0.1	12.6		0.3	17.3
-0.5	-0.5		0.2			0.1		2.4	
0.5		7.4	8.3	17.3	7.0	0.2	0.1	0.5	7.2
0.5	0.5					0.1		1.9	0.3
0.5	1		0.6			2.9		3.0	0.8
0.5	2		0.1			1.3		0.1	
0.5	3		0.2			1.2		0.2	0.2
1		7.4		6.9	16.3				
1	1		0.7			5.1		1.0	0.7
2		0.7		1.4	13.1				
3		0.1		0.7	5.5				
		and 2 further FPs $<1\%$	and 6 further FPs $<1\%$	and 3 further FPs $<1\%$	and 2 further FPs $<1\%$	and 6 further FPs $<1\%$		and 3 further FPs $<1\%$	and 8 further FPs $<1\%$
Σ		16.3	19.8	30.5	42.2	25.9	0.1	11.3	28.2

Table 6.11: Frequency of FP powers (in %) selected as substitutes for the default transformation $\log(t)$ in FPT without default transformation in the main setting with sample size 1000 and light censoring.

			Qualitative type II error	
Effect	X	Crude type II error	with default	w/o default
(Ls)	binary	74.5	95.8	88.4
(Ds)	binary	0.8	3.0	6.2
(Is)	binary	48.7	50.3	59.4
(Iw)	binary	27.9	71.8	36.9
(Bw)	binary	1.1	56.5	44.4
(Ls)	normal	0.0	13.7	13.7
(Ds)	normal	0.0	13.3	23.8
(Is)	normal	6.9	9.9	12.9

Table 6.12: Crude and qualitative type II error (in %) of FPT with and without the default time transformation in the main setting with sample size 1000 and light censoring.

cubic or square-root transformations. In this scenario, the extremely large qualitative type II error reduces to about the half.

This limited investigation indicates that for effects with shapes similar to logarithmic decay or increase, the default transformation achieves its purpose of stabilising selected effect functions as expected. However, if the true effect is of a different shape, use of the default transformation may lead to a considerably inflated qualitative type II error.

6.6 Properties of FPT in multivariable settings

The investigation of the univariate settings shows promising results. This section further explores the performance of FPT in multivariable model building based on a model with five covariates, two of which have a time-varying effect. We investigate six different settings, varying the proportion of censoring (no, light and heavy) and the correlation between covariates ($\rho_{X_1,X_4} = 0$ and 0.5).

6.6.1 Selection and modelling of time-varying effects

Investigation of type I and II error in these settings shows slightly larger error values in settings with correlation between X_1 and X_4 (Table 6.13). However, considering the limited precision of 100 simulation runs, the observed changes are considered to be of minor importance.
For the constant effects of X_3 to X_5, type I error is slightly inflated (3%) in individual settings which may be due to chance. Rolled into one, the type I error over all three variables and six scenarios is 0.9%. The crude type II error rates for the time-varying effects of X_1 and X_2 decrease with increasing sample size, which is conform to the results of the univariate

	Cen-	Type II error				Type I error		
ρ_{X_1,X_4}	soring	X_1		X_2		X_3	X_4	X_5
		crude	qual.	crude	qual.			
0	heavy	7	8	32	32	2	0	0
	light	2	5	26	26	1	0	0
	no	0	3	13	15	1	1	0
0.5	heavy	9	10	34	36	1	2	1
	light	1	7	22	24	0	1	1
	no	0	5	14	16	2	3	0

Table 6.13: Type I and II error (in %) of FPT in multivariable settings.

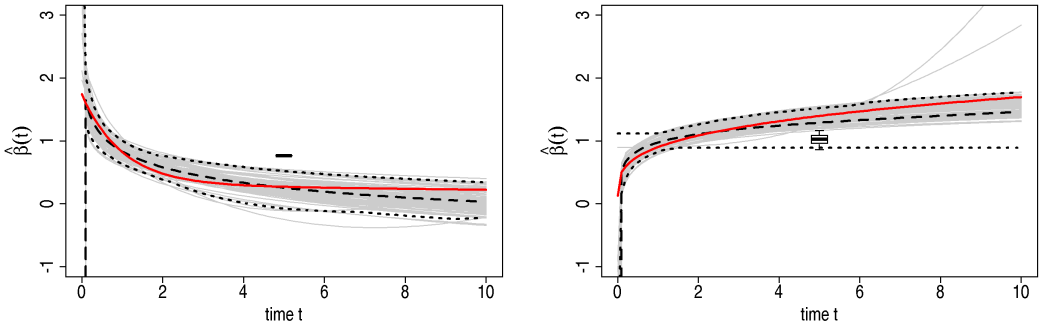


Figure 6.18: True (—) effect, estimated time-varying (—) and time-constant (boxplot) effects, pointwise mean (- -) and 95% empirical confidence intervals (· · ·) for effects of X_1 (left) and X_2 (right) in the multivariable setting with light censoring and ρ_{X_1,X_4}=0.5.

---

investigation. While the crude type II error for X_1, i.e. effect (Ds), is similar to the univariate settings, it decreases to about the half for effect (Is) of X_2. It reaches an acceptable level in the settings without censoring and moderately large values for light and heavy censoring. The qualitative type II error of both effects increases only slightly compared to the crude type II error, as also observed in the univariate investigation.
Looking at the selected effects themselves, a similar picture emerges. Estimated effects are rather close to the true effect, as shown in Figure 6.18 for the two time-varying effects.

In multivariable models, the sequence in which time-varying effects are selected may also influence the final model if the time-varying effects of two variables are approximately equally significant, as observed for the Rotterdam breast cancer series (Section 4.1.3).
Considering the selection sequence of time-varying effects in the simulation study, some changes can be observed between the investigated settings (Table D.5). The 'correct' model including time-varying effects for X_1 and X_2 is selected in the majority of simulation runs, but its selection frequency decreases for correlated variables and larger proportion of censoring. This is due to the decreasing proportion of the sequence X_1X_2. Simultaneously, the proportion of X_2X_1 increases, although less strongly. With increasing proportion of censoring, the proportion of simulation runs in which either of the two is selected increases. In most of these cases, the time-varying effect of the other variable is not significant at the 1% level, but at the 5% level. Furthermore, with no and heavy censoring the proportion of simulation runs with selection sequence X_1X_2 in the settings with correlation is smaller than in settings without correlation.

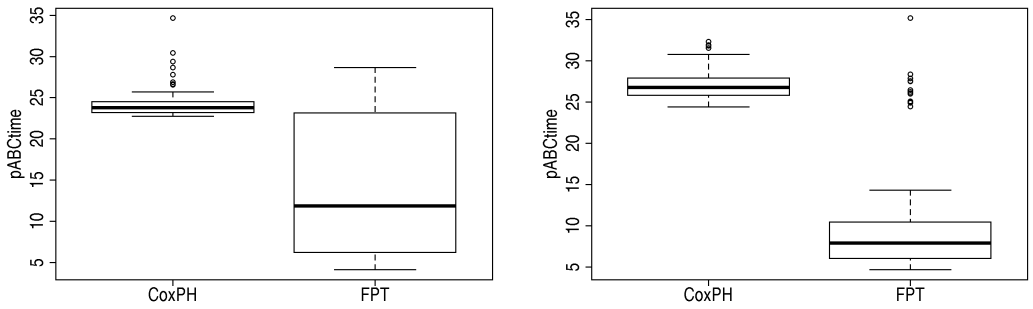


Figure 6.19: pABCtime of effects selected by CoxPH and FPT relative to the true effect function for X_2 in the multivariable setting with heavy (left) and no (right) censoring and $\rho_{X_1,X_4} = 0$.

6.6.2 Difference of effects estimated by FPT to the true effect and comparison to CoxPH

The distance to the true effect function in terms of pABCtime for effect (Ds) of X_1 is analogue to the results obtained for the univariate settings (for median pABCtime see Tables D.6 and D.7 in the Appendix). The pABCtime of both the CoxPH and FPT estimates remain relatively stable over all settings with median values about 60% and 16%, respectively. Hence, FPT is superior to CoxPH.
For the increasing effect (Is) of variable X_2, CoxPH shows a median pABCtime of 25% over all settings, which is similar to the univariate investigations. For FPT, the univariate settings reveal a large range of pABCtime with a median value similar to CoxPH with light and heavy censoring which improves without censoring. In the multivariable framework, variation of pABCtime is still large, but considerably smaller as compared to the univariate case. Furthermore, median values decrease from about 12% for heavy censoring to 8% without censoring (see Figure 6.19). Hence, for this effect FPT reveals its full strength in the multivariable setting.
For the three time-constant effects of X_3 to X_5, pABCtime of CoxPH and FPT are identical with median values between 20% and 5%, depending on the size of the true effect.

6.6.3 Prediction error

The mean prediction error curve in all six scenarios is very similar, with FPT performing slightly better than CoxPH, especially for early times (Figure 6.20). The difference in IPEC to the Kaplan-Meier estimate for both approaches varies between 20% and 30%, depending on the specific parameter setting. Improvement with respect to the Kaplan-Meier estimate

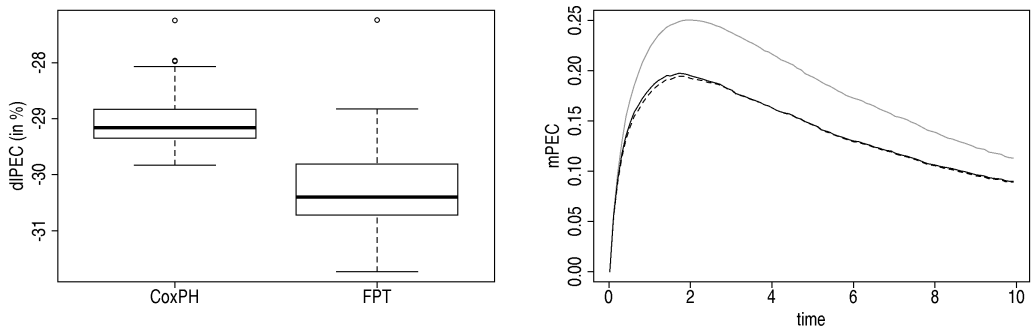


Figure 6.20: Differences in integrated prediction error (dIPEC, left) to the Kaplan-Meier estimate (in %) and mean prediction error curves (mPEC, right) as pointwise mean over all 100 simulation runs for the Kaplan-Meier estimate (—), CoxPH (**—**) and FPT (− −) in the setting with light censoring and $\rho_{X_1,X_4} = 0.5$.

ρ_{X_1,X_4}	Censoring	FPT	CoxPH
0	heavy	-23.63	-23.07
0	light	-25.84	-24.83
0	no	-26.21	-24.76
0.5	heavy	-27.13	-26.47
0.5	light	-30.40	-29.16
0.5	no	-30.88	-29.23

Table 6.14: Median difference (in %) of IPEC over [0,10] to the Kaplan-Meier estimate (dIPEC) for FPT and CoxPH in multivariable settings.

increases with decreasing proportion of censoring and with correlation. In terms of prediction error, FPT is superior to CoxPH in all of the settings (Table 6.14).

6.7 Convergence problems

As already mentioned in Section 6.3, extreme survival times may cause convergence problems. For FPT, convergence problems are observed only for single FP combinations within the selection procedure. To investigate whether these problems distort the estimated effects, we recorded all convergence problems in the univariate simulation study. The percentage of simulation runs in which at least one FP combination fails is given in Table 6.15. In most simulation settings, convergence problems are rare. For normal variable, where large covariate values may intensify problems, non-convergence occurs in up to 30% in single parameter settings.

	Sample	Cen-	Effects									
X		soring	(C)	(Ls)	(Lw)	(Ds)	(Dw)	(Is)	(Iw)	(S)	(Bs)	(Bw)
binary	1000	heavy	3.0	-	-	7.0	-	6.9	-	15.3	4.8	6.7
	1000	light	0.8	0.3	0.9	2.2	0.5	1.2	1.4	3.8	0.4	1.7
	1000	no	0.2	0.2	3.7	0.4	0.3	0.4	0.1	-	-	-
normal	1000	heavy	0.9	0.0	0.2	2.7	1.0	0.2	-	12.1	8.3	7.1
	1000	light	3.2	0.2	0.4	10.0	4.8	1.2	2.9	27.2	32.0	31.4
	1000	no	8.9	-	-	18.1	-	0.1	19.4	-	-	-

Table 6.15: Convergence problems (in %) in univariate settings.

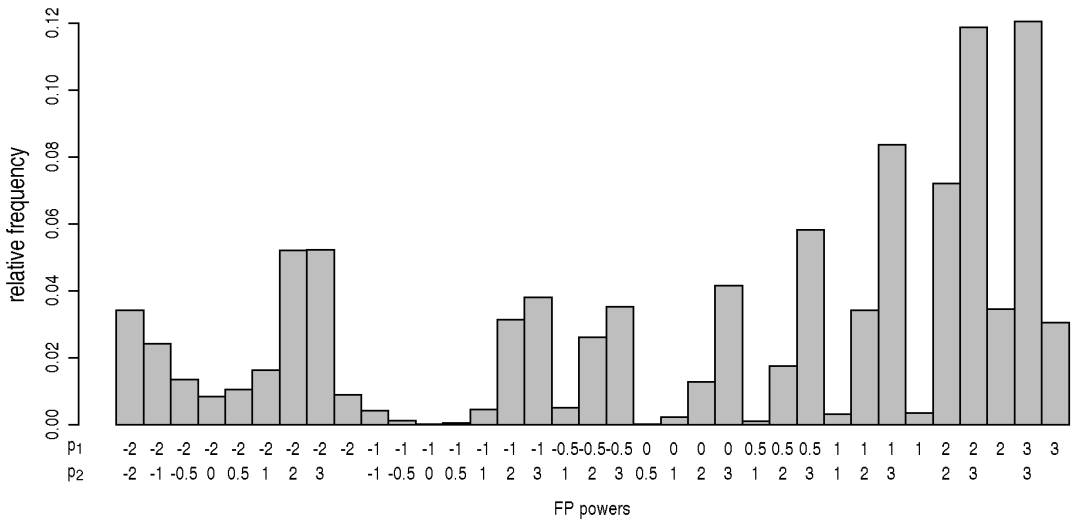


Figure 6.21: Frequency of FP powers that fail to converge relative on all powers that fail.

This leads to the question whether these simulation runs result in systematically different effect estimates than those without convergence problems. To explore this topic, we contrast estimated effect functions of simulation runs with and without convergence problems. However, visual comparison does not reveal any systematic differences. The distance to the true effect, measured by pABCtime, is inconspicuous in all settings, too. Hence, convergence problems in single FP combinations seem to be unproblematic, as they do not affect the estimated effects.

An investigation of the specific FPs, for which convergence problems occur, reveals that FP combinations with powers -2, 2 and 3 most frequently cause problems (see Figure 6.21). Hence, these powers may be less adequate for estimation of time-varying effects. Excluding these powers from the set of power terms would reduce convergence problems to a minimum, but at the expense of considerably limited flexibility of effect estimates. Another

possible solution is the restriction to FP1 functions, as FP2 functions appear more likely to cause convergence problems. Such restrictions, though, are not advisable unless convergence problems occur for a larger number of FPs in a single model, as single failures do not seem to affect the final estimates.

6.8 Difficulties with the Semiparametric Extended Cox model

The Semiparametric Extended Cox model is executed in `R-2.5.1` with package `timereg`, version 1.2.4. This non-parametric approach is potentially more flexible in modelling time-varying effects than parametric approaches such as FPT, and hence is believed to perform better for complex time-varying patterns.
However, in our simulation study the approach turns out to be very sensitive to the choice of the test statistic and the bandwidth and effect estimates tend to be instable especially at larger times. None of these problems can be solved satisfactorily in the context of this investigation. As we cannot be sure to obtain reliable and comparable results, we exclude the approach from the simulation study. Instead we report the problems evolving from different choices of the above mentioned parameters.

6.8.1 Test statistics

The test on time-varying effects is implemented in three versions. The default version ($w = 0$) does not differentiate between regions with large and small variance. Large variances, though, represent uncertainty in estimates and often correspond to few data support. In our simulation study this is supposed to occur for large survival times, where observations are rare. As we believe departure from PH in practical applications to be more relevant when many patients are at risk, we prefer the alternative variance weighted test statistic ($w = 1$). A third version of this test statistic additionally removes the end points on both edges ($w = 2$). The choice of the test statistic, i.e. the weighting scheme, appears to have large influence on test decisions as shown in Section 4.2, where application of the three test statistics in a multivariable analysis results in three different models.

6.8.2 Impact of the bandwidth

Another important parameter of the approach is the bandwidth of the kernel smoother. As for all smoothing methods, the specification of a suitable bandwidth for the kernel smoother of the implemented routine is a challenging task, which is a great practical problem of the method (Cortese et al., 2010). The default bandwidth of the implemented program is 50% of the range of the considered observation period. Scheike and Martinussen (2004) give a

formula for the asymptotic mean-squared error optimal bandwidth. Such plug-in methods, though, are known to be questionable (see e.g. Loader, 1999, pp. 179-182 or Härdle, 1990, p. 92). They depend on the unknown $\beta''(t)$, the second derivative of the time-varying effect $\beta(t)$, which itself must be estimated with a suitable bandwidth. Hence, the problem is simply moved from the estimation of $\beta(t)$ to the estimation of $\beta''(t)$. Martinussen et al. (2002) suggested to try different bandwidths, but do not give advice on situations where results between bandwidths differ largely. Scheike and Martinussen (2004), though, use the smallest possible bandwidth for which the algorithm converges in the estimation of an initial estimate for $\beta''(t)$. A similar choice seems also sensible for the estimation of $\beta(t)$ in practical applications.
Varying the bandwidth of the estimation algorithm can lead to different conclusions about time-varying effects, as pointed out in Section D.3 in the Appendix. The impact of the bandwidth on the shape of final time-varying effects $\hat{\beta}(t)$ using `locfit` is even more pronounced. The default bandwidth of 0.7, though, is deemed to be a sensible choice (Section D.3).

6.8.3 Invertibility problems

For default bandwidth, the estimation procedure issues invertibility warnings in several settings. This suggests that there are some problems in identifying the effects or with stability. If this happens rarely, it may not cause problems. If invertibility problems are more frequent, as in our simulated data sets, they can cause problems.

For example, in our simulation study we observe a considerably inflated type I error of about 34% to 68% for a nominal significance level of 1%. Application of the unweighted test version in the main setting with sample size 1000 and light censoring, results in type I error rates of 18% and 33% for binary and standard normal variable, respectively. This is in contrast to the findings of Scheike and Martinussen (2004), who present a small simulation study with two normal variables with mean 0 and standard deviation 2, one of which has a time-varying effect. For the investigated sample sizes up to 800, they observe a type I error for the unweighted test that is only slightly larger than the nominal significance level of 5%.
The inflated type I error in our simulation study may be caused by the invertibility problems. Spurious time-varying effects are introduced by instable estimates at regions, where data gets sparse. Figure 6.22 (left panel) shows the estimated cumulative effect $\hat{B}(t)$ in a data set with such a spurious time-varying effect. Shortly after time 10, estimates begin to destabilise. At time 13.6, the last subject with covariate level X=1 experiences an event and the estimation algorithm issues non-invertibility warnings for all subsequent event times. Due to the large distances between events (indicated at the axis of abscissae), these instable steps even gain in importance. The corresponding test process (Figure 6.22, right panel) shows a large drop in the time period with rare events. This is caused mainly by the instability of

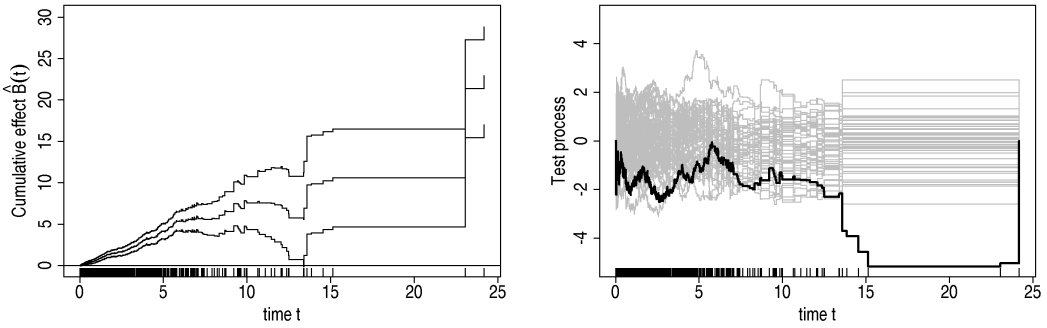


Figure 6.22: Cumulative effect function (left) and test process (right) for a spurious time-varying effect in scenario (C).

effects, but can lead to rejection of the null hypothesis of time-constant effect.

The invertibility problems occur mainly for larger times, where data gets sparse. Hence, the global bandwidth may lead to empty neighbourhoods. The best solution to this problem would possibly be a local smoothing approach within the estimation algorithm. In the context of this work, we are limited to two other possible interventions:

(i) restricting the test and estimation period, or

(ii) increasing the bandwidth.

As the test statistics of the test on PH depend on the considered time interval [0, τ], different choices for τ may lead to different test decisions (Martinussen and Scheike, 2006, p. 123). When restricting the time period under investigation to [0, 10], invertibility problems almost disappear and type I error decreases to about 6% and 1% for binary and standard normal variable, respectively. For the unweighted test, it is about 1.5-2% for both distributions. Imagine the example in Figure 6.22 to be considered on [0, 10] only. As data support in this period is sufficiently large, the reduction of invertibility problems is evident. Yet, as some time-varying effects may not become apparent till later time points, these larger times can be of particular interest. Furthermore, restriction of the test and estimation intervals results in limited comparability to FPT.

Increasing the bandwidth is another possible solution, as it may avoid empty neighbourhoods. In our simulation study, this helps to reduce the type I error for most parameter combinations. However, with binary variable, the variance weighted test statistic ($w = 1$) still shows a considerably inflated type I error. These results emphasise the impact of the chosen bandwidth on the test. Although the type I error reduces with larger bandwidth, the estimation algorithm still issues a considerable amount of invertibility warnings. Hence, re-

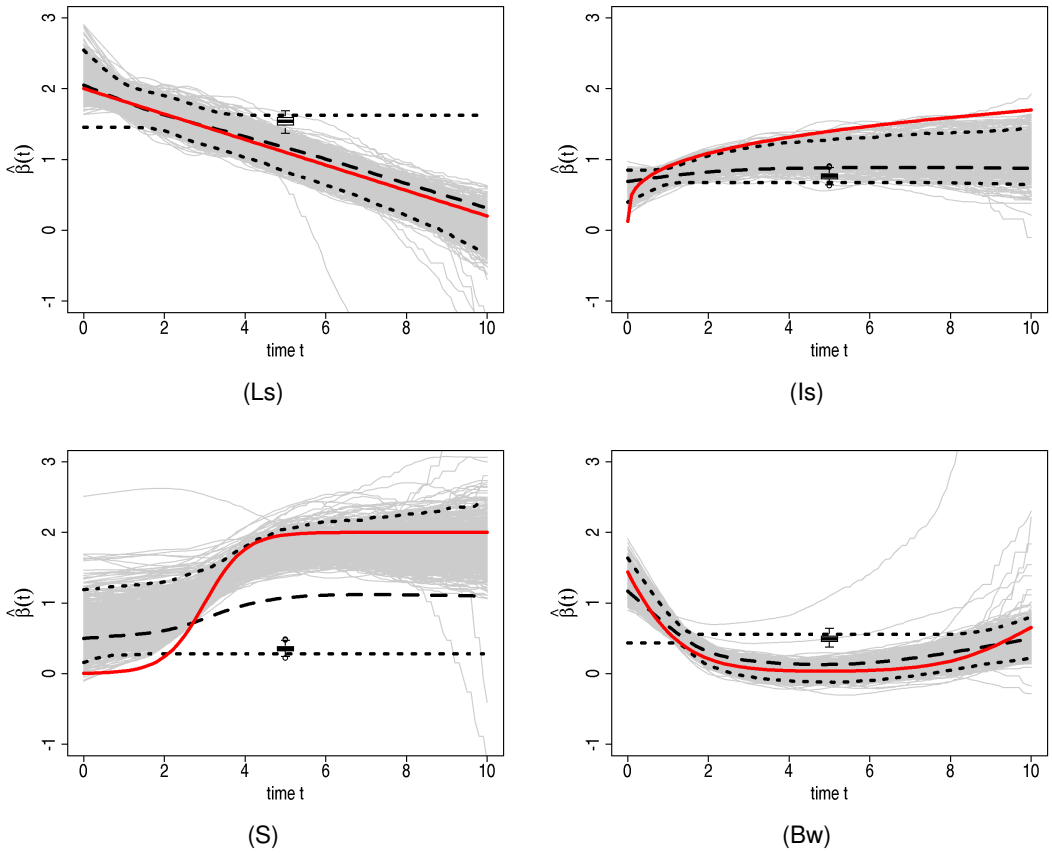


Figure 6.23: True (—) effect, estimated time-varying (—) and time-constant (boxplot) effects, pointwise mean (- -) and 95% empirical confidence intervals (· · ·) for the Semiparametric Extended Cox model in scenarios (Ls), (ls), (S) and (Bw) with standard normal variable, $n = 1000$ and light censoring.

liability of the results is questionable. Details on the type I error and the influence of the bandwidth on estimated effects are given in the Appendix in Section D.3.

6.8.4 Flexibility of effect estimates

Despite these problems and critical choices, effect estimates are very flexible. To give an impression of the potential flexibility, Figure 6.23 shows the estimated effects for scenarios (Ls), (ls), (S) and (Bw). These plots show that the Semiparametric Extended Cox model is flexible enough to describe also complex patterns. However, one should be aware of possible locally extreme behaviour of estimates. This is unfavourable especially for simple patterns such as linear time-varying effects (Figure 6.23, top left), where too wiggly estimates may skew the underlying pattern.

Consequently, the Semiparametric Extended Cox model is a flexible approach, but requires its parameters to be chosen with care, which requires sufficient experience and effort. Furthermore it is sensitive to peculiarities in the data, which may introduce problems in estimation and stability of effects.

6.9 Summary

Summarising, FPT is shown to outperform a simple CoxPH model in nearly all investigated settings. The test procedure approximately holds the nominal significance level in case of PH. With time-varying effects, the performance strongly depends on the specific scenario and parameter settings. In general, the power is adversely affected by sample size and the proportion of censoring.
With binary variable, the test procedure has difficulties in detecting some time-varying patterns such as the linear decreasing effects or the increasing effects, but shows reasonably good power for most other scenarios. It is known that considering FPs of degree 2 may reduce the power of the hierarchical closed test procedure, which might explain the poor crude type II error in some of the scenarios. A small investigation on this issue shows ambiguous results. While omitting the test on FP2 leads to a considerably improved power for the increasing effects, but results for the linear decreasing effect hardly change.
For standard normal variable, using the same baseline settings and effects than with binary variable, the performance of FPT is considerably better. In general, the power is extremely good, except for some difficulties with the increasing effects. These good results may be explained by the distributions of X and T. For extreme X values, often smaller values of T are generated. As discussed in Section 6.3, the combination of extreme event times and a (strong) time-varying effect may influence the likelihood considerably. Hence, these extreme times help in identifying time-varying effects. The same rationale, though, explains convergence problems for some FP combinations that may occur in situations with too many extreme times. However, these convergence problems are shown to be unproblematic with respect to final effect estimates. Simultaneously, an exclusive relation between large X values and small event times may introduce frailty effects, resulting in a systematic underestimation of effect sizes as observed for the increasing effects.

Heterogeneity of selected functions decreases with increasing effective sample size. This is also reflected in the qualitative type II error. The scenarios (Ls), (Lw), (Is) and (Iw), which already have large crude type II error rates, necessarily show a large qualitative type II error. Yet, even with normal variable, the qualitative type II error in these scenarios is moderate to large. A possible source of inflated qualitative type II error is the choice of $\log(t)$ as a default time transformation. A small investigation indicates that usage of the default transformation

may lead to a considerably inflated qualitative type II error, if the true effect differs from the logarithmic shape. For $\log$-like effects, on the contrary, the default transformation achieves its purpose of stabilising selected effect functions and controls the qualitative type II error.
Investigation of scenarios (Ls)/(Lw) and (Ds)/(Dw), which reflect similar time-varying patterns, but different effect size, shows that the crude type II error and the shape of selected functions are not affected by the strength of the effect, except for single parameter settings. The improvement in ABCtime and prediction error relative to CoxPH, on the contrary, reduce for the weaker effects.
The pointwise means of estimated effects are, for reasonable type II error rates, in most scenarios close to the true effects. The exceptions are the sigmoid and bathtub effects, where FPT is not flexible enough to describe the complex patterns exactly. These effects explore the limitations of the FPT approach, because they are no members of the FP class. Hence, FPT is not able to describe them correctly. With sigmoid effect (S), for example, FPT detects the presence of a time-varying effect in each simulation run, but similarly fails to describe it each time. Correspondingly, this scenario shows the largest spectrum of selected curve shapes. Yet, the gain relative to time-constant CoxPH effects is distinct. For the bathtub effects, the performance of FPT is much better, although the gain relative to CoxPH is less pronounced, due to the shape of the true effects.

In the presence of time-varying effects, FPT shows an improvement in terms of pABCtime as well as prediction performance compared to CoxPH, if the crude type II error is not too large. In parameter settings with large crude type II error, the prediction performance is naturally similar to the CoxPH model.

The multivariate investigations show no convincing differences between settings with and without correlation, despite of a slightly improved PEC in the settings with correlation. On the covariate level, results for the decreasing effect (Ds) resemble those in the univariate investigations, while the increasing effect (Is) even improves in terms of the type II error. Hence, for the increasing effect, FPT reveals its full strength in the multivariable investigation with improvement in pABCtime compared to CoxPH in all settings.

Chapter 7

Discussion

In the analysis of larger studies with long-term follow-up, standard techniques such as the Cox model (Cox, 1972) may be inappropriate due to violation of the PH assumption caused by time-varying effects. Ignoring the presence of these time-varying effects results in incorrect models with misleading effect estimates and possibly false conclusions thereof. However, beyond detecting the time-dependency of effects, appropriate modelling of their shape is at least as important, because 'wrong' shapes of time-varying effects can lead to false conclusions just as well as erroneously assuming PH.

The literature on methods for testing and modelling of time-varying effects is broad. Some basic 'traditional' approaches are artificial time-dependent covariates or partitioning of the time axis. The former has already been proposed by Cox (1972) in his original paper and uses a pre-defined transformation of time which is the basis for testing on non-PH. Yet, this method depends strongly on the choice of the time transformation. Partitioning of the time axis, or piecewise constant effects, is another popular technique. It partitions the time-axis into several time-intervals and estimates separate (time-constant) effects for each of these intervals. However, this method is highly dependent on the number and position of intervals and requires a sufficient amount of events in each interval.
Recently, Schemper et al. (2009) proposed average hazard ratios by weighted Cox regression for situations with very small sample size or high-dimensional data, where more complex models incorporating time-varying effects may be less powerful.

Besides these approaches, some more advanced techniques have been proposed such as splines and fractional polynomials. Several of the approaches have theoretical or technical drawbacks, such as the absence of multivariable model building strategies or selection of time-varying effects, missing supply of programs or poor usability. Furthermore, we are not aware of larger simulation studies on properties of the approaches or comprehensive comparisons of different techniques that could guide to appropriate tools for selection and

modelling of time-varying effects.

One possible reason for the lack of simulation studies may be the complexity of such a task, which already starts with the generation of appropriate data. The standard method for simulation of survival data under the PH assumption (see e.g. Bender et al., 2005) is in general not applicable with time-varying effects, as the integral and inverse cannot be solved analytically any more.
Alternative proposals allowing for time-varying effects are rare and we deem none of them optimal for various reasons. Hence, we generalised the inversion method (Bender et al., 2005) which is simpler and more intuitive. This generalised algorithm is based on numerical inversion and integration and thus allows the simulation of survival times for settings including time-dependent components.

To give guidance on different techniques for modelling time-varying effects, we compared five recent approaches based on promising techniques, such as splines and fractional polynomials, including frequentist as well as Bayesian inference. As Hand (2001) notes: “It may be possible for an expert to tune method A to achieve results superior to method B, but what we really want to know is, whether someone untutored in the niceties of method A can do this. Or does method B, presented as a black box and requiring no tuning, generally outperform an untuned method A?” Therefore, we used the default settings for the different parameters of the approaches when applying them to a real-life example and a simulated data set to gain insights into the performance and usability of the approaches.

Since it does not include selection of time-varying effects, the Reduced Rank model considerably overfits the data at hand and does not even convince in terms of the apparent prediction error. The Dynamic Cox model, on the contrary, performs quite well with respect to selecting and modelling time-varying effects as well as predicting survival probabilities. Unfortunately, the implemented algorithm is not able to handle data sets exceeding a few hundred observations without categorising survival time. Furthermore, it is implemented in an older `S-Plus` version and is not compatible to current versions of the software, which limits the practical applicability of the method. This approach can also be considered as a special case of the FPT approach on which the main focus of this work lies.

The Semiparametric Extended Cox model is as a non-/semiparametric approach based on cumulative regression functions very flexible in estimating even complex time-varying effects. Yet, the choice of the test statistic and the bandwidth for the kernel smoother has a large impact on results. Especially the decision on the bandwidth is a great practical problem (Cortese et al., 2010). As the implemented program provides only the final estimates of the cumulative effect functions, an additional transformation is required by the user to obtain the estimated time-varying effects. In our simulation study, we were forced to exclude this approach, as it showed severe instability, which could not be sufficiently repaired by adaptation

of the smoothing parameters. These findings may be due to specific patterns in the data, but demonstrate the need for a critical check of the derived model to exclude a distortion of the algorithm by special patterns in the data or suboptimal parameter choices.

The Empirical Bayes model can easily be applied also by users less experienced in Bayesian modelling, as it does not require decisions on the mixing of Markov chains and shows similar performance than an equivalent full Bayes approach (Kneib and Fahrmeir, 2007). The spline based technique for modelling of time-varying effects is very flexible and, as the smoothing parameters are simultaneously estimated, does not require user based decisions on these critical parameters. Furthermore, the default number and location of knots seem to work pretty well in our examples. Yet, the approach can be rather time-consuming, especially with a larger number of covariates. As it does not include automated selection strategies, the candidate models are compared with respect to a goodness-of-fit criterion such as the Akaike information criterion (AIC) or Bayesian information criterion (BIC). This fact makes it difficult to evaluate the approach in simulation studies where automated selection procedures are required.

The FPT approach is due to the limited class of FPs potentially less flexible than nonparametric or spline based approaches. The global FP functions are likewise the weakness and the strength of this approach. They provide concise formulae of the functional form of time-varying effects, which are easy to interpret. The global nature also ensures estimates that are mostly robust to locally extreme patterns in the data and thus helps to avoid overfitting and allows better generalisability. The same characteristic, though, may let relevant local patterns with rapid changes over a short time-period remain uncovered with FPs of maximum degree 2. Yet, FPs provide sufficient flexibility to describe a broad variety of functional forms (Royston and Sauerbrei, 2008, chap. 4.5). Govindarajulu et al. (2007) even found that FPs are least biased in some settings compared to several competitive smoothing methods (e.g. splines) for fitting non-linear covariate effects. In a similar framework Holländer and Schumacher (2006) show that FPs perform best in terms of type I error and mean absolute error. Within the scope of time-varying effects, Ng'Andu (1997) shows that the time-dependent covariate test, on which the FPT algorithm is based, is one of the best among several alternative tests on the PH assumption.

However, all investigated approaches may show boundary artefacts in the estimated time-varying effects. Kernel smoothers and splines of degree $\geq$2 are known to be prone to extreme behaviour beyond the boundary knots, while the FPs may show artefacts due to their global nature. For the latter especially the default transformation $\log(t)$ may show artefacts towards zero. Being aware of this characteristic, analysts would not change their interpretation of the effect. For example, a logarithmic decay of the effect (converging to infinity for $t \downarrow 0$) basically indicates a very strong initial effect that diminishes strongly within

a short time.

To sum up, all approaches under investigation have their assets and drawbacks. As the Empirical Bayes model, the Semiparametric Extended Cox model and FPT all perform quite well in the examples and do not suffer from severe theoretical or technical problems, they seem preferable. Yet, the Semiparametric Extended Cox model has been developed mainly for testing purposes. With a special emphasis on the modelling of time-varying effects, this approach may be less adequate. Furthermore, to our personal experience the Empirical Bayes model and FPT are more user-friendly and less sensitive to parameters set by the user. If the true effect shows considerable curvature within a short time-period, the global functions of the FPT approach will likely be not flexible enough to describe this effect and the spline-based Empirical Bayes model may be preferable. On the contrary, the latter approach may be more prone to locally extreme behaviour in the data and may have some drawbacks in interpretability and generalisability of estimated effects.

After thus exploring and comparing available proposals in examples, we turned to the second focus of this work, the assessment of the properties of the FPT approach. We conducted a large simulation study based on a variety of different time-varying effects, which are of potential interest in medical applications. For the univariate settings, we simulated data for ten different effects, up to four different sample sizes, varying proportion of censoring and two distributions of the variable (binary and standard normal). The simulation study shows that FPT approximately holds the nominal significance level in case of PH. The power, on the contrary, strongly depends on the specific effect and parameter settings. In general, it improves in settings with standard normal variable compared to binary variable and with increasing sample size. The time-varying effects under investigation include members of the FP class, similar functions and complex non-FP functions, which are useful for assessing the performance of FPT in situations where the true effect is too complex to be fully recovered. For the latter class of functions, the test on the presence of time-varying effects performs extremely well. As expected, FPs of maximum degree 2 are not flexible enough to describe such curvatures. FPs of higher degree could solve this problem, but at the cost of much higher computational effort and increased model complexity which is not always desirable. However, estimated effects show at least the general trend of the true effects in most situations. For the non-linear decreasing effects which are similar to the default transformation, FPT performs extremely good in terms of both testing and modelling. For other shapes, it shows some difficulties. These results are mainly due to two characteristics of the FPT algorithm: (i) the test on FP2 functions in the hierarchical closed test procedure and (ii) the choice of $\log(t)$ as a default transformation. The test on FP2 functions implies wasting degrees of freedom, if the true function is less complex. Restricting the algorithm to FPs of maximum degree 1 improves the power for the increasing effects. Hence, if subject matter knowledge suggests a simple monotone shape of the underlying effects, restricting

the maximum complexity may improve the power of the algorithm. Yet, this restriction can similarly lead to a loss in power if the true effect is more complex. Furthermore, the default transformation $\log(t)$ was detected as a source of inflated qualitative type II error, if the true effect differs from the logarithmic shape. For $\log$-like effects, on the contrary, it achieves its purpose of stabilising selected functions and controls the qualitative type II error.

The multivariable settings contain five variables, two of which have a time-varying effect, and vary in the proportion of censoring and the correlation structure. In these settings, the algorithm similarly holds the nominal significance level. While the results for the decreasing time-varying effect are analogue to the univariate investigation, the power even improves for the increasing effect. These results suggest that the algorithm reveals its full strength particularly in multivariable analyses. Our investigations reveal no indication of a loss in power for correlation between variables with and without a time-varying effect.

Summing up, the overall power of FPT is satisfactory and the algorithm performs well in modelling most time-varying effects. Furthermore, it is shown to be superior to standard CoxPH analyses in terms of ABCtime and prediction error in nearly all parameter settings, even if the true effect is too complex to be fully described by FPs of maximum degree 2. Although the individual effect estimates vary more or less, depending on the scenario, the pointwise mean over all estimates is in most scenarios reasonably close to the true effect.

As a third aspect of this work, we investigated the stability of the FPT algorithm by using bootstrap replications of the Rotterdam breast cancer series. Comfortingly, this revealed stable mean estimates. Although the majority of selected effects in general is in good agreement with the true effect (in the simulation study) and the reference effect in the original data set (bootstrap investigation), individual effects may show deviant behaviour. This suggests to base the selection and modelling of time-varying effects on bootstrap techniques to account for model selection uncertainty. We proposed a modification of the FPT algorithm, called BootstrapFPT, that selects time-varying effects only, if they are significant in more than a pre-specified amount of bootstrap samples and/or out-of-bag samples. The shape of the final time-varying effect may, for example, be chosen as the most frequently selected FP transformation or the pointwise mean over all significant time-varying effects. This strategy is in the style of bagging and model averaging techniques. For example, Sabanés Bové and Held (2010) recently proposed a model averaging approach to account for model selection uncertainty for Bayesian FPs in modelling non-linear covariate effects in a linear regression framework. They average over the set of possible models weighted by the posterior model probabilities instead of relying on the model with the highest posterior probability only. However, averaging the individual effects, rather than the complete model, allows better interpretability of single effects. Although first results in the Rotterdam breast cancer series are promising, the BootstrapFPT approach requires further investigation and

fine-tuning to allow reliable conclusions about its performance.

This leads us to outstanding work and topics of interest warranting further research, such as a potential improvement of the FPT approach by allowing for updating of FP functions in analogy to Berger et al. (2003), who use a backfitting-type procedure. The FPT approach utilises a forward selection approach for selecting time-varying effects in multivariable settings. Once a time-varying effect has been selected, its FP powers remain fixed and are not updated any more. The backfitting-type procedure of Berger et al. (2003), on the contrary, iteratively updates the FP functions, adjusting for all other (potentially time-varying) effects, until the selected FPs do not change any more. Such a modification may further improve the FPT approach.
Another future goal is the extension of the simulation study to the complete MFPT algorithm (Sauerbrei et al., 2007). This algorithm combines the MFP approach for selection of covariates and functional forms of covariates with the FPT approach for modelling time-varying effects. Such investigations are necessary to assess interactions between selection and modelling of time-varying effects and inclusion of covariates and non-linear functional forms. Mismodelling the latter issues can result in spurious time-varying effects. Hence, a good model building strategy should address all three issues.
Among the five approaches under investigation, such an extension has been proposed only for FPT, the Empirical Bayes model and the Semiparametric Extended Cox model. However, with the latter the focus lies on mere testing, rather than modelling of a potential non-linear effect. The Empirical Bayes model and FPT also tackle the modelling task. Yet, especially with a large number of potential covariates, the model building procedure for the former may become extremely time consuming, as it considers all three modelling variants simultaneously. The MFPT approach, on the contrary, uses a hierarchical structure dealing first with variable selection and identification of non-linear effects before investigating time-varying effects, which gives a clear priority to non-linear effects. However, Binquet et al. (2008) argue that such a step-wise procedure may lead to biased estimates and/or incorrect conclusions and hence propose to consider both aspects simultaneously, as proposed by Abrahamowicz and MacKenzie (2007). They state that a priori testing on log-linear effects may induce (i) an increased number of spurious non-linear effects and/or (ii) incorrect exclusion of covariates whose time-varying effects appear non-significant when ignoring the time-dependency. The latter issue is directly addressed by step 2 of the MFPT algorithm, which restricts the investigation period to a short-term interval to add effects that have been overlooked for exactly this reason. However, we cannot disprove the former claim theoretically. Investigation of this point would be subject of the aforementioned simulation study.

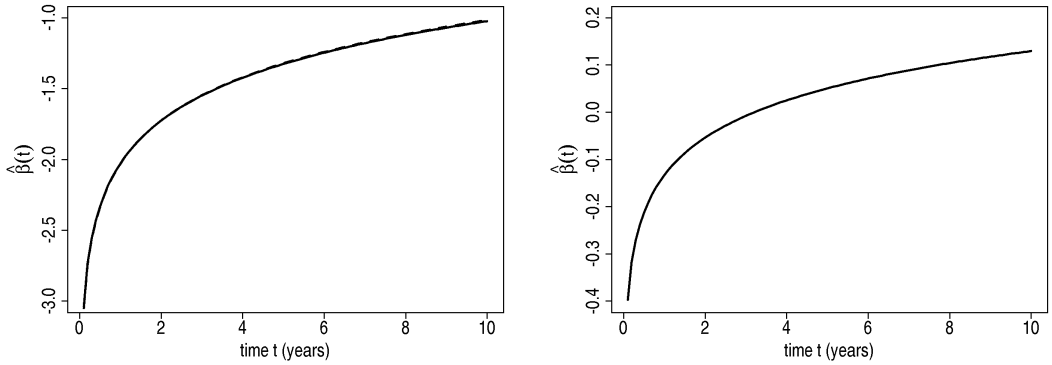
Over all, investigation of time-varying effects is most important in large studies with long-term follow-up and several covariates. In addition, other modelling issues such as selection of covariates and non-linear covariate effects have to be addressed. This can be done, for

example, by the first two steps of the MFPT algorithm. Yet, the focus of this thesis is on the detection and modelling of time-varying effects for a given model.
The simulation study on properties of the FPT approach shows that the algorithm yields satisfactory power for stronger time-varying effects. The pointwise means of effect estimates are reasonably close to the true effects. Even if the time-varying effects are too complex to be fully recovered by FPs, the performance of the FPT procedure is still superior to a standard CoxPH model, simply because of accounting for the general trend of the time-varying effect. For several time-varying effects, a forward selection procedure is proposed which may be further improved by bootstrap based selection techniques and/or a backfitting-type algorithm. In a comparison of five recent approaches for modelling time-varying effects in single examples, the only serious competitor to FPT in terms of selection and modelling of time-varying effects as well as usability seems to be the Empirical Bayes model. Consequently, we deem the FPT approach a promising, flexible and easy to use technique for selection and modelling of time-varying effects with satisfactory properties.

Appendix A

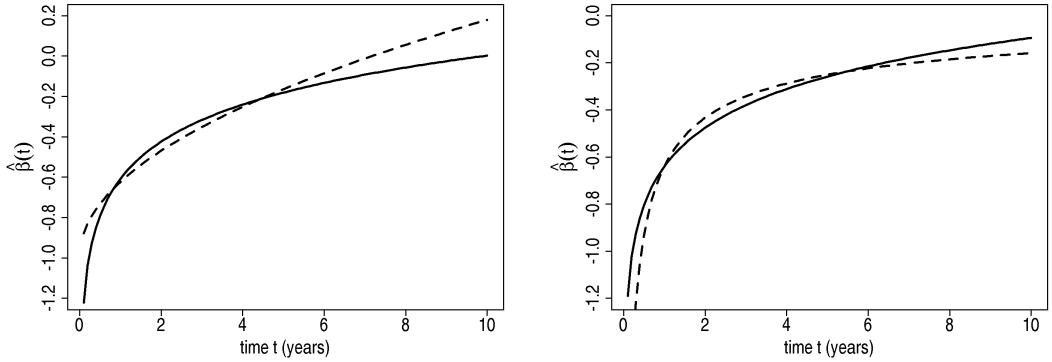
Results of the comparison of different approaches

This chapter includes supplementary material to the analyses of the Rotterdam breast cancer series and the simulated data example presented in Chapter 4.1.3.
Figure A.1 contrasts the time-varying effects selected by MFPT with and without default time-transformation $\log(t)$ in the analysis of the Rotterdam breast cancer series. All subsequent tables and figures give additional information about the selection procedures of the five competitive approaches in a simulated data set as presented in Chapter 4.2.



X_{5e}^2: virtually identical functions

$\log(X_6)$: identical functions



X_8: different function, shape differs slightly

X_9: different function, similar shape

Figure A.1: Comparison of the FPT algorithm with (—) and without (- -) default transformation $\log(t)$.

	best FP2		best FP1		default	p values (test vs. best FP2)		
Variable	powers	deviance	power	deviance	deviance	constant	default	best FP1
Iteration 1								
$\boldsymbol{X_1}$	-1 -1	**113.297**	0	97.490	97.490	0.0000	0.0012	**0.0004**
X_2	-2 0	22.610	0.5	18.624	15.851	0.0002	0.0800	0.1362
X_3	-0.5 -0.5	7.420	-2	2.281	0.298	0.1153	0.0681	0.0766
X_4	3 3	5.349	2	4.242	1.214	0.2533	0.2472	0.5748
X_5	-1 3	9.311	-0.5	5.868	5.197	0.0538	0.2495	0.1788
Iteration 2								
$\boldsymbol{X_2}$	-2 0	**20.016**	0.5	16.038	13.264	**0.0005**	0.0802	0.1368
X_3	-0.5 -0.5	9.938	0.5	4.098	3.055	0.0415	0.0757	0.0539
X_4	3 3	4.921	2	4.065	0.595	0.2955	0.2284	0.6520
X_5	-0.5 3	16.278	-0.5	13.118	11.582	0.0027	0.1955	0.2060
Iteration 3								
X_3	1 2	10.144	0.5	4.670	3.637	0.0381	0.0894	0.0648
X_4	3 3	6.613	2	5.644	1.600	0.1578	0.1708	0.6160
X_5	-0.5 3	**12.448**	-0.5	9.868	7.719	0.0143	0.1927	0.2752

Table A.1: Forward selection based on deviance differences for FPT in a simulated data set. Covariates for which time-varying effects are selected are in bold.

Variable	best power(s)	p value
Iteration 1		
$\boldsymbol{X_1}$	**-2**	**<0.001**
$\boldsymbol{X_2}$	**0**	**<0.001**
X_3	0	0.038
X_4	2	0.035
X_5	3	0.234
Iteration 2		
$\boldsymbol{X_1}$	**-0.5**	**<0.001**
$\boldsymbol{X_2}$	**0**	**<0.001**
$\boldsymbol{X_3}$	**0**	**0.007**
X_4	2	0.036
X_5	-0.5	0.134

Variable	best power(s)	p value
Iteration 3		
$\boldsymbol{X_1}$	**-0.5**	**<0.001**
$\boldsymbol{X_2}$	**0**	**<0.001**
$\boldsymbol{X_3}$	**0**	**0.007**
X_4	2	0.036
X_5	-0.5	0.134

Table A.2: Backfitting algorithm based on likelihood ratio tests for the Dynamic Cox model in a simulated data set. Covariates with significant time-varying effects are in bold.

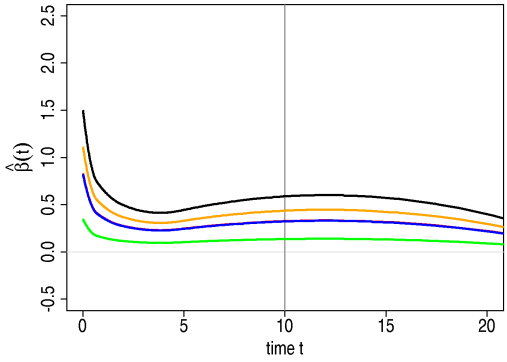
Variable	AIC_c
Iteration 0	
PH	2767.53
Iteration 1	
$\boldsymbol{X_1}$	**2667.20**
X_2	2751.37
X_3	2767.36
X_4	2766.98
X_5	2766.25
Iteration 2	
$\boldsymbol{X_2}$	**2654.99**
X_3	2663.29
X_4	2667.26
X_5	2666.08

Variable	AIC_c
Iteration 3	
$\boldsymbol{X_3}$	**2650.93**
X_4	2651.90
X_5	2655.80
Iteration 4	
$\boldsymbol{X_4}$	**2647.38**
X_5	2651.66
Iteration 5	
X_5	2648.79

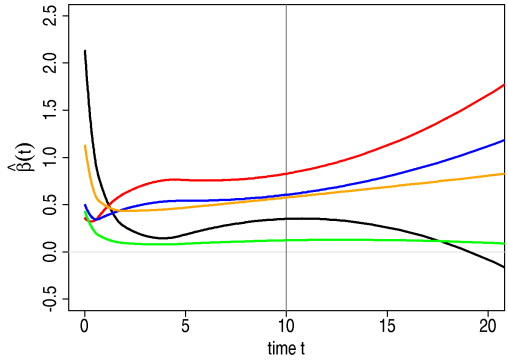
Table A.3: Forward selection procedure based on the AIC_c for the Empirical Bayes model in a simulated data set. Covariates for which time-varying effects are selected are in bold.

Variable	p value
Iteration 1	
X_1	0.000
X_2	0.000
$\boldsymbol{X_3}$	**0.167**
X_4	0.001
X_5	0.102
Iteration 2	
X_1	0.000
X_2	0.000
X_4	0.000
$\boldsymbol{X_5}$	**0.011**
Iteration 3	
X_1	0.000
X_2	0.000
X_4	0.000

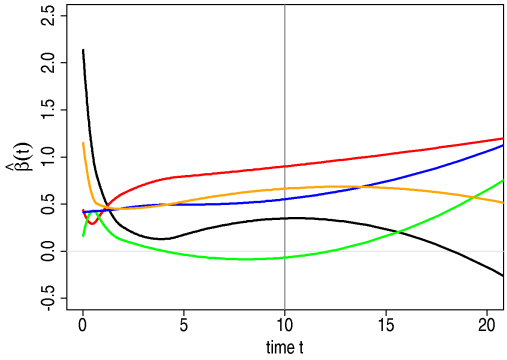
Table A.4: Backward elimination based on p value of the variance weighted Kolmogorov-Smirnov type test for the Semiparametric Extended Cox model in a simulated data set. Covariates for which time-varying effects are eliminated from the model are in bold.



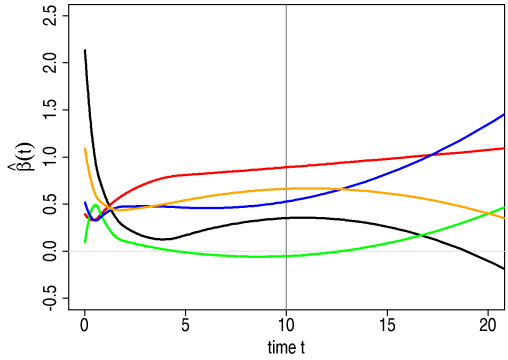
(a) Rank 1: AIC=5536.64



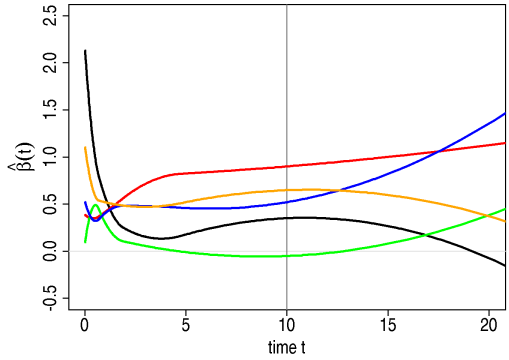
(b) Rank 2: AIC=5443.81



(c) Rank 3: AIC=5439.60



(d) Rank 4: AIC=5444.55



(e) Rank 5: AIC=5447.94

Figure A.2: Estimated effects for different ranks of the Reduced Rank model in a simulated data set for covariates X_1 (—), X_2 (—), X_3 (—), X_4 (—), X_5 (—). The rank 3 model has the smallest AIC and $r = 3$ is therefore chosen as the optimal rank.

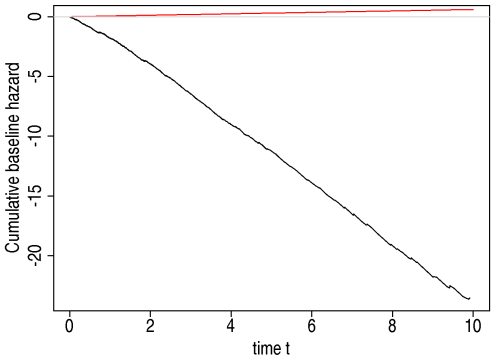


Figure A.3: True cumulative baseline hazard (—) and cumulative baseline estimated by the Semiparametric Extended Cox model (—) in a simulated data example.

Appendix B

Categorisation of survival times[1]

Some methods for modelling time-varying effects, such as the FPT algorithm, require an artificial enlargement of the data set for computational reasons. The observations under risk are split into episodes at specified time points, ideally at each event time. As the investigation of time-varying effects is most sensible with long-term follow-up and requires large data sets in order to have some power to detect interaction with time, splitting at each event time may cause technical problems. In the Rotterdam breast cancer series, for example, a split at each event time would lead to more than 2.2 million data lines. To reduce the computational effort, Sauerbrei et al. (2007) propose to 'categorise' survival times in half-year periods up to year 15 and a final period $>$15 years. This results in 31 distinct time points for categorised survival time t and a manageable 35,698 data lines.

To investigate the influence of categorisation of survival time on the three steps of the MFPT algorithm including variable selection, selection of functional forms and selection of time-varying effects, we apply the algorithm to two data sets with survival times categorised in 1, 3, 6, 12 and 24 month intervals.

In the Rotterdam breast cancer series (with uncategorised survival time and variables not centred around their mean), the final model derived by MFPT contains eight covariates, including two non-linear and two time-varying effects. Models selected on the data with survival time categorised in intervals up to 12 months are nearly identical (see Table B.1). The binary variables X_{3a}, X_{3b}, X_4, X_8 and X_9 and the continuous variables X_1 (linear), X_{5e} (power 2) and X_6 (power 0) are selected. For $log(X_6 + 1)$ and X_{3a} a time-varying effect of the type $\beta(t) = \beta_0 + \beta_1 \log(t)$ is also included. Minor exceptions from this model are observed for 12 month categorisation (power 3 instead of 2 for X_{5e}) and for categorisation in 1 month intervals (time-varying effect of X_{3a} excluded and replaced by two time-varying effects for X_{5e} and X_9). The latter difference is a result of similar deviance differences for

[1] Large parts of this section match the manuscript Buchholz et al. (2009) submitted for publication.

X_{3a} (17.518, original time) and X_{5e} (17.497, original time) in the second iteration of the third step of the MFPT algorithm. Minor changes of the data result in a larger value for X_{5e}. Even for categorisation in 2 year intervals (only eight distinct event times), most parts of the model are identical.

Using the model selected with MFPT in the original data, we compare the influence of categorisation on the eight estimated time-constant and the four time-varying (β_0 and β_1 for X_{3a} and $\log(X_6 + 1)$) coefficients. Altogether differences are small, but get larger with increasing interval length. With exception of X_{3a}, the relative difference is below 1% for coefficients when using 1 month categorisation. Differences of the four time-varying coefficients are between 2.2% and 4.2%. For half year intervals, differences are between 1.9% and 5.1% (except for X_{3b} with 12%) for time-constant coefficients and between 8.8% and 16.4% for the time-varying coefficients. With 1 year intervals, the relative difference is below 10% for most estimates, but large for some. For 2 year intervals, though, estimates differ severely. In general, time-constant coefficients estimated with categorised survival time are closer to zero than those for original time, that is the effects of factors are underestimated. On the contrary, most estimated time-varying coefficients show increased effect sizes. They also show larger deviations from original time than the time-constant coefficients, because categorised times are given by mean survival times within the categorisation intervals. That is, different categorised survival times will be assigned to subjects for varying interval widths which has a larger impact on time-varying than on time-constant effects. The time-varying effects for X_{3a} and $\log(X_6 + 1)$ are shown in Figure B.1. As can easily be seen, the functions are quite similar. For X_{3a}, the functions all lie within the pointwise confidence intervals for the original time. For $\log(X_6 + 1)$, estimates for 1 to 6 month intervals lie within the confidence intervals, while the estimated function for 12 month categorisation lies only slightly outside the confidence interval up to approximately three years and inside it for later time points. The function for 2 year intervals is not covered by the confidence interval up to approximately four years.

For very large data sets, e.g. registry data with hundred thousands of observations and many covariates, a split at each event time would produce an exploding number of records, i.e. memory problems increase considerably, and the expanded data set may be too large for the model to be fitted. This is the case for the Whitehall I study (Marmot et al., 1984), a prospective, cross-sectional epidemiological cohort study assessing risk factors for death of male British Civil Servants employed in London. This data set with 17,260 patients (2269 distinct event times) leads to 30,852,449 records for a split at each event time, requiring more than 2.7 GB of memory. With data sets of such magnitude, even larger compute servers with more than 4 gigabyte of RAM may fail to provide the required amount of memory unless they use a 64-bit operating system. Thus, for very large data sets, powerful computers with special system requirements are necessary, so that many data analysts may not be able to

Table B.1: Rotterdam data. Influence of length of time interval on selection of variables and transformations by using MFPT. For steps 1 and 2 the selected FP powers are given, "-" marking variables that have not been selected. FP powers selected in step 1 are kept fixed in step 2 (indicated by "$\sim$"). For step 3 the FP powers of selected time-varying effects are displayed, with "-" indicating that no time-varying effect was selected (constant effect). Fields marked by $\cdot$ indicate, that the variable was not selected in steps 1 and 2 and is therefore no candidate in step 3.

		Original	Interval length (months)				
		data	1	3	6	12	24
continuous factors							
X_1	Step 1	1	1	1	1	1	1
	Step 2	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$
	Step 3	-	-	-	-	-	-
X_{5e}	Step 1	2	2	2	2	3	3
	Step 2	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$
	Step 3	-	0	-	-	-	-
X_6	Step 1	-	-	-	-	-	-
	Step 2	0	0	0	0	0	0
	Step 3	0	0	0	0	0	0
X_7	Step 1	-	-	-	-	-	-
	Step 2	-	-	-	-	-	-
	Step 3	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
binary factors							
X_2	Step 1	-	-	-	-	-	-
	Step 2	-	-	-	-	-	-
	Step 3	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
X_{3a}	Step 1	1	1	1	1	1	1
	Step 2	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$
	Step 3	0	-	0	0	0	-
X_{3b}	Step 1	-	-	-	-	-	-
	Step 2	1	1	1	1	1	-
	Step 3	-	-	-	-	-	$\cdot$
X_4	Step 1	1	1	1	1	1	1
	Step 2	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$
	Step 3	-	-	-	-	-	-
X_8	Step 1	1	1	1	1	1	1
	Step 2	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$
	Step 3	-	-	-	-	-	-
X_9	Step 1	1	1	1	1	1	1
	Step 2	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$
	Step 3	-	0	-	-	-	-

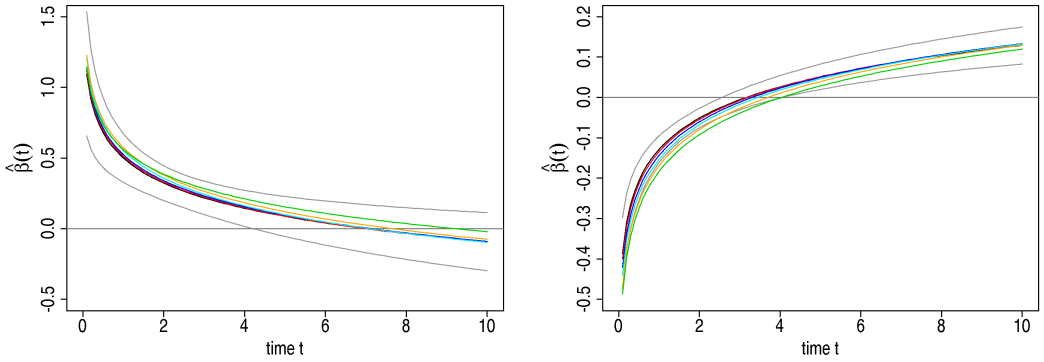


Figure B.1: Rotterdam data. Effect of different categorisation intervals on the estimated functions of the two selected time-varying effects for X_{3a} (left) and $\log(X_6 + 1)$ (right). Shown are effects estimated in the original data (—) with pointwise 95% confidence intervals (—) together with estimates based on categorised survival times in 1 (—), 3 (—), 6 (—), 12 (—), and 24 (—) months.

fit the models with uncategorised survival time any more.

With respect to selected models and parameter estimates, the MFPT procedure gives similar results for small categorisation intervals. Selected models are quite stable with differences only for single factors. Estimated effects are also very similar and effect sizes tend to be underestimated with categorised time. Even for moderately large data sets, such as the Rotterdam breast cancer series, interval lengths of up to three or six months give acceptable results with differences in estimated coefficients mainly below 5% for time-constant effects. Intervals larger than six months are not advisable any more in this data set. For the larger Whitehall data set, differences diminish (data not shown). Selected models are nearly identical for all categorisation lengths. For time-constant effects, differences in estimated coefficients are negligible for smaller intervals. Up to three or six months, differences are below 1.5%, and even for 12 and 24 months most differences remain below 5%. Time-varying effects, though, are more affected by categorisation. However, functions are still very similar for categorisations up to about three or six months. The results from these two data sets indicate, that the influence of categorisation of survival time seems to decrease for increasing sample size, and that the MFPT procedure is to some extent robust to categorisation. Thus, we believe that for extremely large data sets, where categorisation of survival time is required the most, categorisation in reasonably small intervals will allow an expansion of the data without harming estimates too much. Based on our limited experience in these two studies, we propose that about 50 to 100 distinct event times can give results with sufficient precision.

Appendix C

Details on generated survival times

C.1 Univariate settings

The distribution of simulated survival times vary over the different parameter settings, depending on the effect, the proportion of censoring and the distribution of the variable. The median survival times range between 0.8 and 6.7 (Table C.1). For binary variable, it tends to be smaller for decreasing effects, that have a strong effect initially. For normal variable, on the contrary, the median survival time is relatively similar for the different effects and varies only with varying proportion of censoring. With the chosen parameter settings for the baseline hazards, median survival times increase with increasing proportion of censoring. The median follow-up time (calculated by the reverse Kaplan-Meier method, Altman et al., 1995) simultaneously decreases from light to heavy censoring (Table C.1).

Furthermore, for most scenarios, generated uncensored event times tend to be larger for normal X, than for binary variable. To illustrate this, Figure C.1 shows several quantiles of uncensored event times. The lines connect quantiles of uncensored event times of binary

	Sample	Cen-	Effects									
X	size	soring	(C)	(Ls)	(Lw)	(Ds)	(Dw)	(Is)	(Iw)	(S)	(Bs)	(Bw)
binary	1000	no	1.5	0.8	1.4	1.3	1.7	1.5	2.2	-	-	-
binary	1000	light	2.3	1.2	2.2	2.1	2.5	2.2	3.2	3.0	3.2	2.5
binary	1000	heavy	4.5	-	-	5.0	-	3.9	-	4.4	6.6	5.8
normal	1000	no	2.2	-	-	2.1	-	2.0	2.2	-	-	-
normal	1000	light	3.3	3.0	3.2	3.1	3.3	2.9	3.0	3.1	3.2	3.2
normal	1000	heavy	6.5	5.8	6.3	6.4	6.7	5.2	-	4.8	5.5	6.3

Table C.1: Median survival time (Kaplan-Meier method) in univariate settings of the simulation study.

X	Sample size	Cen-soring	Effects (C)	(Ls)	(Lw)	(Ds)	(Dw)	(Is)	(Iw)	(S)	(Bs)	(Bw)
binary	1000	light	6.9	6.9	6.9	6.9	6.9	6.9	6.9	6.9	6.9	6.9
binary	1000	heavy	3.5	3.5	-	3.5	-	3.5	-	3.5	3.5	3.5
normal	1000	light	7.0	7.0	7.0	7.0	7.0	7.0	6.9	7.0	6.9	6.9
normal	1000	heavy	3.5	3.5	3.5	3.5	3.5	3.5	-	3.5	3.5	3.5

Table C.2: Median follow-up time (reverse Kaplan-Meier method) in univariate settings of the simulation study.

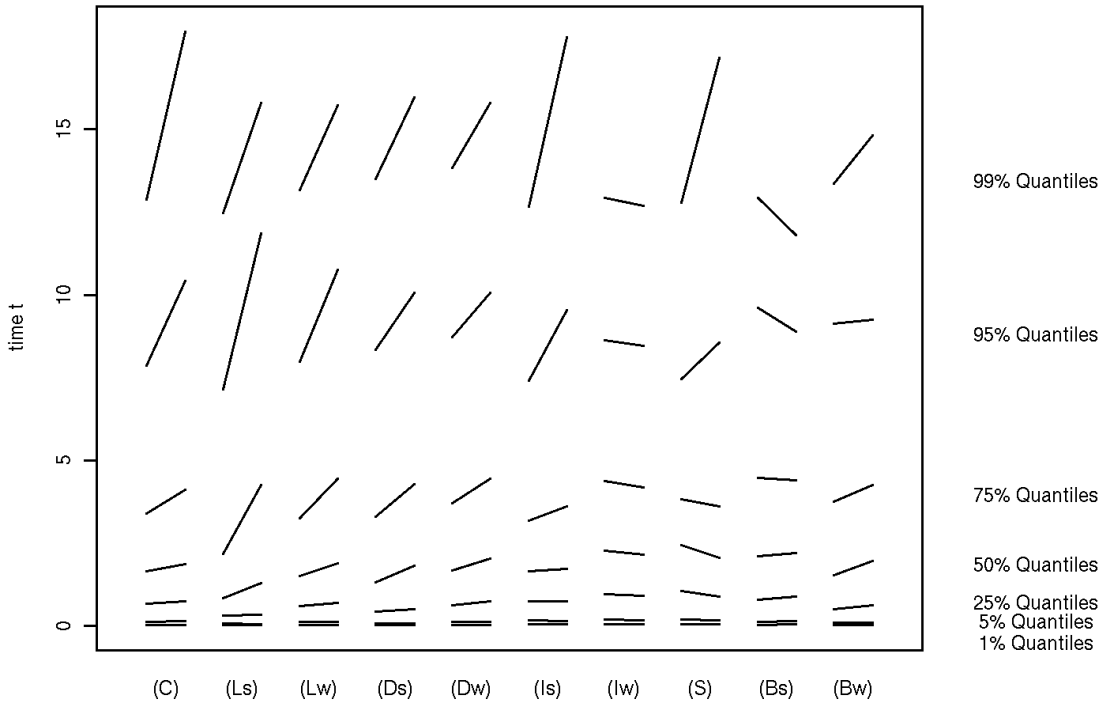


Figure C.1: Quantiles of event times for simulated data sets in the main setting ($n = 1000$, light censoring). Lines link the quantile of event times for binary X with the corresponding quantile of the same scenario with standard normal X.

---

and standard normal variables per effect. For most scenarios, generated uncensored event times tend to be larger for normal X. This can be explained by the larger variation of standard normal X values. The large majority of values scatters around zero, i.e. the impact of the variable on survival is less pronounced, leading to generation of larger survival times. Simultaneously, though, some extremely large values of X may occur. This leads to another phenomenon that can be observed for scenarios with strong effect ((Ls), (Ds), (Is), (S) and

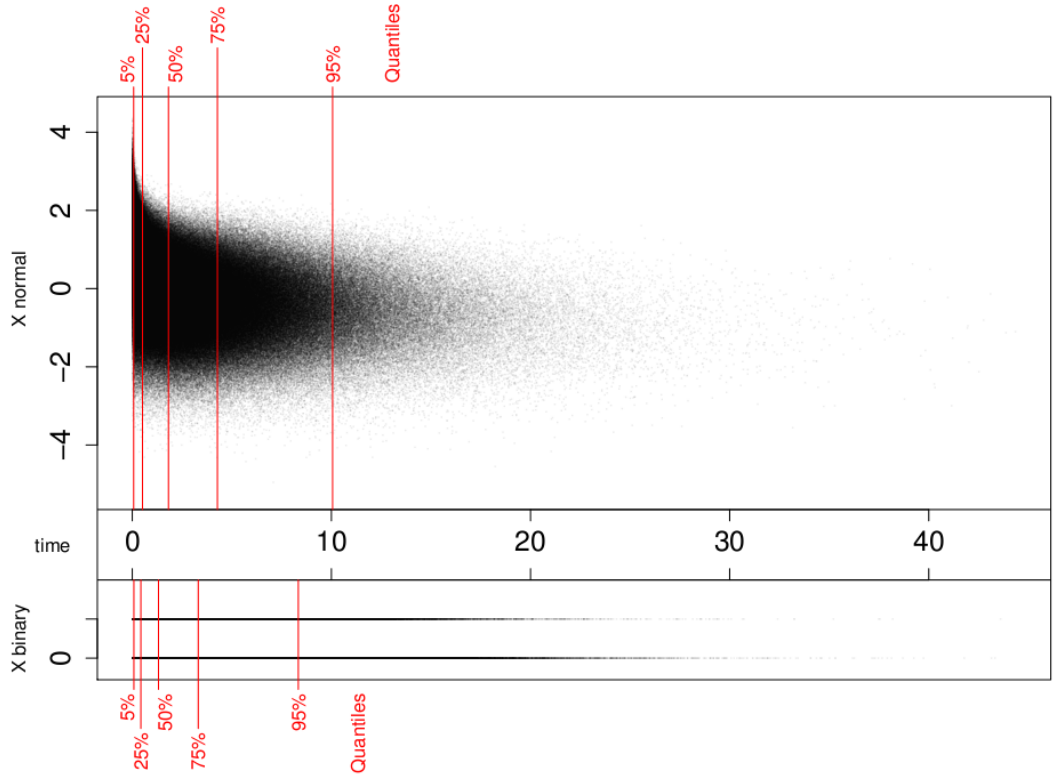


Figure C.2: Density scatterplot of uncensored event times vs. values of the covariate X (binary and standard normal) in scenario (Ds) with sample size 1000 and light censoring.

(Bw)). In addition to the general trend of increased event times for normal variable, an increased number of very small times is generated. These occur mainly for large values of X, as exemplarily shown in Figure C.2 for effect (Ds). The density scatter plot shows the combinations of uncensored event times and X for both normal and binary variable. The grey lines mark the 5%, 25%, 50%, 75% and 95% quantiles of uncensored event times, respectively. It can easily be seen, that large values of X solely involve extremely small event times. A strong effect even enforces this relationship. Small and moderate values of X on the contrary, result in generation of larger event times. Hence, the more extreme values of standard normal variables compared to binary X lead to more extreme event times on both sides.

A different behaviour is observed for the scenarios with effects lw and Bs, which increase for larger event times. For these effects, event times for normal variable are smaller than for binary X.

Changes in generated event times also affect the censoring rates. The censoring rates over all data sets differ between 20% and 35%, depending on the specific scenario. As shown in Table C.3, the smallest censoring rate of 20.3% for light censoring, is observed for the

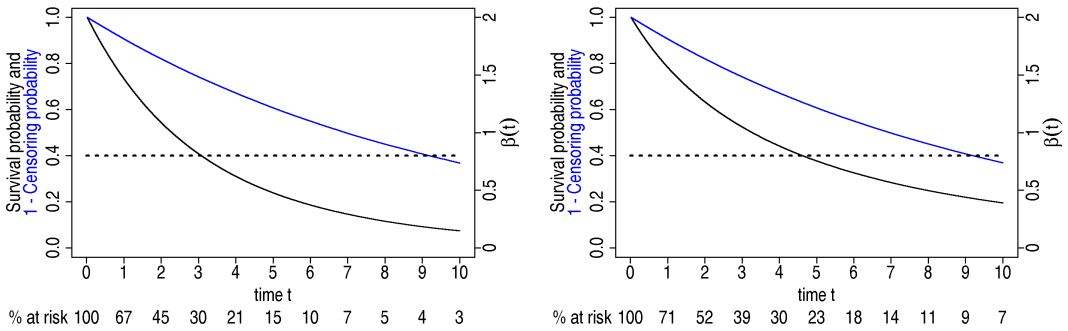
X	Sample size	Censoring	Effects									
			(C)	(Ls)	(Lw)	(Ds)	(Dw)	(Is)	(Iw)	(S)	(Bs)	(Bw)
binary	1000	light	25.8	20.3	25.6	25.6	27.9	24.8	30.2	27.9	31.0	27.9
binary	1000	heavy	57.0	-	-	57.4	-	-	-	58.0	63.6	60.4
normal	1000	light	35.2	32.6	32.8	31.7	32.7	34.8	30.1	34.9	31.2	32.0
normal	1000	heavy	64.7	58.5	63.0	62.8	65.2	60.9	-	61.8	60.9	63.4

Table C.3: Censoring rate over simulated data sets per scenario (in %) in univariate settings.

combination of binary variable and strong linear decreasing effect (Ls), which corresponds to a very strong average population effect of 1.73 (Table 6.1). The largest proportion of censoring (35.2% and 64.7% for light and heavy censoring, respectively) occurs for constant effect (C) and standard normal variable. In general, censoring rates are larger for standard normal variables. This can be explained by the larger event times generated for these scenarios. Because the censoring times are unaffected by the effect and covariate values, the proportion of censoring increases.

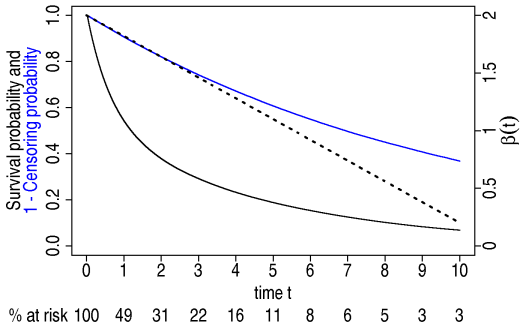
Figures C.3 and C.4 show the survival probabilities and the probabilities of not being censored (reverse Kaplan-Meier) for the different effects with sample size 1000 and light censoring. The censoring distribution is identical in all scenarios, while the event distributions change depending on the distribution of X and the effect. In general, survival probabilities for standard normal variable tend to be larger than for binary X, as explained before. The decrease in survival probability early in time, is more or less pronounced, depending on the effect. Effects of same type, i.e. linear decreasing, non-linear decreasing, increasing or bathtub, show very similar survival and censoring distributions.

For constant effect (C), the decrease is moderate (Figures C.3a and C.3b). Effects like (Ds) and (Dw), which diminish quickly, or (Iw) which rises slowly, show larger survival probabilities with a less pronounced initial decrease. Similar survival curves are observed for the bathtub effects (Bs) and (Bw), which are comparable to the decreasing effects early in time. The increase later in time has less influence on survival probability, which already is rather low. A steeper increase than for the constant effect can be observed for the linear decreasing effects (Ls) and (Lw) and the effect (Is), which rises quickly. The survival probability for the sigmoid effect (S), initially shows a moderate decrease, but steadily falls to a low level due to the sudden increase of the effect up to year five.
In general, with the current choices of baseline hazards, the survival probability tends to be largest for scenarios with heavy censoring, followed by the data sets with light censoring. Uncensored settings show the smallest survival probability. Survival and censoring distributions are exemplarily shown for effects (C), (Ds) and (Is) with standard normal variable (Figure C.5).

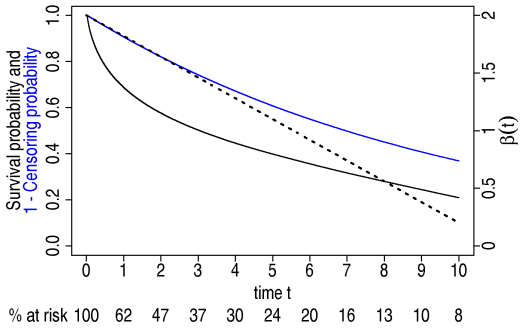


(a) (C) X binary

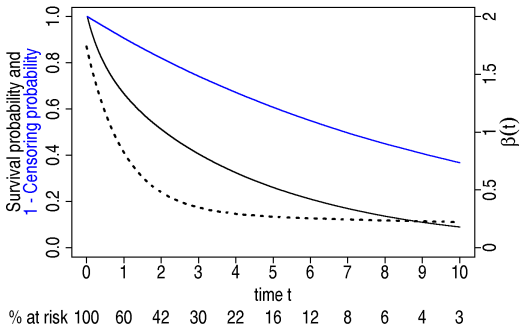
(b) (C) X normal



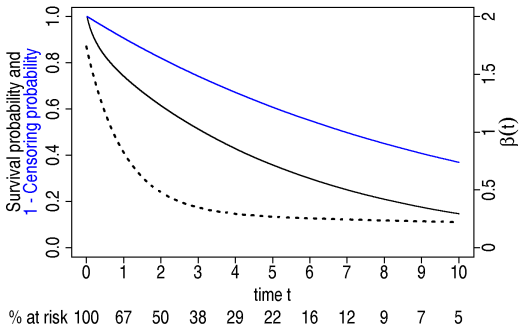
(c) (Ls) X binary



(d) (Ls) X normal

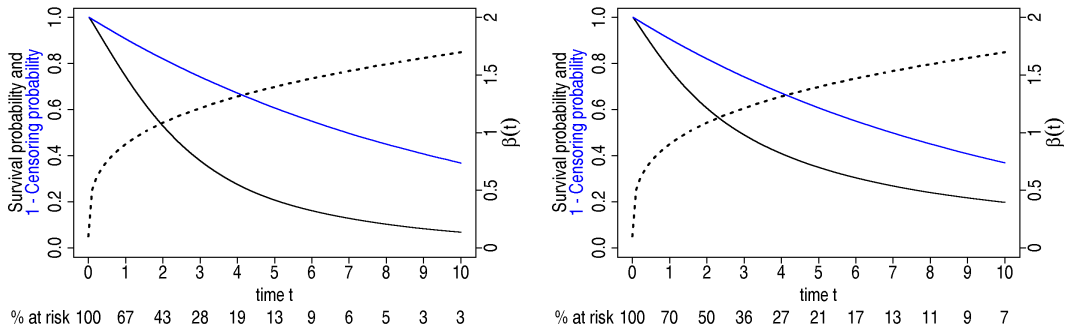


(e) (Ds) X binary



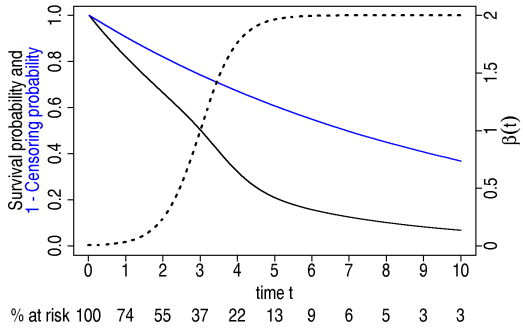
(f) (Ds) X normal

Figure C.3: Survival probability (—) and 1-censoring probability (—) for effects (C), (Ls) and (Ds) (· · ·) with light censoring.

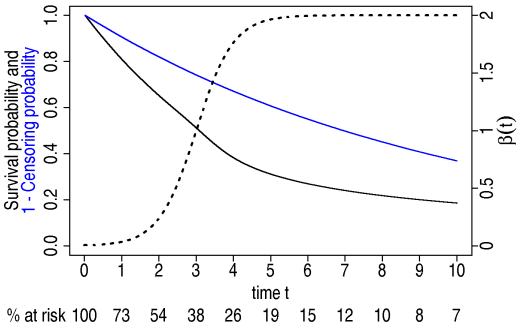


(a) (ls) X binary

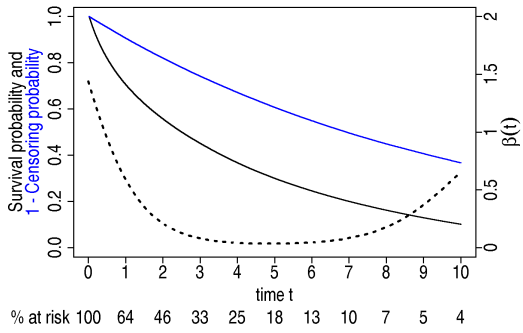
(b) (ls) X normal



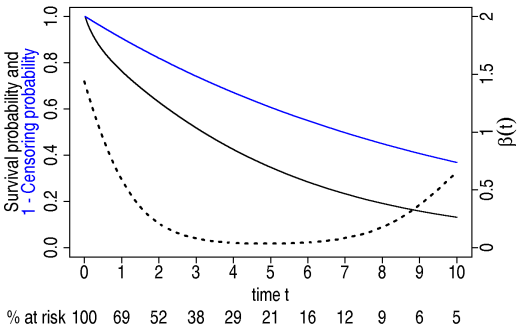
(c) (S) X binary



(d) (S) X normal

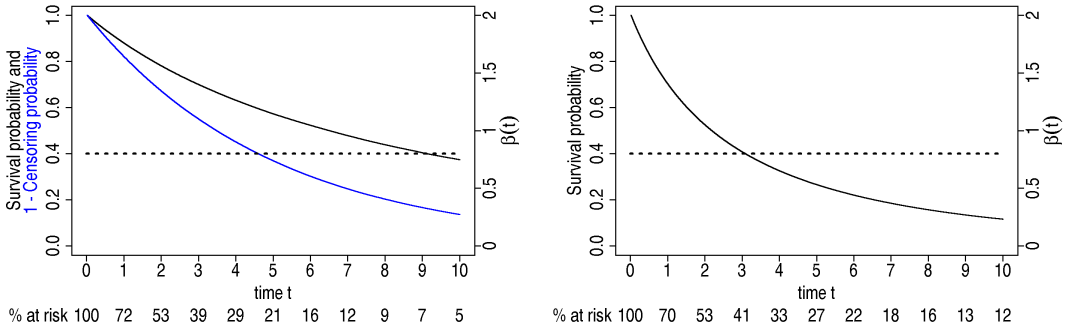


(e) (Bw) X binary



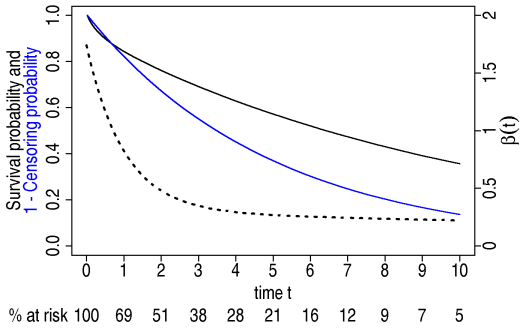
(f) (Bw) X normal

Figure C.4: Survival probability (—) and 1-censoring probability (—) for effects (ls), (S) and (Bw) (· · ·) with sample size 1000 and light censoring.

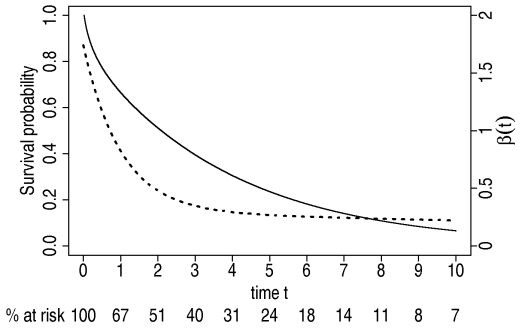


(a) (C) heavy censoring

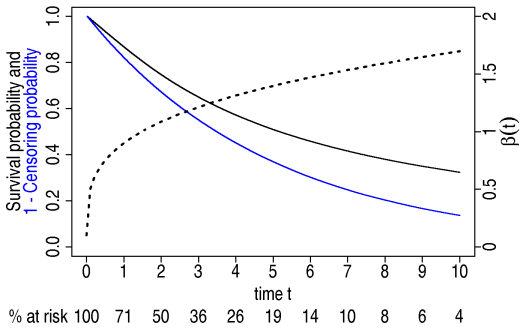
(b) (C) no censoring



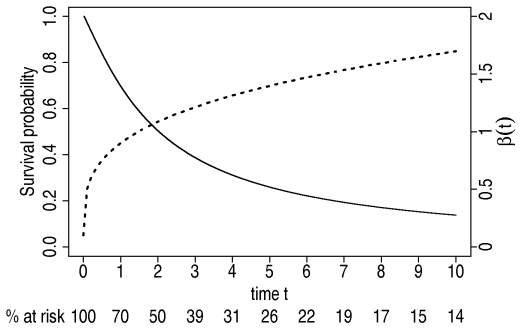
(c) (Ds) heavy censoring



(d) (Ds) no censoring



(e) (Is) heavy censoring



(f) (Is) no censoring

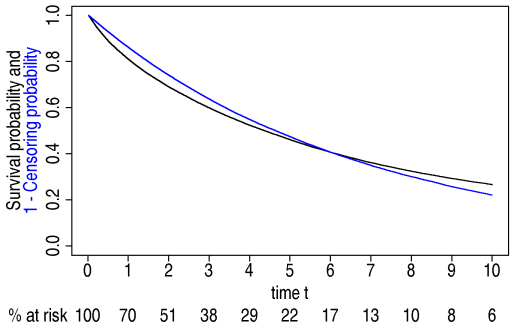
Figure C.5: Survival probability (—) and 1-censoring probability (—) for effects (C) (· · ·, top), (Ds) (· · ·, middle) and (Is) (· · ·, bottom) with heavy (left) and no (right) censoring and sample size 1000 with standard normal variable.

C.2 Multivariable settings

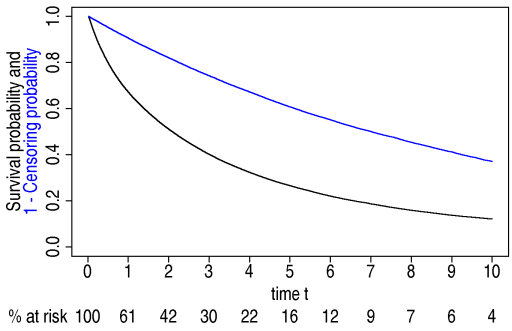
In the multivariable settings, median survival and follow-up time strongly depend on the parameter settings (Table C.4). With the current choices, it increases with larger proportion of censoring, where light censoring corresponds to on average 27% of censoring and heavy censoring to about 50%.
Survival and censoring distributions are virtually identical for settings with and without correlation between X_1 and X_4. Kaplan-Meier and reverse Kaplan-Meier curves are exemplarily shown for the uncorrelated settings in Figure C.6.

ρ_{X_1,X_4}	Censoring	Median survival time	Median follow-up time	Mean proportion of censoring
0	heavy	4.4	4.6	49.7
0	light	2.1	7.0	27.1
0	no	1.2	-	0.0
0.5	heavy	4.3	4.6	49.5
0.5	light	2.1	7.0	27.3
0.5	no	1.2	-	0.0

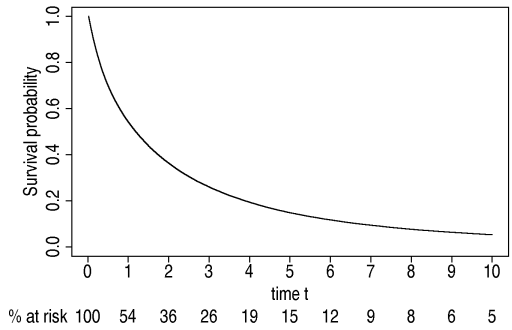
Table C.4: Median survival and follow-up time (Kaplan-Meier and reverse Kaplan-Meier method) and mean proportion of censoring in multivariable settings.



(a) heavy censoring



(b) light censoring



(c) no censoring

Figure C.6: Survival probability (—) and 1-censoring probability (—) in multivariable settings without correlation.

Appendix D

Supplementary information on the simulation study

Following are tables and figures with results of the simulation study as supplementary material to the results presented in Chapter 6.

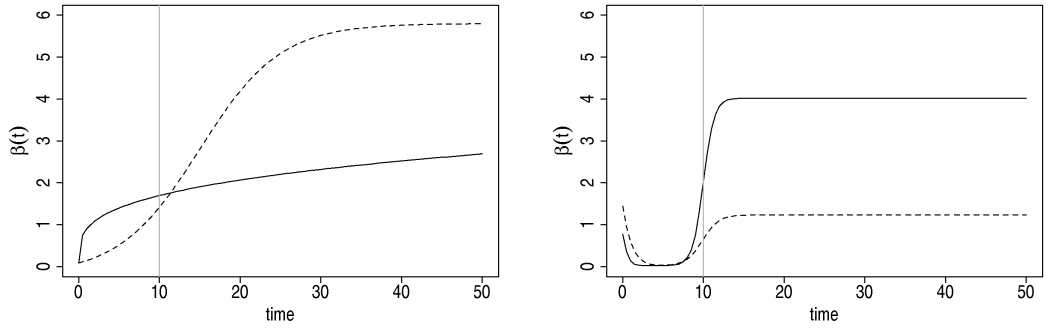
D.1 CoxPH

	Sample	Cen-	Effects									
X	size	soring	(C)	(Ls)	(Lw)	(Ds)	(Dw)	(Is)	(Iw)	(S)	(Bs)	(Bw)
binary	250	light	0.81	-	-	0.78	-	1.01	-	-	-	-
	500	light	0.81	-	-	0.77	-	1.01	-	-	-	-
	1000	heavy	0.80	-	-	0.72	-	1.08	-	0.90	0.32	0.53
	1000	light	0.80	1.78	0.80	0.76	0.52	1.01	0.33	0.69	0.24	0.52
	1000	no	0.80	1.84	0.82	0.80	0.40	0.96	0.30	-	-	-
	3000	light	0.80	1.77	0.81	0.76	0.39	1.01	0.34	-	-	-
normal	250	light	0.80	-	-	0.76	-	0.76	-	-	-	-
	500	light	0.80	-	-	0.75	-	0.76	-	-	-	-
	1000	heavy	0.79	1.67	0.81	0.77	0.38	0.67		0.32	-0.08	0.47
	1000	light	0.80	1.53	0.75	0.75	0.51	0.76	0.04	0.36	0.06	0.49
	1000	no	0.80	-	-	0.71	-	0.79	0.11	-	-	-
	3000	light	0.80	-	-	0.75	-	0.77	0.04	-	-	-

Table D.1: Mean of effects estimated by CoxPH in univariate settings.

D.2 FPT

D.2.1 Univariate settings



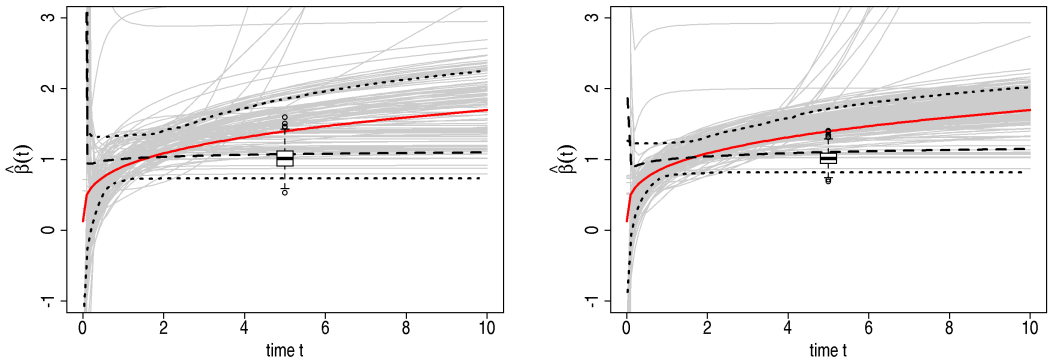
(Iw) Weak increasing effect

(Bs, Bw) Bathtub-shaped effect with strong (—) and weak (- -) end

Figure D.1: Time-varying effects with artificial asymptote.

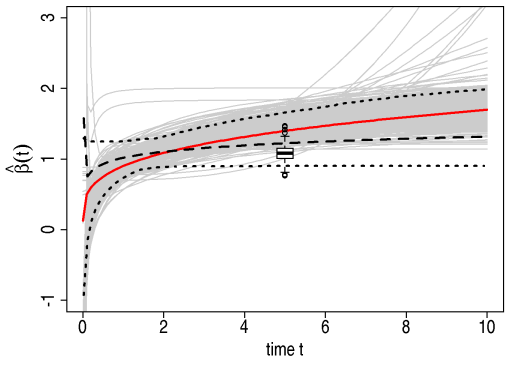
FP powers		frequency (in %)			
		X binary		X standard normal	
p_1	p_2	heavy censoring	light censoring	heavy censoring	light censoring
0		16.0	0.6		
0	0	0.7	0.6	8.3	14.6
0	0.5	0.1	0.9	7.6	9.7
0.5		16.8	2.0	2.5	0.8
0.5	0.5	0.1	0.3	1.4	0.3
1		48.4	59.9	14.4	1.9
1	1			0.3	1.9
1	2	0.4		22.0	58.3
1	3			4.8	5.1
2		1.8	21.5		
2	2	5.1	1.2	37.5	6.6
2	3	4.0	3.9	0.8	
3	3	2.8	8.2		
		and 11 further FPs with frequency $<$1%	and 8 further FPs with frequency $<$1%	and 3 further FPs with frequency $<$1%	and 2 further FPs with frequency $<$1%

Table D.2: FP powers of estimated time-varying effects for sigmoid effect (S) with binary and standard normal variable.

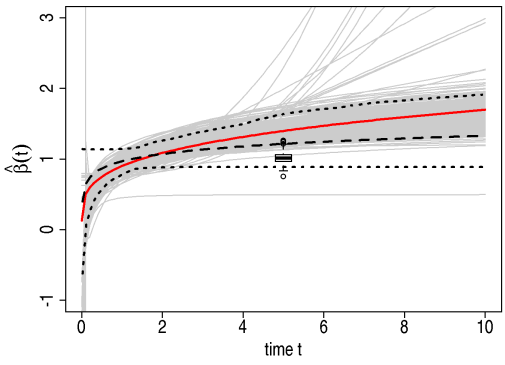


(a) $n = 250$, light censoring

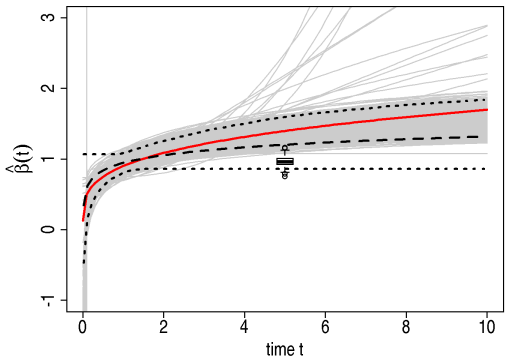
(b) $n = 500$, light censoring



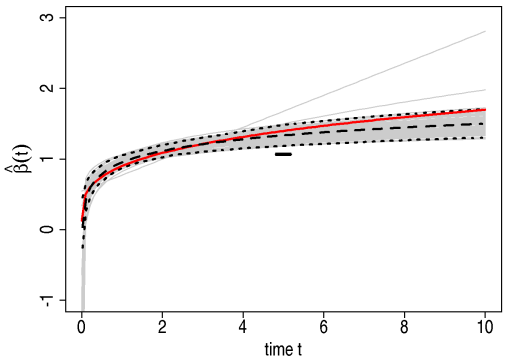
(c) $n = 1000$, heavy censoring



(d) $n = 1000$, light censoring

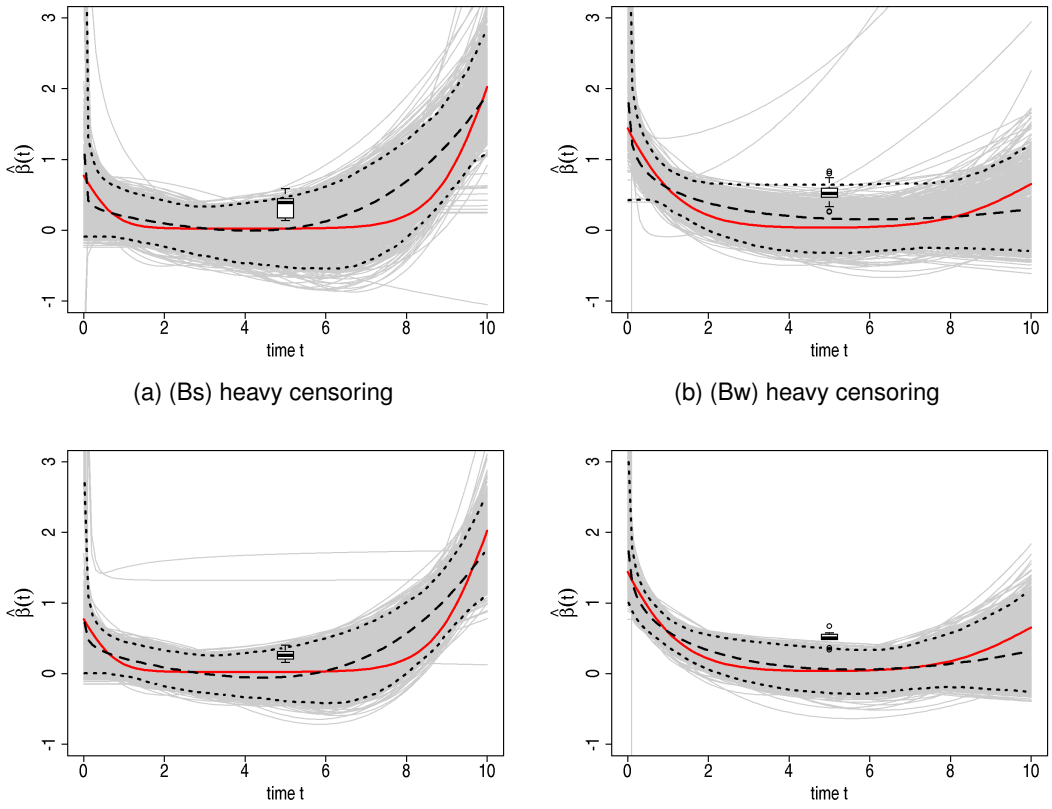


(e) $n = 1000$, no censoring



(f) $n = 3000$, light censoring

Figure D.2: True (—) effect, estimated time-varying (—) and time-constant (boxplot) effects, pointwise mean (- -) and 95% empirical confidence intervals (· · ·) for effect (ls).



(a) (Bs) heavy censoring

(b) (Bw) heavy censoring

(c) (Bs) light censoring

(d) (Bw) light censoring

Figure D.3: True (—) effect, estimated time-varying (—) and time-constant (boxplot) effects, pointwise mean (- -) and pointwise 95% empirical confidence intervals (· · ·) for effects (Bs) (left) and (Bw) (right) for light (bottom) and heavy (top) censoring with sample size 1000 and binary variable.

	Sample	Cen-	FPT				CoxPH			
	size	soring	[0,2.5]	[0,5]	[0,7.5]	[0,10]	[0,2.5]	[0,5]	[0,7.5]	[0,10]
(C)	250	light	-5.95	-7.95	-8.26	-7.98	-5.96	-7.95	-8.26	-8.02
(C)	500	light	-6.02	-8.08	-8.38	-8.13	-6.03	-8.07	-8.39	-8.15
(C)	1000	heavy	-3.91	-6.20	-7.54	-8.20	-3.92	-6.21	-7.54	-8.21
(C)	1000	light	-6.06	-8.12	-8.43	-8.19	-6.07	-8.12	-8.44	-8.20
(C)	1000	no	-7.45	-8.50	-8.14	-7.78	-7.44	-8.49	-8.14	-7.79
(C)	3000	light	-6.07	-8.15	-8.48	-8.22	-6.09	-8.15	-8.47	-8.24
(Ls)	1000	light	-47.09	-43.10	-36.88	-32.66	-47.00	-43.04	-36.82	-32.63
(Ls)	1000	no	-48.96	-38.82	-32.71	-29.91	-48.95	-38.80	-32.71	-29.92
(Ls)	3000	light	-47.23	-43.22	-36.98	-32.75	-47.02	-43.10	-36.88	-32.71
(Lw)	1000	light	-8.53	-10.03	-9.62	-8.93	-8.43	-9.99	-9.59	-8.89
(Lw)	1000	no	-10.30	-10.55	-9.65	-9.09	-10.22	-10.51	-9.62	-9.07
(Lw)	3000	light	-8.66	-10.11	-9.71	-9.01	-8.46	-10.02	-9.63	-8.93
(Ds)	250	light	-11.56	-9.93	-8.43	-7.52	-10.91	-9.73	-8.24	-7.31
(Ds)	500	light	-12.14	-10.34	-8.90	-7.98	-10.93	-9.79	-8.37	-7.44
(Ds)	1000	heavy	-7.09	-6.65	-5.85	-5.34	-6.24	-6.32	-5.54	-5.01
(Ds)	1000	light	-12.23	-10.44	-9.01	-8.10	-10.85	-9.80	-8.41	-7.48
(Ds)	1000	no	-15.65	-12.53	-10.80	-10.01	-14.06	-11.69	-10.05	-9.28
(Ds)	3000	light	-12.28	-10.49	-9.06	-8.14	-10.89	-9.84	-8.45	-7.52
(Dw)	1000	light	-2.84	-2.41	-2.03	-1.76	-2.91	-2.51	-2.11	-1.82
(Dw)	1000	no	-2.59	-2.44	-2.14	-1.97	-2.54	-2.42	-2.09	-1.90
(Dw)	3000	light	-3.09	-2.72	-2.41	-2.17	-2.61	-2.51	-2.23	-2.00
(ls)	250	light	-6.49	-11.38	-12.38	-11.99	-6.52	-11.40	-12.43	-12.14
(ls)	500	light	-6.55	-11.52	-12.54	-12.18	-6.58	-11.53	-12.57	-12.27
(ls)	1000	heavy	-4.42	-10.59	-14.79	-16.71	-4.47	-10.60	-14.78	-16.76
(ls)	1000	light	-6.61	-11.60	-12.65	-12.31	-6.59	-11.57	-12.62	-12.33
(ls)	1000	no	-7.73	-10.52	-10.30	-9.79	-7.66	-10.48	-10.28	-9.78
(ls)	3000	light	-6.70	-11.69	-12.75	-12.43	-6.57	-11.60	-12.66	-12.38
(lw)	1000	light	-0.06	-0.36	-0.86	-1.25	-0.01	-0.29	-0.81	-1.16
(lw)	1000	no	-0.09	-0.44	-0.70	-0.79	0.00	-0.35	-0.64	-0.74
(lw)	3000	light	-0.13	-0.45	-0.93	-1.35	-0.02	-0.29	-0.82	-1.18
(S)	1000	heavy	0.02	-3.99	-12.15	-15.94	1.40	-2.97	-10.66	-14.06
(S)	1000	light	0.10	-3.25	-6.46	-7.02	2.02	-1.75	-4.78	-5.42
(Bs)	1000	heavy	-0.51	-0.23	-0.10	-0.07	-0.63	-0.21	0.06	0.04
(Bs)	1000	light	-0.73	-0.43	-0.33	-0.30	-0.70	-0.38	-0.26	-0.25
(Bw)	1000	heavy	-3.97	-2.94	-2.22	-1.89	-3.34	-2.75	-2.01	-1.61
(Bw)	1000	light	-6.71	-4.99	-4.06	-3.59	-5.54	-4.50	-3.56	-3.06

Table D.3: Median difference (%) of IPEC to the Kaplan-Meier estimate (dIPEC) over different intervals [0,τ] for FPT and CoxPH in univariate settings with binary X.

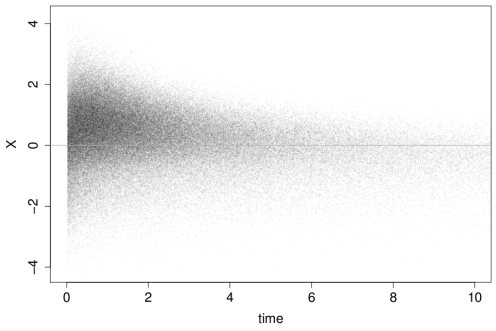
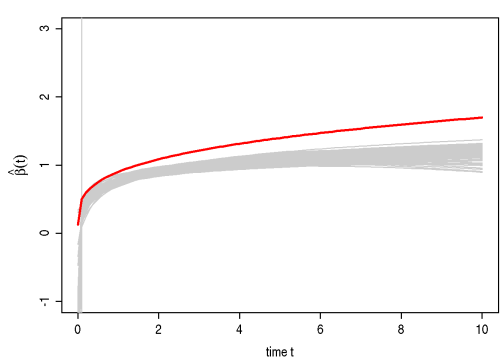
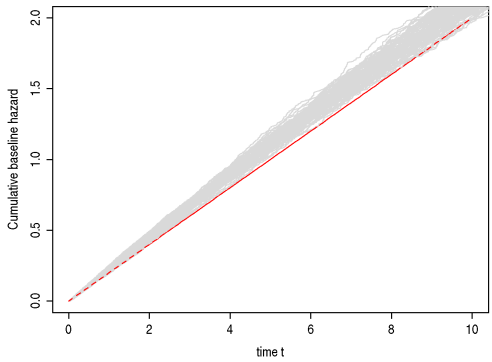


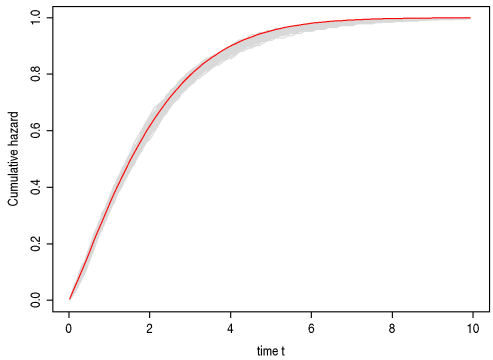
Figure D.4: Relation between the values of the standard normal variable X and uncensored event times for effect (ls) with sample size 3000 and light censoring.



(a) Estimated effects

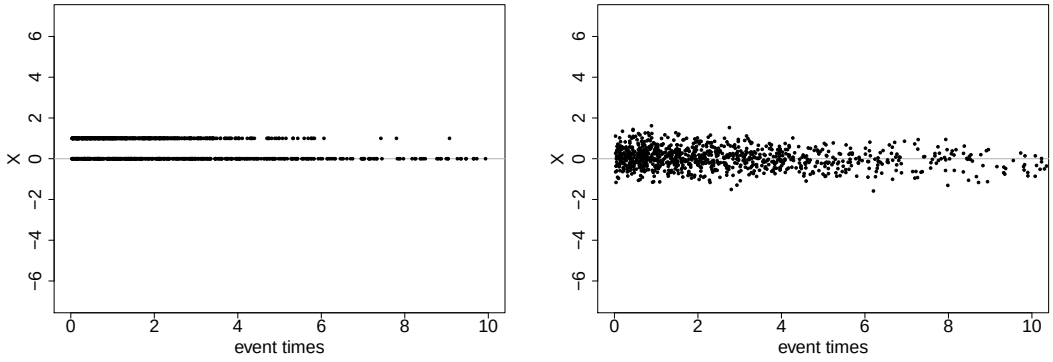


(b) Estimated cumulative baseline hazards



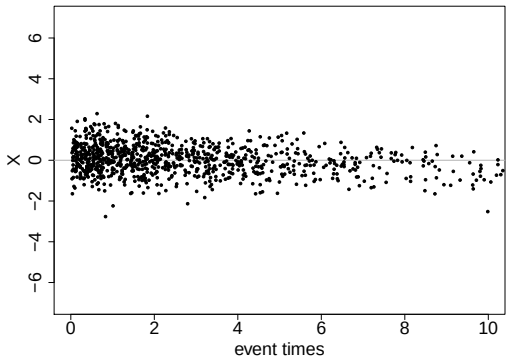
(c) Estimated cumulative hazard functions

Figure D.5: Estimated effects (top left) and cumulative baseline hazards (top right) and cumulative hazard functions (bottom) for effect (ls) with standard normal variable, $n = 3000$ and light censoring. FPT considerably underestimates the effect size, but simultaneously overestimates the cumulative baseline hazard. Shown are the true effect, cumulative baseline hazard and cumulative hazard (—), respectively, together with the estimates of all simulation runs (—).

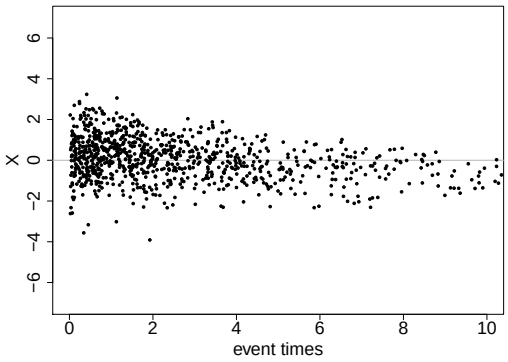


(a) X binary

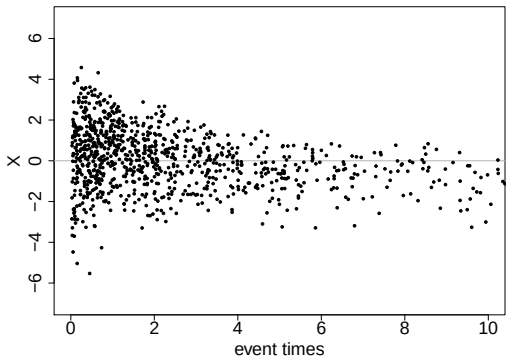
(b) $X \sim N(0, 0.25)$



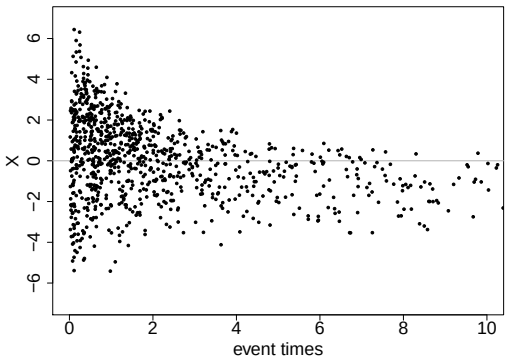
(c) $X \sim N(0, 0.5)$



(d) $X \sim N(0, 1)$

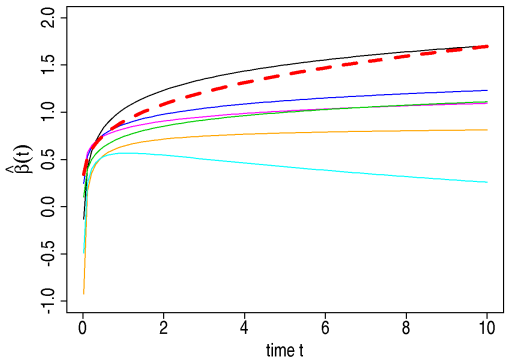


(e) $X \sim N(0, 2)$

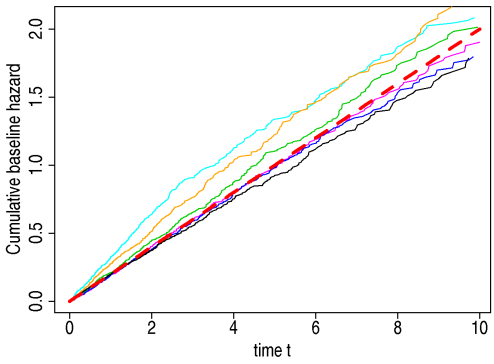


(f) $X \sim N(0, 4)$

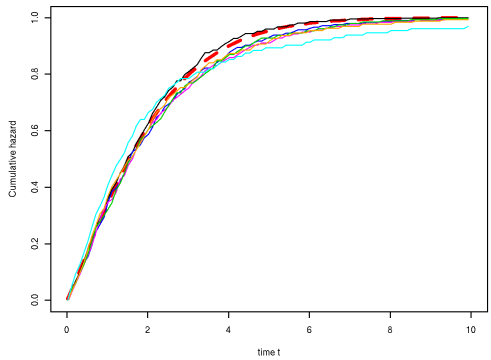
Figure D.6: Relation between the values of the variable X and uncensored event times for effect (Is) for different variances of X in one data set with sample size 1000 and light censoring.



(a) Estimated effects



(b) Estimated cumulative baseline hazards



(c) Estimated cumulative hazards

Figure D.7: Estimated effects (top left), cumulative baseline hazards (top right) and cumulative hazards (bottom) for effect (ls) with $n = 1000$ and light censoring with changing variance of the covariate: X binary with $P(X = 1) = 0.5$ (—), $X \sim N(0, 0.25)$ (—), $X \sim N(0, 0.5)$ (—), $X \sim N(0, 1)$ (—), $X \sim N(0, 2)$ (—), $X \sim N(0, 4)$ (—) and true effect, cumulative baseline hazard and cumulative hazard (- -), respectively.

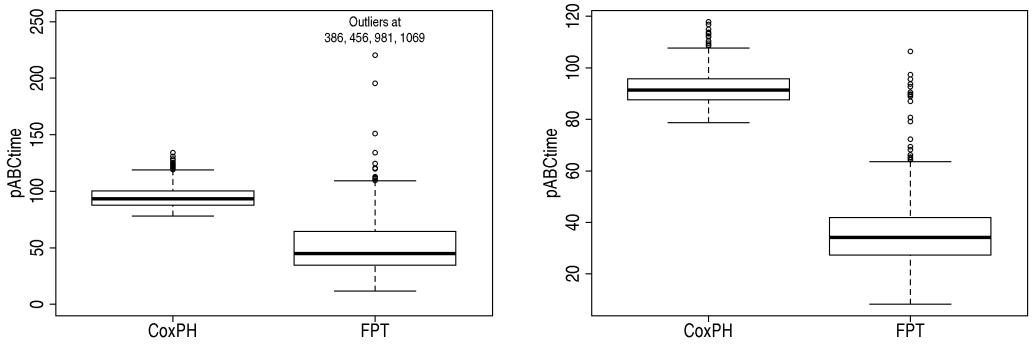


Figure D.8: pABCtime of effects selected by CoxPH and FPT relative to the true effect function for effect (Bw) with binary variable, sample size 1000 and heavy (left) or light (right) censoring.

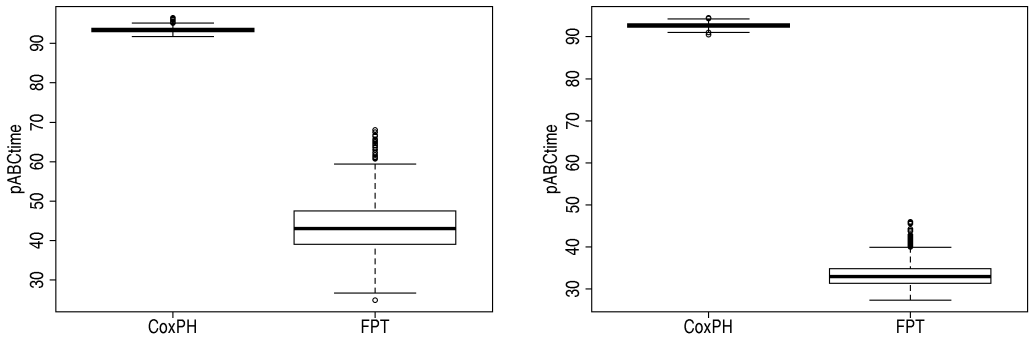


Figure D.9: pABCtime of effects selected by CoxPH and FPT relative to the true effect function for effect (S) with binary variable, sample size 1000 and heavy (left) or light (right) censoring.

	Sample	Cen-	FPT				CoxPH			
	size	soring	[0,2.5]	[0,5]	[0,7.5]	[0,10]	[0,2.5]	[0,5]	[0,7.5]	[0,10]
(C)	250	light	-17.50	-22.54	-23.98	-25.05	-17.50	-22.53	-23.98	-25.05
(C)	500	light	-17.59	-22.62	-24.08	-25.17	-17.59	-22.62	-24.07	-25.17
(C)	1000	heavy	-11.56	-17.17	-20.31	-21.74	-11.57	-17.18	-20.32	-21.74
(C)	1000	light	-17.65	-22.69	-24.16	-25.26	-17.65	-22.67	-24.13	-25.24
(C)	1000	no	-21.27	-24.34	-25.81	-26.25	-21.27	-24.32	-25.81	-26.25
(C)	3000	light	-17.67	-22.71	-24.18	-25.28	-17.68	-22.70	-24.16	-25.28
(Ls)	1000	heavy	-79.24	-81.70	-77.25	-70.51	-78.66	-81.38	-76.94	-69.48
(Ls)	1000	light	-86.86	-84.95	-78.25	-70.35	-84.92	-83.68	-77.32	-68.66
(Lw)	1000	heavy	-17.96	-22.32	-23.17	-21.99	-17.75	-22.25	-23.12	-21.78
(Lw)	1000	light	-24.19	-26.61	-25.55	-23.78	-23.49	-26.25	-25.30	-23.38
(Ds)	250	light	-34.65	-25.97	-21.44	-18.70	-30.78	-24.11	-19.50	-16.26
(Ds)	500	light	-34.80	-26.08	-21.59	-18.86	-30.89	-24.23	-19.60	-16.34
(Ds)	1000	heavy	-26.31	-20.87	-17.64	-15.46	-23.02	-19.47	-16.17	-13.48
(Ds)	1000	light	-34.87	-26.19	-21.68	-18.97	-30.78	-24.27	-19.69	-16.41
(Ds)	1000	no	-39.18	-29.00	-24.05	-21.57	-33.93	-26.56	-21.59	-18.61
(Ds)	3000	light	-34.95	-26.25	-21.75	-19.04	-30.83	-24.34	-19.74	-16.41
(Dw)	1000	heavy	-6.62	-6.29	-5.60	-5.03	-5.30	-5.74	-5.24	-4.66
(Dw)	1000	light	-9.85	-8.06	-6.57	-5.61	-9.51	-7.94	-6.04	-4.48
(Is)	250	light	-13.04	-19.12	-20.55	-20.45	-13.02	-19.09	-20.56	-20.53
(Is)	500	light	-13.13	-19.23	-20.62	-20.40	-13.07	-19.14	-20.61	-20.59
(Is)	1000	heavy	-7.71	-15.09	-18.09	-18.73	-7.63	-14.74	-17.69	-18.56
(Is)	1000	light	-12.78	-18.84	-20.35	-20.36	-13.12	-19.21	-20.66	-20.65
(Is)	1000	no	-16.30	-20.90	-21.79	-21.95	-16.04	-20.74	-21.80	-22.12
(Is)	3000	light	-13.36	-19.40	-20.71	-20.40	-13.13	-19.23	-20.69	-20.68
(Iw)	1000	light	0.06	0.04	-0.03	-0.04	0.13	0.07	0.01	-0.03
(Iw)	1000	no	0.08	-0.16	-0.23	-0.22	0.16	-0.10	-0.20	-0.23
(Iw)	3000	light	-0.05	-0.04	-0.12	-0.12	0.15	0.08	-0.01	-0.06
(S)	1000	heavy	-1.90	-6.19	-12.94	-16.55	3.90	-0.70	-5.50	-7.99
(S)	1000	light	-0.36	-4.16	-8.42	-10.98	4.59	-0.27	-4.20	-6.37
(Bs)	1000	heavy	-0.09	-0.68	-1.38	-2.01	0.15	-0.28	-0.69	-0.93
(Bs)	1000	light	-1.41	-0.76	-0.58	-0.56	-0.56	-0.26	-0.03	0.08
(Bw)	1000	heavy	-13.97	-9.49	-6.78	-5.48	-10.23	-7.71	-4.71	-2.96
(Bw)	1000	light	-20.19	-13.72	-10.52	-8.89	-15.98	-11.66	-8.08	-5.84

Table D.4: Median difference (%) of IPEC to the Kaplan-Meier estimate (dIPEC) over different intervals [0,τ] for FPT and CoxPH in univariate settings with standard normal X.

D.2.2 Multivariable settings

Selection	No censoring		Light censoring		Heavy censoring	
sequence	$\rho_{X_1,X_4} = 0$	$\rho_{X_1,X_4} = 0.5$	$\rho_{X_1,X_4} = 0$	$\rho_{X_1,X_4} = 0.5$	$\rho_{X_1,X_4} = 0$	$\rho_{X_1,X_4} = 0.5$
PH	-	-	-	-	2	1
X_1	13	14	26	22	29	32
$\mathbf{X_1 X_2}$	**82**	**75**	**61**	**64**	**42**	**36**
$X_1 X_2 X_3$	1	2	-	-	1	1
$X_1 X_2 X_4$	1	3	-	-	-	-
$X_1 X_3$	-	-	-	-	1	-
$X_1 X_4$	-	-	-	-	-	1
$X_1 X_5 X_2$	-	-	-	1	-	-
X_2	-	-	1	-	5	7
$\mathbf{X_2 X_1}$	**3**	**6**	**11**	**12**	**20**	**20**
$X_2 X_1 X_5$	-	-	-	-	-	1
$X_2 X_3$	-	-	1	-	-	-
$X_2 X_4$	-	-	-	-	-	1
$X_4 X_2$	-	-	-	1	-	-

Table D.5: Selection sequence for time-varying effects in the FPT approach in multivariable settings.

ρ_{X_1,X_4}	Censoring	X_1	X_2	X_3	X_4	X_5
0	heavy	18.41	11.90	17.78	10.71	8.27
0	light	15.52	9.74	17.81	9.43	7.51
0	no	15.28	7.92	14.17	8.78	7.04
0.5	heavy	19.96	12.02	22.27	17.72	8.84
0.5	light	14.94	9.03	20.49	12.98	7.12
0.5	no	16.32	8.06	14.42	10.23	6.48

Table D.6: Median pABCtime for FPT in multivariable settings.

ρ_{X_1,X_4}	Censoring	X_1	X_2	X_3	X_4	X_5
0	heavy	57.74	23.81	20.13	11.40	8.04
0	light	59.02	24.87	19.59	10.21	8.71
0	no	62.40	26.76	14.79	9.08	7.14
0.5	heavy	57.63	23.77	21.35	16.33	8.22
0.5	light	59.36	25.01	20.82	12.61	7.29
0.5	no	63.63	26.74	15.12	9.91	6.67

Table D.7: Median pABCtime for CoxPH in multivariable settings.

D.3 Semiparametric Extended Cox model

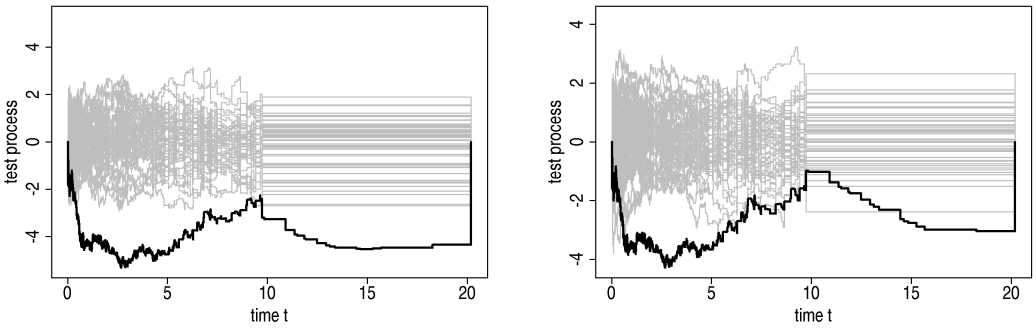
To explore the impact of the bandwidth on invertibility problems in our simulation study, the bandwidth is increased from the default 0.5 to 0.7, 0.9, 1 and 1.5. For binary variable and unweighted test, a type I error of approximately 1% is observed for bandwidth 0.9 that decreases further for larger bandwidths. With weighted test statistic, though, even with a bandwidth of 5, the type I error is with about 6% still considerably inflated. With normal variable, both test statistics approximately hold the nominal significance level of 1% for bandwidth 1.5. These results emphasise the impact of the chosen bandwidth on the test process (Figure D.10) and thus on the test results.

Although the type I error seems more or less sensible for these choices of bandwidth, the estimation algorithm still issues a considerable amount of invertibility warnings. Hence, reliability of the results is questionable.

To evaluate the impact of the bandwidth on estimated effect, reconsider the simulated data set with five standard normal variables used in Section 4.2. Varying the bandwidth of the estimation algorithm with test version $w = 1$ leads to different conclusions about time-varying effects.

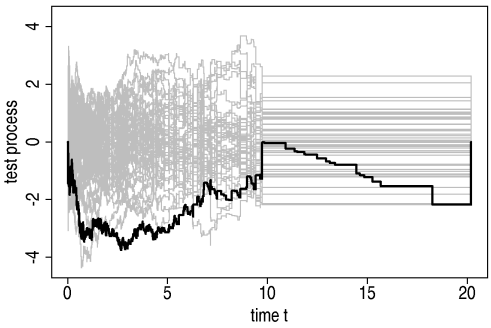
For a bandwidth of 0.4, i.e. 40% of the range of the observation period, the backward elimination algorithm selects time-varying effects for X_1, X_2, X_4 and X_5. Increasing the bandwidth to 0.5, 0.6 or 0.7 results in time-varying effects for X_1, X_2 and X_4, while for bandwidths 0.9 and 1 only the effect of X_1 is assumed to be time-varying. Although 0.4 is the smallest bandwidth, for which the algorithm converged, the results obtained with the default bandwidth 0.5 appear to be more stable. Using this bandwidth, the model selection procedure correctly detects the two time-varying effects for X_1 and X_2, but additionally selects one false positive time-varying effects for X_4. However, none of the other bandwidths results in the correct model either. With varying bandwidths for a fixed model, that is investigating estimated effects without considering the test on time-varying effects, the effect estimates are virtually identical over the first 20 years, but differ largely for later times where data gets sparse (Figure D.11).

The impact of the bandwidth for smoothing of the final estimate $\hat{B}(t)$, i.e. for the local smoother `locfit`, on the shape of estimated time-varying effects $\hat{\beta}(t)$ is more pronounced even early in time. Figure D.12 shows the smoothed effects $\hat{\beta}(t)$ for five different bandwidths between 0.2 and 1.2. As expected, the small bandwidths result in rather wiggly curves, while the larger ones oversmooth the effects and eliminate most of the curvature. The default bandwidth 0.7, though, seems to be a good compromise between both extremes, reflecting the slope of $\hat{B}(t)$ without including too much noise.

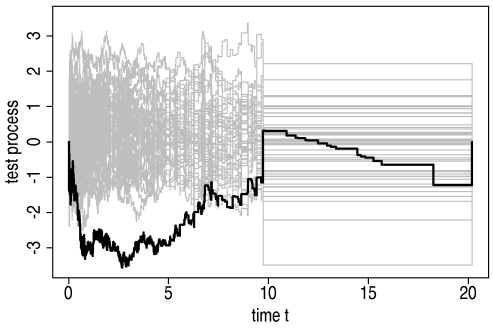


(a) Bandwidth=0.5

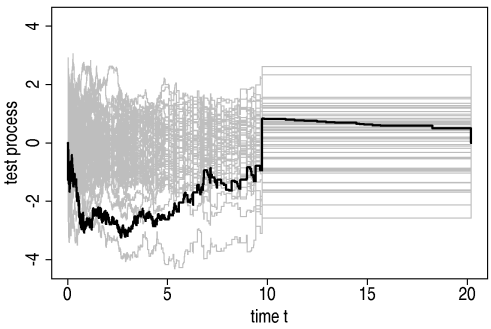
(b) Bandwidth=0.7



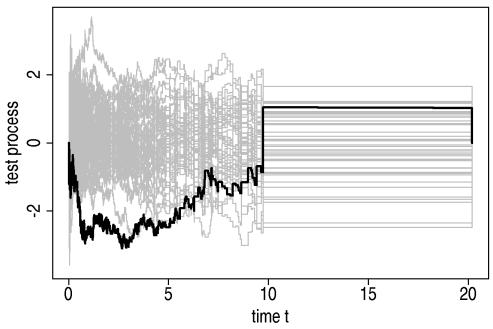
(c) Bandwidth=0.9



(d) Bandwidth=1

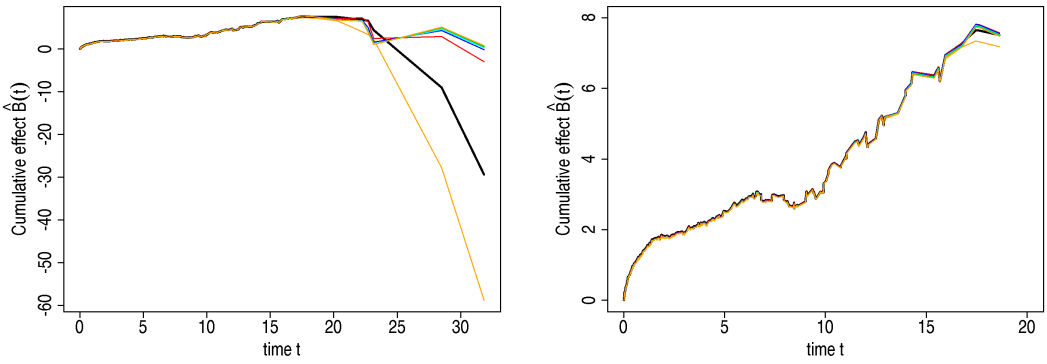


(e) Bandwidth=1.5



(f) Bandwidth=5

Figure D.10: Test process for different bandwidths within the Semiparametric Extended Cox algorithm in a univariate setting with constant effect (C). The test decision changes with changing bandwidth. For bandwidths 0.9, 1.5 and 5, the test on PH correctly decides on a constant effect, while for bandwidths 0.5, 0.7 and 1 it is false positive.



(a) Overview of the complete time period

(b) Detailed view up to t =20

Figure D.11: Cumulative effects obtained for different bandwidths within the Semiparametric Extended Cox algorithm in the simulated data. Given are the cumulative time-varying effects $\hat{B}(t)$ over the complete observed time period (left) and a detailed view on the period [0,20] (right) with default bandwidth 0.5 (—) and bandwidths 0.4 (—), 0.6 (—), 0.7 (—), 0.8 (—), 0.9 (—) and 1.0 (—) for variable X_1 of the simulated data set presented in Section 4.2 with true effect function (Ds).

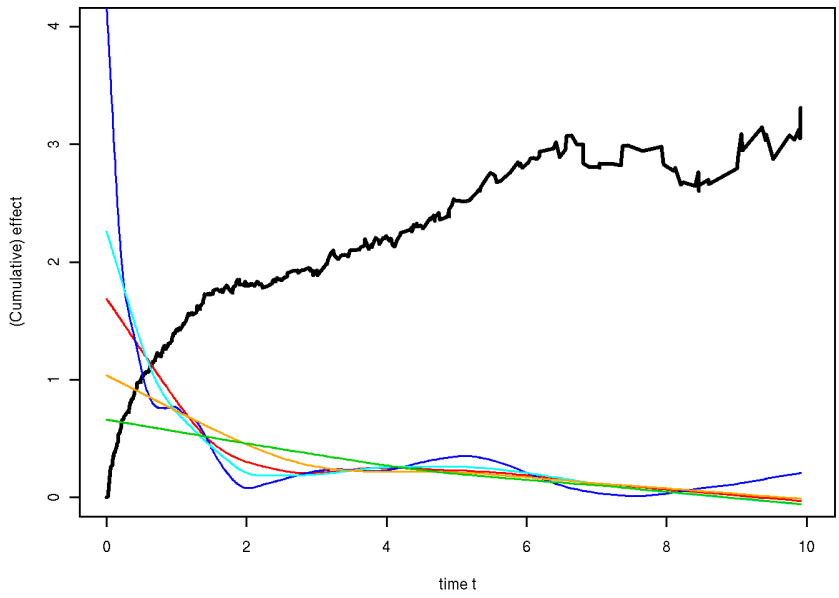


Figure D.12: Time-varying effects obtained for different bandwidths in the final smoothing step of the Semiparametric Extended Cox model. Given are the cumulative time-varying effect (—) and its smoothed derivatives $\hat{\beta}(t)$ with default bandwidth 0.7 (—) and bandwidths 0.2 (—), 0.5 (—), 0.9 (—) and 1.2 (—) for variable X_1 of the simulated data set presented in Section 4.2 with true effect function (Ds).

Bibliography

Abrahamowicz, M., T. MacKenzie, and J. M. Esdaile (1996). Time-dependent hazard ratio: Modeling and hypothesis testing with application in lupus nephritis. *Journal of the American Statistical Association 91*, 1432–1439.

Abrahamowicz, M. and T. A. MacKenzie (2007). Joint estimation of time-dependent and nonlinear effects of continuous covariates on survival. *Statistics in Medicine 26*, 392–408.

Altman, D., B. De Stavola, S. Love, and K. Stepniewska (1995). Review of survival analyses published in cancer journals. *British Journal of Cancer 72*, 511–518.

Anderson, J. A. and A. Senthilselvan (1982). A two-step regression model for hazard functions. *Applied Statistics 31*, 44–51.

Augustin, N., W. Sauerbrei, and M. Schumacher (2005). The practical utility of incorporating model selection uncertainty into prognostic models for survival data. *Statistical Modelling 5*, 95–118.

Belitz, C., A. Brezger, T. Kneib, and S. Lang (2009). *BayesX - Software for Bayesian inference in structured additive regression models. Version 2.00*. Available from http://www.stat.uni-muenchen.de/ bayesx.

Bender, R., T. Augustin, and M. Blettner (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine 24*, 1713–1723.

Berger, U., J. Schäfer, and K. Ulm (2003). Dynamic Cox modelling based on fractional polynomials: Time–variations in gastric cancer prognosis. *Statistics in Medicine 22*(7), 1163–1180.

Beyersmann, J., A. Latouche, A. Buchholz, and M. Schumacher (2009). Simulating competing risks data in survival analysis. *Statistics in Medicine 8*, 956–971.

Biganzoli, E., P. Boracchi, L. Mariani, and E. Marubini (1998). Feed forward neural networks for the analysis of cenored survival data: a partial logistic regression approach. *Statistics in Medicine 17*, 1169–1186.

Binder, H., W. Sauerbrei, and P. Royston (2010). Measures and simulation designs for evaluating multivariable model building with continuous covariates. Manuscript.

Binquet, C., M. Abrahamowicz, K. Astruc, J. Faivre, C. Bonithon-Kopp, and C. Quantin (2009). Flexible statistical models provided new insights into the role of quantitative prognostic factors for mortality in gastric cancer. *Journal of Clinical Epidemiology 62*, 232–240.

Binquet, C., M. Abrahamowicz, A. Mahboubi, V. Jooste, J. Faivre, C. Bonithon-Kopp, and C. Quantin (2008). Empirical study of the dependence of the results of multivariable flexible survival analyses on model selection strategy. *Statistics in Medicine 27*, 6470–6488.

Bland, J. M. and D. G. Altman (1998). Bayesians and frequentists. *British Medical Journal 317*, 1151.

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*, 123–140.

Brown, D., G. Kauermann, and I. Ford (2007). A partial likelihood approach to smooth estimation of dynamic covariate effects using penalised splines. *Biometrical Journal 49*, 1–12.

Buchholz, A., N. Holländer, and W. Sauerbrei (2008). On properties of predictors derived with a two-step bootstrap model averaging approach - a simulation study in the linear regression model. *Computational Statistics & Data Analysis 52*, 2778–2793.

Buchholz, A., W. Sauerbrei, and P. Royston (2009). Investigation of time-varying effects in survival analysis may require categorisation of time: does it matter? Manuscript.

Buckland, S., K. Burnham, and N. Augustin (1997). Model selection: an integral part of inference. *Biometrics 53*, 603–618.

Cai, Z. and Y. Sun (2003). Local linear estimation for time-dependent coefficients in cox's regression model. *Scandinavian Journal of Statistics 30*(1), 93–111.

Coradini, D., P. Daidone, Maria Grazia andBoracchi, E. Biganzoli, S. Oriana, G. Bresciani, C. Pellizarro, G. Tomasic, G. Di Fronzo, and E. Marubini (2000). Time-dependent relevance of steroid receptors in breast cancer. *Journal of Clinical Oncology 18*, 2702–2709.

Cortese, G., T. H. Scheike, and T. Martinussen (2010). Flexible survival regression modelling. *Statistical Methods in Medical Research 19*, 5–28.

Costa, M. and J. Shaw (2009). Parametrization and penalties in spline models with an application to survival analysis. *Computational Statistics and Data Analysis 53*, 657–670.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B: Methodological 34*, 187–220.

D'Agostino, Ralph B., S. (2009). The delayed-start study design. *The New England Journal of Medicine 361*, 1304–1306.

Dunkler, D., M. Schemper, and G. Heinze (2010). Gene selection in microarray survival studies under possibly non-proportional hazards. *Bioinformatics 26*, 784–790.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association 78*, 316–330.

Efron, B. and R. Tibshirani (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association 92*(438), 548–560.

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

Foekens, J. A., H. A. Peters, M. P. Look, H. Portengen, M. Schmitt, M. D. Kramer, N. Brünner, F. Jänicke, M. E. M. van Gelder, S. C. Henzen-Logmans, W. L. van Putten, and J. G. Klijn (2000). The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients. *Cancer Research 60*(3), 636–643.

Gao, D., G. K. Grunwald, J. S. Rumsfeld, L. Schooley, T. MacKenzie, and A. L. Shroyer (2006). Time-varying risk factors for long-term mortality after coronary artery bypass graft surgery. *Annals of Thoracic Surgery 81*, 793–799.

Gerds, T. A. and M. Schumacher (2007). Efron-type measures of prediction error for survival analysis. *Biometrics 63*(4), 1283–1287.

Govindarajulu, U. S., D. Spiegelman, S. W. Thurston, B. Ganguli, and E. A. Eisen (2007). Comparing smoothing techniques in cox models for exposure-response relationships. *Statistics in Medicine 26*, 3735–3752.

Grambsch, P. M. and T. M. Therneau (1994). Proportional hazards tests and diagnostics based on weighted residuals (Corr: 95V82 p668). *Biometrika 81*, 515–526.

Gray, R. (1994). Spline-based tests in survival analysis. *Biometrics 50*, 640–652.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association 87*, 942–951.

Hand, D. (2001). Measuring diagnostic accuracy of statistical prediction rules. *Statistica Neerlandica 55*, 3–16.

Haneuse, S. J.-P., K. D. Rudser, and D. L. Gillen (2008). the separation of timescales in bayesian survival of the time-varying effect of a time-dependent exposure. *Biostatistics 9*, 400–410.

Härdle, W. (1990). *Smoothing Techniques: With Implementation in S*. Springer.

Hastie, T. J. and R. J. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B 55*, 757–796.

He, J., D. L. McGee, and X. Niu (2010). Application of the bayesian dynamic survival model in medicine. *Statistics in Medicine 29*, 347–360.

Heinzl, H. and A. Kaider (1997). Gaining more flexibility in cox proportional hazards regression models with cubic spline functions. *Computer Methods and Programs in Biomedicine 54*(3), 201–208.

Hennerfeind, A., A. Brezger, and L. Fahrmeir (2006). Geoadditive survival models. *Journal of the American Statistical Association 101*, 1065–1075.

Hess, K. R. (1994). Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Statistics in Medicine 13*, 1045–1062.

Hilsenbeck, S. G., P. M. Ravdin, C. A. de Moor, C. K. Chamness, Gary C.and Osborne, and G. M. Clark (1998). Time-dependence of hazard ratios for prognostic factors in primary breast cancer. *Breast Cancer Research and Treatment 52*, 227–237.

Hofner, B., T. Kneib, W. Hartl, and H. Küchenhoff (2010). Building cox-type structured hazard regression models with time-varying effects. *Statistical Modelling: An International Journal*, to appear. Available as Technical Report at http://epub.ub.uni-muenchen.de/3232/.

Holländer, N., N. Augustin, and W. Sauerbrei (2006). Investigation on the improvement of prediction by bootstrap model averaging. *Methods of Information in Medicine 45*, 44–50.

Holländer, N. and M. Schumacher (2006). Estimating the functional form of a continuous covariate's effect on survival time. *Computational Statistics and Data Analysis 50*, 1131–1151. (abstr 534).

Kalbfleisch, J. and R. Prentice (2002). *The statistical analysis of failure time data* (Second edition ed.). Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons.

Keiding, N., P. K. Andersen, and J. P. Klein (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine 16*, 215–224.

Klein, J. P. and M. L. Moeschberger (2005). *Survival Analysis. Techniques for Censored and Truncated Data.* New York: Springer.

Kneib, T. (2006). *Mixed model based inference in structured additive regression*. Dr. Hut, München.

Kneib, T. and L. Fahrmeir (2007). A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics 34*, 207–228.

Leemis, L. M., L.-H. Shih, and K. Reynertson (1990). Variate generation for accelerated life and proportional hazards models with time dependent covariates. *Statistics & Probability Letters 10*, 335–339.

Lehr, S. and M. Schemper (2007). Parsimonious analysis of time-dependent effects in the cox model. *Statistics in Medicine 26*, 2686–2698.

Liestøl, K., P. K. Andersen, and U. Andersen (1994). Survival analysis and neural nets. *Statistics in Medicine 13*, 1189–1200.

Loader, C. (1999). *Local regression and likelihood*. Springer.

Loader, C. (2007). *locfit: Local Regression, Likelihood and Density Estimation.* R package version 1.5-4.

Lunn, A. and S. Davies (1998). A note on generating correlated binary variables. *Biometrika 85*, 487–490.

MacKenzie, T. and M. Abrahamowicz (2002). Marginal and hazard ratio specific random number data generation: Applications to semi-parametric bootstrapping. *Statistics and Computing 12*, 245–252.

Marmot, M., M. Shipley, and G. Rose (1984). Inequalities in death - specific explanations of a general pattern? *Lancet 323*, 1003–1006.

Martinussen, T. and T. Scheike (2006). *Dynamic Regression Models for Survival Data*. New York: Springer-Verlag Inc.

Martinussen, T., T. H. Scheike, and I. M. Skovgaard (2002). Efficient estimation of fixed and time-varying covariate effects in multiplicative intensity models. *Scandinavian Journal of Statistics 29*(1), 57–74.

McKeague, I. W. and M. Tighiouart (2000). Bayesian estimators for conditional hazard functions. *Biometrics 56*, 1007–1015.

Moreau, T., J. O'Quigley, and M. Mesbah (1985). A global goodness-of-fit statistic for the proportional hazards model. *Applied Statistics 34*, 212–218.

Ng'Andu, N. H. (1997). An empirical comparison of statistical tests for assessing the proportional hazards assumption of cox's model. *Statistics in Medicine 16*, 611–626.

O'Quigley, J. and F. Pessione (1991). The problem of a covariate-time qualitative interaction in a survival study. *Biometrics 47*, 101–115.

Perperoglou, A. (2005). *coxvc: Cox models with time varying effects of the covariates and Reduced Rank models*. R package version 1-1-1.

Perperoglou, A., S. le Cessie, and H. C. van Houwelingen (2006a). A fast routine for fitting Cox models with time varying effects of the covariates. *Computer Methods and Programs in Biomedicine 81*, 154–161.

Perperoglou, A., S. le Cessie, and H. C. van Houwelingen (2006b). Reduced-rank hazard regression for modelling non-proportional hazards. *Statistics in Medicine 25*, 2831–2845.

Porzelius, C. and H. Binder (2009). *peperr: Parallelised Estimation of Prediction Error*. R package version 1.1-3.

Porzelius, C., M. Schumacher, and H. Binder (2010). A general, prediction error based criterion for selecting model complexity for high-dimensional survival models. *Statistics in Medicine 29*, 830–838.

Putter, H., M. Sasako, H. H. Hartgrink, C. J. H. van de Velde, and J. C. van Houwelingen (2005). Long-term survival with non-proportional hazards: results from the Dutch Gastric Cancer Trial. *Statistics in Medicine 24*, 2807–2821.

Quantin, C., M. Abrahamowicz, T. Moreau, G. Bartlett, T. MacKenzie, M. A. Tazi, L. Lalonde, and J. Faivre (1999). Variation over time of the effects of prognostic factors in a population-based study of colon cancer: comparison of statistical models. *American Journal of Epidemiology 150*, 1188–1200.

Royston, P. and D. G. Altman (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling (Disc: P453-467). *Applied Statistics 43*, 429–453.

Royston, P. and W. Sauerbrei (2003). Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation. *Statistics in Medicine 22*, 639–659.

Royston, P. and W. Sauerbrei (2005). Building multivariable regression models with continuous covariates in clinical epidemiology–with an emphasis on fractional polynomials. *Methods of Information in Medicine 44*, 561–571.

Royston, P. and W. Sauerbrei (2007). Improving the robustness of fractional polynomial models by preliminary covariate transformation: A pragmatic approach. *Computational Statistics & Data Analysis 51*, 4240–4253.

Royston, P. and W. Sauerbrei (2008). *Multivariable Modelling: A pragmatic approach based on fractional polynomials for continuous variables*. John Wiley & Sons.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics 11*, 735–757.

Ruppert, D., M. Wand, and R. Carroll (2003). *Semiparametric Regression*. Cambridge University Press, U.K.

Sabanés Bové, D. and L. Held (2010). Bayesian fractional polynomials. *Statistics and Computing*, to appear. DOI 10.1007/s11222-010-9170-7.

Sauerbrei, W., N. Holländer, and A. Buchholz (2008). Investigation about a screening step in model selection. *Statistics and Computing 18*(2), 195–208.

Sauerbrei, W. and P. Royston (1999). Building multivariable prognostic and diagnostic models: Transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society, Series A: Statistics in Society 162*(1), 71–94.

Sauerbrei, W., P. Royston, and M. Look (2007). A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. *Biometrical Journal 49*, 453–473.

Sauerbrei, W. and M. Schumacher (1992). A bootstrap resampling procedure for model building: Application to the Cox regression model. *Statistics in Medicine 11*(16), 2093–2109.

Scheike, T. (2009). *timereg: timereg package for Flexible regression models for survival data.* R package version 1.2-4.

Scheike, T. H. and T. Martinussen (2004). On estimation and tests of time-varying effects in the proportional hazards model. *Scandinavian Journal of Statistics 31*, 51–62.

Schemper, M., S. Wakounig, and G. Heinze (2009). The estimation of average hazard ratios by weighted cox regression. *Statistics in Medicine 28*, 2473–2489.

Shepherd, B. E. (2008). The cost of checking proportional hazards. *Statistics in Medicine 27*, 1248–1260.

StataCorp (2007). *Stata Statistical Software: Release 10*. College Station, TX: StataCorp LP.

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

Therneau, T. M. and P. M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag Inc.

Tian, L., D. Zucker, and L. Cai (2005). On the cox model with time-varying regression coefficients. *Journal of the American Statistical Association 100*, 172–183.

van Buuren, S., H. Boshuizen, and D. Knook (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine 18*, 681–694.

Verweij, P. J. M. and H. C. van Houwelingen (1995). Time-dependent effects of fixed covariates in cox regression. *Biometrics 51*, 1550–1556.

Xu, R. and J. O'Quigley (2000). Estimating average regression effect under non-proportional hazards. *Biostatistics 1*, 423–439.