



LUDWIG-  
MAXIMILIANS-  
UNIVERSITÄT  
MÜNCHEN

INSTITUT FÜR STATISTIK



Benjamin Hofner, Thomas Kneib,  
Wolfgang Hartl & Helmut Küchenhoff

# Building Cox-Type Structured Hazard Regression Models with Time-Varying Effects

Technical Report Number 027, 2008  
Department of Statistics  
University of Munich

<http://www.stat.uni-muenchen.de>



# Building Cox-Type Structured Hazard Regression Models with Time-Varying Effects

*This is a preliminary version of an article that has been accepted for publication in **Statistical Modelling: An International Journal**.  
Please use the journal version for citation.*

Benjamin Hofner\*

Institut für Medizininformatik, Biometrie und Epidemiologie  
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany;

Thomas Kneib

Institut für Statistik  
Ludwig-Maximilians-Universität München, Germany;

Wolfgang Hartl

Department of Surgery, Campus Großhadern  
Ludwig-Maximilians-Universität München, Germany;

Helmut Küchenhoff

Institut für Statistik  
Ludwig-Maximilians-Universität München, Germany;

In recent years, flexible hazard regression models based on penalised splines have been developed that allow us to extend the classical Cox-model via the inclusion of time-varying and nonparametric effects. Despite their immediate appeal in terms of flexibility, these models introduce additional difficulties when a subset of covariates and the corresponding modelling alternatives have to be chosen. We present an analysis of data from a specific patient population with 90-day survival as the response variable. The aim is to determine a sensible prognostic model where some variables have to be included due to subject-matter knowledge while other variables are subject to model selection. Motivated by this application, we propose a two-stage stepwise model building

---

\*benjamin.hofner@imbe.med.uni-erlangen.de, Waldstraße 6, 91054 Erlangen, Germany;

strategy to choose both the relevant covariates and the corresponding modelling alternatives within the choice set of possible covariates simultaneously. For categorical covariates, competing modelling approaches are linear effects and time-varying effects, whereas nonparametric modelling provides a further alternative in case of continuous covariates. In our data analysis, we identified a prognostic model containing both smooth and time-varying effects.

*Key words: hazard regression, mixed models, model building, prognostic model, P-splines, time-varying effects*

## 1 Introduction

Knowledge on factors which determine the prognosis of critically ill, septic patients is of utmost clinical importance. Traditional prognostic models mostly used binary outcome variables reflecting intensive care unit or hospital mortality. However, the latter variables are highly sensitive to variations in the rate of inter-institutional patient transfer (Kahn, Kramer and Rubenfeld, 2007), and derived prognostic models, are, therefore, likely to be biased. This bias may be overcome by using survival time as dependent variable, for example in a hazard regression model. Besides the question how to model the prognosis of septic patients, the more difficult remaining problem is the exact specification of a regression model. In our application, it turned out that both non-linear and time-varying effects are present, requiring a flexible regression model allowing for these possibilities. Moreover, we have to decide which of the covariates to include in the final model, taking into account that some covariates are preset due to subject matter knowledge. Still the precise form of the effect (e.g. linear vs. non-linear) may be unknown and has to be determined from the data in a model selection procedure.

We start our methodological considerations with the popular Cox proportional hazards model (Cox, 1972) that specifies the impact of covariates  $v$  on continuous survival times in a regression model for the hazard rate of an observation  $i$ ,  $i = 1, \dots, n$ , via

$$\lambda_i(t) = \lambda(t, v_i) = \lambda_0(t) \exp(v_i' \gamma), \quad (1)$$

where  $\gamma$  is a vector of regression coefficients and  $\lambda_0(t)$  is a baseline hazard rate common to all observations. The two most distinctive features of the Cox model are the proportionality of the hazard rates for different observations and estimation of the regression

coefficients based on the partial likelihood. Proportional hazards result from the clear separation of temporal effects in the baseline hazard rate and (multiplicative) covariate effects in the linear predictor  $v_i'\gamma$ . This specification is quite restrictive and often in conflict with data observed in practice where covariate effects may well be time-varying leading to non-proportional hazards. Moreover, restricting the attention to linear effects can be misleading when some of the true covariate effects are sufficiently non-linear.

Therefore, recent approaches as in Kneib and Fahrmeir (2007) or Kauermann, Xu and Vaida (2008) replace the linear predictor with a more flexible predictor comprising time-varying and nonparametric effects leading to the hazard rate model

$$\lambda_i(t) = \lambda_0(t) \exp \left( \sum_{l=1}^L g_l(t)u_{il} + \sum_{j=1}^J f_j(x_{ij}) + v_i'\gamma \right), \quad (2)$$

where  $g_l(t)$  are time-varying effects of covariates  $u_l$ ,  $f_j(x_j)$  are smooth effects of continuous covariates  $x_j$ , and  $v'\gamma$  represents linear effects. In the application our research is based on, the major aim is to determine a sensible prognostic model, i.e. a model that contains all covariates that are relevant for prognosis in an appropriate way. Note that we are not interested in individual predictions of the survival time of specific patients but instead aim at determining an interpretable model containing all relevant prognostic factors. Therefore, we require a model building procedure allowing for enough flexibility in the model specification while avoiding unnecessary complexity. In addition, we have to account for the constraint that some of the covariates are preset, i.e. they have to be part of the final model due to subject-matter knowledge. In this case, we only have to decide between modelling alternatives, whereas exclusion of the covariates is not an option. In summary, we require both a flexible model as in (2) and a model selection strategy to identify prognostic factors based on a set of candidate covariates.

We propose a semiparametric approach where both smooth model terms and time-varying effects are specified as penalised splines allowing for considerably more flexibility, especially in the time-varying effects as compared to approaches based on artificial covariates (e.g. Therneau and Grambsch, 2000) or fractional polynomials (Sauerbrei and Royston, 1999; Sauerbrei, Royston and Look, 2007). Representing the log-baseline hazard rate as a further penalised spline yields the possibility to base estimation (and more generally inference) on the full instead of the partial likelihood. Thus, the penalised spline approach provides a direct estimate for the full survival time distribution. Hennerfeind, Brezger

and Fahrmeir (2006) derive simulation based Bayesian estimation schemes, whereas Kneib and Fahrmeir (2007) consider mixed model based penalised likelihood estimation, extending previous suggestions by Cai, Hyndman and Wand (2002) and Kauermann (2005). Both approaches provide a fully automated estimation procedure, including estimation of smoothing parameters corresponding to the penalised spline functions.

Our model building strategy relies on a two-stage stepwise selection procedure that operates on the level of the covariates on the first stage and within modelling alternatives for a given covariate on the second stage. For example, for binary covariates, the effect may either be time-constant, time-varying or the covariate may have no effect at all. For continuous covariates, a further modelling possibility is provided by a flexible, nonparametric effect. In each selection step, *all* modelling alternatives of *all* covariates not already included in the model are compared based on an optimality criterion that fits the purpose of the data analysis. For example, classical criteria such as Akaike's information criterion (AIC) or the Bayesian information criterion (BIC) can be considered but can be replaced by prediction-oriented measures such as the integrated Brier score (Graf et al., 1999) if prediction is the ultimate purpose of the analysis. In summary, the algorithm combines discrimination between time-constant and time-varying effects as well as selection between linear and nonlinear effects. In addition, it also allows comparison with pre-specified models in terms of the final optimality criterion.

The rest of the paper is organised as follows: Section 2 describes the specification of structured hazard regression models in more detail and outlines mixed model based inference. Section 3 presents the two-stage model selection procedure, which is evaluated in a simulation study in Section 4. Section 5 presents our application together with some adaptations of the basic model building strategy. The final Section 6 discusses and summarises our findings. Details on mixed model based penalised likelihood estimation are presented in Appendix A.

## 2 Structured Hazard Regression

### 2.1 Model Specification

In structured hazard regression models, the hazard rates are related to additive predictors by an exponential link, i.e.

$$\lambda_i(t) = \exp(\eta_i(t)), \quad i = 1, \dots, n, \quad (3)$$

with

$$\eta_i(t) = g_0(t) + \sum_{l=1}^L g_l(t)u_{il} + \sum_{j=1}^J f_j(x_{ij}) + v_i'\gamma. \quad (4)$$

As in the introduction,  $g_l(t)$ ,  $l = 1, \dots, L$ , are time-varying effects of covariates  $u_{il}$ , and  $f_j(x_{ij})$ ,  $j = 1, \dots, J$ , are nonparametric, smooth effects of continuous covariates  $x_{ij}$ , but note that (4) contains the log-baseline effect  $g_0(t) = \log(\lambda_0(t))$  in addition. This modification allows us to derive the full likelihood not only for the regression effects but also for parameters describing the log-baseline hazard.

Let  $T_i$ ,  $i = 1, \dots, n$ , denote the true survival time of observation  $i$  and  $C_i$  the corresponding censoring time. Then, the observed data in case of right-censored observations is given by the censored survival time  $t_i = \min(T_i, C_i)$  and the indicator of non-censoring  $\delta_i = \mathbb{1}(T_i \leq C_i)$ . Under the usual assumption of non-informative and independent censoring, the likelihood contribution for observation  $i$  with hazard rate  $\lambda_i(t)$  is given by

$$L_i = \lambda_i(t_i)^{\delta_i} \exp\left(-\int_0^{t_i} \lambda_i(t) dt\right).$$

Taking logarithms and assuming (conditional) independence between the observations, the complete data log-likelihood is obtained as

$$\sum_{i=1}^n \left[ \delta_i \eta_i(t_i) - \int_0^{t_i} \lambda_i(t) dt \right].$$

Since the log-likelihood involves an integral over the hazard rate, the estimation scheme discussed in the following section will rely on numerical integration to compute the likelihood and its derivatives. In the software we employed for our application, the trapezoidal rule is used to evaluate this integral. The software supports three standard settings for choosing the integration points: (i) using all available (distinct) data points, (ii) using a quantile based grid (employing 60 quantiles and five equidistant points between these

quantiles as default), and (iii) setting up an equidistant grid (with 300 points as default). While the two former choices adapt automatically to the distribution of the data, the latter is more stable with respect to gaps and holes in this distribution that may lead to numerical problems and instable estimates. Kneib and Fahrmeir (2007) compared equidistant and quantile based integration points in a simulation and found only minor differences. In our application, we therefore considered quantile based integration points, corresponding to the standard choice. Note that the numerical integration can be avoided by data augmentation leading to an approximate Poisson likelihood (Kauermann, 2005). However, this possibility comes at the cost of a severely increased sample size while numerical integration allows to work with the original data.

The specification of a structured hazard regression model contains several unknown functions, including the log-baseline hazard  $g_0(t)$ , the time-varying effects  $g_i(t)$  and the non-parametric effects  $f_j(x_j)$ . A flexible yet parsimonious approach to model these functions is to approximate them by polynomial splines, i.e. to assume

$$f(x) = \sum_{m=1}^M \beta_m B_m(x) \quad \text{or} \quad g(t) = \sum_{m=1}^M \beta_m B_m(t), \quad (5)$$

where we dropped the function index for the sake of simplicity. The basis functions  $B_m$  are B-splines of degree  $q$  defined over a grid of equally spaced knots  $\kappa_0 < \kappa_1 < \dots < \kappa_s$ ,  $M = q + s$  (Dierckx, 1993). The basis function approach allows us to replace the nonparametric estimation problem by a parametric analogon where the regression coefficients defining the splines are added to the vector of unknowns. Eilers and Marx (1996) propose to use a rich basis to enable flexibility of the resulting functions but to supplement pure likelihood based estimation with a penalty term that penalises wiggly functions. Instead of a squared  $k$ -th derivative penalty as employed in the popular smoothing spline approach, Eilers and Marx (1996) use a simple approximation based on  $k$ -th order differences of adjacent parameters. The penalty associated with a parameter vector  $\beta = (\beta_1, \dots, \beta_M)'$  is therefore given by

$$\text{pen}(\beta) = \frac{1}{2\tau^2} \sum_{m=k+1}^M \Delta_k(\beta_m)^2 \quad (6)$$

where  $\Delta_1(\beta_m) = \beta_m - \beta_{m-1}$  for first order differences and  $\Delta_2(\beta_m) = \Delta_1(\Delta_1(\beta_m)) = \beta_m - 2\beta_{m-1} + \beta_{m-2}$  for second order differences. Note that in contrast to the standard formulation of penalty terms, we include  $\tau^2$  as an inverse smoothing parameter. Hence, large values of  $\tau^2$  correspond to rough function estimates, whereas small values enforce

smoothness. Automatic determination of  $\tau^2$  is of course an integral part of estimating structured hazard regression models. We will facilitate this task based on a mixed model representation where  $\tau^2$  turns into a random effects variance, providing an explanation for our parameterisation. Mixed model based inference in semiparametric regression models has gained wide attention in recent years, see the introductory paper by Wand (2003) and Ruppert, Wand and Carroll (2003) for a general overview.

## 2.2 Mixed Model Based Inference

Estimation of structured, Cox-type survival models is based on the penalised log-likelihood criterion

$$\sum_{i=1}^n \left[ \delta_i \eta_i(t_i) - \int_0^{t_i} \lambda_i(t) dt \right] - \sum_{l=0}^L \text{pen}(\beta^{(l)}) - \sum_{j=1}^J \text{pen}(\beta^{(j)}).$$

The parameter vectors  $\beta^{(l)}$ ,  $l = 0, \dots, L$  and  $\beta^{(j)}$ ,  $j = 1, \dots, J$  each consist of  $M$  elements and represent the coefficients of the basis functions for  $g_l(\cdot)$  and  $f_j(\cdot)$ , respectively. Based on derivatives of the penalised log-likelihood, we can devise a Newton-Raphson-type algorithm for penalised likelihood estimation. The crucial quantities in obtaining these penalised likelihood estimates are the smoothing parameters  $\tau_l^2$  and  $\tau_j^2$  that control the compromise between fidelity to the data (in terms of the likelihood) and smoothness (in terms of the penalty terms). A full inferential procedure that provides both estimates for the regression coefficients and the smoothing parameters can be derived based on a mixed model interpretation of structured hazard regression models (Cai et al., 2002; Kauer- mann, 2005; Kneib and Fahrmeir, 2007). Therefore, we first have to note that the penalty term associated with a penalised spline can be equivalently interpreted as a special random effects distribution. The quadratic penalty (6) can be represented as a quadratic form

$$\text{pen}(\beta) = \frac{1}{2\tau^2} \beta' K \beta,$$

where the penalty matrix  $K = D'D$  is composed as the cross-product of an appropriate difference matrix  $D$ . Comparing the penalty to a Gaussian distribution with density

$$p(\beta|\tau^2) \propto \exp\left(-\frac{1}{2\tau^2} \beta' K \beta\right) \quad (7)$$

reveals that the penalty essentially equals the negative log-density of a Gaussian random effects distribution. Within that distribution, the smoothing parameter  $\tau^2$  turns into a random effects variance.



A difficulty arising with random effects distribution (7) is its partial impropriety arising from the rank-deficiency of the penalty matrix  $K$ . This rank-deficiency reflects that a  $(k - 1)$ -th order polynomial remains unpenalised when applying a  $k$ -th order difference penalty. Since standard mixed models require proper random effects distributions, a reparameterisation into fixed effects representing the unpenalised part and random effects representing the penalised part has to be applied to all vectors of regression coefficients associated with nonparametric or time-varying effects. Afterwards, the model (4) can be written as

$$\eta_i(t_i) = w(t_i)' \gamma + z(t_i)' \nu \quad (8)$$

where  $\gamma$  comprises the collection of all fixed effects corresponding to covariates  $w(t_i)$ ,  $\nu$  contains random effects corresponding to covariates  $z(t_i)$ , and all random effects variances are collected in the vector  $\theta$  of length  $L + J$ . In terms of the mixed model representation (8), the likelihood and log-likelihood will be denoted as  $L(\gamma, \nu)$  and  $l(\gamma, \nu)$  but actually coincide with the quantities derived above.

The advantage of the mixed model representation is the availability of algorithms for the joint determination of the random effects and of their variances. The approach is based on penalised likelihood estimation for the random effects and marginal likelihood estimation for the variances. The latter employs a Laplace approximation to the marginal likelihood, yielding a simple Newton-Raphson algorithm (see Kneib and Fahrmeir (2007) and Appendix A for details).

The mixed model based estimation procedure is implemented in the software package BayesX, freely available from <http://www.stat.uni-muenchen.de/~bayesx>.

### 3 Model Building

Based on the estimation scheme from the previous section, it is possible to obtain estimates in a model with fixed model structure, i.e. given the modelling specification for the different covariates. However, in practice, several competing modelling strategies exist for different types of covariates. In particular, categorical covariates could be included with linear or time-varying effects. For continuous covariates, a further alternative is to include them with a smooth effect as a penalised spline. In the following section, we develop a two-stage stepwise procedure that provides a means of model building, taking into account

these alternative modelling strategies. The discrimination between modelling alternatives is based on suitable model comparison criteria. Some specific examples will be discussed in Section 3.2.

### 3.1 Two-Stage Stepwise Procedure

Consider the following two-stage stepwise procedure:

**Starting Model:** Define a starting model. Typically this will be the empty model containing only the baseline hazard rate. A non-empty starting model with preset variables can be obtained by modifications described in Section 5.1.

**Initial Choice Set:** Define an initial choice set of covariates not already included in the starting model.

**Model Comparison Criterion:** Choose an appropriate model comparison criterion. See the Section 3.2 for some suggestions.

- [i] **Modelling Alternatives:** For each covariate in the choice set, define a set of modelling alternatives, e.g. linear effect vs. time-varying effect in case of categorical variables, or linear effect vs. nonparametric effect vs. time-varying effect in case of continuous covariates.
- [ii] **Estimation of Models:** For each of the covariates in the choice set and for each modelling possibility, estimate the hazard regression model obtained from the current model by adding the covariate in the respective modelling possibility and store the model comparison criterion.
- [iii] **Selection Step with Stopping Criterion:** If the optimal model comparison criterion obtained in step [ii] improves the model comparison criterion of the current model, replace the current model with the best-fitting model from step [ii], delete the corresponding covariate from the choice set, and go to step [iv]. Otherwise terminate the algorithm.
- [iv] **Backward Deletion:** Perform a backward deletion step on the current model, i.e. estimate all hazard regression models obtained from the current model by dropping one covariate at a time. If an improvement of the model comparison criterion can be achieved, make the reduced model with optimal model comparison criterion the working model and add the deleted variable to the choice set. Continue with step [i].

The selection procedure is called two-stage since it differentiates between inclusion of variables on the first stage and different modelling possibilities for the covariates on the second stage. It proceeds in a stepwise fashion, where each forward step for inclusion of additional terms is followed by a backward deletion step.

A “toy example” that illustrates the application of our two-stage stepwise procedure can be found in Table 1. In this example, we apply the conditional AIC ( $AIC_c$ , see Section 3.2) as model comparison criterion. We start with an empty model containing only the baseline hazard rate and three variables in the *initial choice set* with either two or three *modelling alternatives* (step [i]). For each variable and each modelling possibility, the model is fitted and the  $AIC_c$  is calculated (step [ii]). “Apache II score” modelled as smooth term has the minimal  $AIC_c$ , and thus is added to the *starting model*. In the next iteration, only two variables are left in the choice set. “Palliative operation for malignant disease” is added as linear term as it is the minimizer of  $AIC_c$ . In the last step of this example, age is chosen as linear term. The inclusion of the variable, in each step, improves the  $AIC_c$  of the previous step (i.e.  $AIC_c$  decreases) (step [iii]). Between step 1 and 2 no *backward deletion step* (step [iv]) is needed, as only one variable is included in the model so far. The backward step after step 2 was performed in the usual manner but did not lead to a better model.

Of course, the proposed selection scheme can be modified at some points if recommended by the application at hand. One might, for example, think of starting with a full model instead of the empty model but this approach will suffer from two drawbacks: Firstly, it is not clear which model should be the full model. In particular, for continuous covariates it is unclear whether a full model should contain nonparametric or time-varying effects. Secondly, the full model would typically be overly complex. This would lead to a higher computational burden compared to the proposed strategy. Moreover, it may often be impossible to identify the full model from given data if, for example, the percentage of censoring is high or the number of possible covariates is large.

## 3.2 Model Comparison Criteria

The model comparison criterion that is probably applied most frequently in practice is Akaike’s information criterion (AIC). Since estimation in our structured hazard regression

model is based on mixed model methodology, it seems plausible to base the model building procedure on an AIC for mixed models. However, two different versions of AIC are available: The marginal AIC is based on the marginal likelihood of the mixed model with the random effects integrated out, i.e.

$$\text{AIC}_m = -2 \log \left[ \int L(\gamma, \nu) d\nu \right] + 2[\text{dim}(\gamma) + \text{dim}(\theta)]. \quad (9)$$

It consists of the log-marginal likelihood as pure model fit criterion and the number of fixed effects,  $\text{dim}(\gamma)$ , plus the number of smoothing variances,  $\text{dim}(\theta)$ , as a measure of model complexity. In contrast, the conditional AIC

$$\text{AIC}_c = -2l(\gamma, \nu) + 2 \text{df} \quad (10)$$

is based on the conditional likelihood  $L(\gamma, \nu)$  in combination with the effective degrees of freedom  $\text{df}$  as a complexity measure. The degrees of freedom are defined as the trace of the product of unpenalised Fisher information and inverse penalised Fisher information, see the appendix for details. They replace the usual parameter count since the effective dimensionality reduction induced by the random effects distribution has to be taken into account.

Vaida and Blanchard (2005) (in the context of mixed models) and Wager, Vaida and Kauermann (2007) (in the context of mixed model based penalised spline smoothing) compare both AIC versions and also devise some general rules to judge the appropriateness of either of the two. If the number of random effects is fixed and all observations share the same set of random effects, the conditional AIC is more appropriate. This situation arises in the context of mixed model based smoothing where the random effects are given by the coefficients of the basis functions. Since all observations share the same random effects, there is no necessity to obtain a marginal model comparison measure without the random effects. Moreover, the random effects represent the penalised part of the nonparametric function and integrating them out corresponds to marginalising over parts of the function. Since, however, the nonparametric functions are of major interest in our analyses, a conditional model comparison measure is recommended. Moreover, the conditional AIC coincides with the classical AIC from the smoothing literature as outlined for example in Hastie and Tibshirani (1990).

Note, however, that still the marginal AIC is frequently employed in mixed model based smoothing since it is the sole model comparison criterion delivered by most mixed model

implementations in statistical software. However, it is more appropriate for clustered data, where the random effects are introduced to model correlations between the elements of one cluster. For example, Kauermann et al. (2008) apply the marginal AIC in structured hazard regression models with time-varying effects expressed as random effects and additional random effects for cluster-specific effects (i.e., frailties) to select between different candidate models. Greven and Kneib (2008) investigate conditional and marginal AIC in linear mixed models and detect theoretical defects for both of them. However, their results do not immediately carry over to structured hazard regression models and our empirical experience suggests deviations from the theoretical behaviour predicted by Greven and Kneib (2008). We therefore decided to stick to the conditional AIC in our application but provide some further remarks in the discussion section.

A straightforward alternative to the conditional AIC is the conditional BIC where the penalty parameter is replaced by the logarithm of the sample size  $n$ , yielding

$$\text{BIC}_c = -2l(\gamma, \nu) + \ln(n) \text{ df}. \quad (11)$$

Of course, any other criterion that can be used to evaluate the fit of a model can also be applied. The corrected AIC (Hurvich, Simonoff and Tsai, 1998), for example, can be applied to force the selection of sparser models since it uses a larger penalty for the degrees of freedom when the number of parameters approaches the sample size. The corrected AIC has also been successfully applied in the context of hazard regression models with random effects (Therneau and Grambsch, 2000), making it suitable for mixed model based penalised spline estimation as considered in this paper.

The (integrated) Brier score can be used to assess the predictive power of competing models and thus to distinguish between different models (Gerds and Schumacher, 2006). The mean squared error of prediction in survival models, i.e. the Brier score, can be estimated as

$$\widehat{\text{Bs}}(t) = 1/n \sum_{i=1}^n \left[ I(T_i^* > t) - \hat{S}(t|x_i^*) \right]^2 \cdot w_i(t), \quad (12)$$

where the Brier score is evaluated for an arbitrary but fixed value  $t$ ,  $\hat{S}(t|x_i^*)$  is the survivor function estimated on the training data and evaluated for a new observations  $x_i^*$ ,  $i = 1, \dots, n$ . The indicator function  $I(\cdot)$  is 1, if the time-point  $t$  is before the observed event time  $T_i^*$  of the new observation  $x_i^*$  and 0 otherwise. The weights  $w_i(t)$  depend on the censoring distribution (for details see Gerds and Schumacher, 2006). The Brier score now

can be plotted against time. To have a summary measure, the integral over time of the Brier score can be used (Graf et al., 1999).

Further examples of available model comparison criteria include scoring rules as discussed in Gneiting and Raftery (2007).

## 4 Simulation Study

To assess the properties of the proposed algorithm, we conducted a small simulation study employing the conditional AIC (10) and the conditional BIC (11) as model comparison criteria. We generated 100 independent data sets, each with 400 observations, for both criteria. The simulations apply a generalized version of the sampling algorithm proposed by Bender, Augustin and Blettner (2005) and are based on the hazard rate

$$\lambda(t, \mathbf{x}) = \exp\left(\log(t + 0.01) + 0.9 \cdot x_1 + \sin(-x_2^2 - 0.6x_2^3) + 0.5\sqrt{t} \cdot x_3\right)$$

with time  $t$  and covariates  $\mathbf{x} = (x_1, \dots, x_5)'$ . Thus, we have five possible covariates but only three have an effect on the survival time  $T_i$ . To be more specific,  $x_1$  enters in linear form, the effect of  $x_2$  has a flexible, nonlinear form, and  $x_3$  has a time-varying effect. The covariates are sampled independently from uniform distributions on  $[-1, 1]$ , i.e.

$$x_1, \dots, x_5 \stackrel{\text{i.i.d.}}{\sim} U[-1, 1].$$

The observed survival time is then defined as  $t_i = \min(T_i, C_i)$ , where censoring is introduced through Expo(8) distributed censoring times  $C_i$ . This leads to roughly 80% to 90% non-censored observations.

Table 4 summarizes the modelling alternatives selected for each of the covariates based on the conditional AIC and the conditional BIC. Since the different modelling alternatives are mutually exclusive, all rows sum up to 100 (the total number of simulation replications). Comparing results for  $AIC_c$  and  $BIC_c$ , we find the expected behaviour that  $BIC_c$  favors sparser models. In particular, the spurious covariates  $x_4$  and  $x_5$  are never selected by  $BIC_c$ , whereas a small number of false decisions is found for  $AIC_c$ . On the other hand, the  $BIC_c$  also tends to select a linear effect for  $x_3$  instead of the correct time-varying effect. Hence, despite finding the relevant covariates in most simulation replications,  $BIC_c$  tends to favor models without time-varying effects. Based on  $AIC_c$ , we find a similar tendency towards linear effects for  $x_3$  (42 out of 100) but the number of correct decisions increases

quite a bit (from 11 to 43). It is also noteworthy that based on  $AIC_c$  the modelling alternative chosen most frequently is the correct one for all covariates. Furthermore, 100 out of 100 replications yielded the correct choice of a flexible effect for  $x_2$ . Both  $AIC_c$  and  $BIC_c$  eventually identify the influential covariates in all replications but sometimes mistake the correct form of the influence (both,  $AIC_c$  and  $BIC_c$ ) and may introduce a small number of spurious covariates ( $AIC_c$ ).

[Table 1 about here.]

While Table 4 presents marginal frequencies of modelling alternatives for the covariates, it is of course also of interest to identify complete model specifications (i.e. combinations of modelling alternatives for the covariates) that are frequently chosen by our procedure. Table 5 shows the frequency of replications where the correct model was chosen together with the models that are chosen most frequently. Based on  $BIC_c$ , only a very small number of different models has been identified (in at least two simulation replications). Only nine replications led to the correct model. Note however, that this small number is simply resulting from the fact that only 11 replications identified the time-varying effect of  $x_3$  and therefore 11 is also the maximal number of correct model identifications that could be achieved. The model most frequently selected by  $BIC_c$  is correct apart from the misspecification for  $x_3$ . Based on  $AIC_c$ , the number of different models deemed as optimal (in at least two replications) strongly increases. The model chosen most frequently now (in 14 out of 100 cases) is actually the correct one, followed by models where either the effect of  $x_3$  is mistaken to be linear (14 cases) or the effect of  $x_1$  is mistaken to be time-varying (11 cases). All other modelling possibilities appear less than five times within the simulation. Again, it turns out that all influential variables are found in most of the simulation replication.

[Table 2 about here.]

Inspection of the estimated effects (omitted for the sake of brevity) of  $x_1$  (truly linear) and  $x_2$  (truly smooth) showed that the true effects were estimated very well. Minor deviations, especially in the case where time-varying effects were selected for  $x_1$ , can at least partly be attributed to the heavily skewed distribution of the observed survival times  $t$ . The skewness of the marginal distribution of  $t$  is also responsible for difficulties

regarding the detection and estimation of time-varying effects. Since the right tail of the distribution is very sparse, estimation of the effect is highly variable in this region or might bias the whole estimation (e.g. Gray, 1992). Problems with the detection and estimation of time-varying effects can be observed in many other publications, even in very manageable simulation settings. For example, Abrahamowicz and MacKenzie (2007) estimate an interaction surface of a smooth and a time-varying effect were the other three covariates are *a priori* known to have a linear effect on the log hazard. Despite this prior knowledge, there is considerable variation in the estimation of the effects. Kauermann et al. (2008) use a setting similar to ours, where the log-baseline hazard and time-varying effects are both modeled using penalised splines. Their simulation setting consists of one time-varying effect for a binary covariate and an additional frailty term, which can be simultaneously modeled with their proposed stacked Laplace-EM algorithm. Again, high variability for the estimates of time-varying effects is observed.

In summary, our simulation study supports that the two-stage stepwise model building procedure has the potential to detect correct functional forms in survival models. The correct detection of time-varying effects is less powerful, but this seems to be attributable to the increased difficulty in this task that has already been observed in previous attempts. One of the main advantages of the two-stage procedure lies in the fact that all resulting models are easily interpretable since each covariate enters in exactly one functional form and no combinations such as time-varying and flexible effects of the same covariate are considered. Note also that the final model delivered by the two-stage stepwise approach should only be considered a candidate model delivered by an automatic model building strategy. Instead of investigating the possibility to detect the “true model” (which will probably not exist in reality anyway), the results of our simulation study should provide some evidence that the proposed procedure detects reasonable models for further investigation.

The model comparison criterion employed should be chosen to meet the requirements of the study performed. If variable selection is of particular interest,  $BIC_c$  may be used to identify these variables, despite its drawback of favoring simpler, linear models over models with time-varying effects. If, on the other hand, it is more important to correctly identify the functional forms of the covariate effects of the influential variables,  $AIC_c$  may be favored, despite its tendency to introduce some spurious covariates. Since we



are particularly interested in time-varying effects in our application, we considered the (conditional) AIC instead of (conditional) BIC in the following. Note also that it may well be useful to investigate the performance of a specific comparison measure (or other settings in the two-stage stepwise procedure) in simulations designed to meet the data under consideration. Although general recommendations seem to be difficult to derive, this will then allow to choose a sensible model building strategy for the given application.

## **5 Application: Prognostic model for surgical patients with severe sepsis**

Our analysis was based on data from a database, which was initiated in 1993 in the Intensive Care Unit, Department of Surgery, Campus Großhadern, LMU Munich, Germany, for benchmarking and quality control. The documentation period started on March 1st, 1993, and lasted until February 28th, 2005. We could obtain relevant covariates reflecting the state of the patient on admission day, and 90-day survival time in 462 patients with severe sepsis. We aimed to build a prognostic model to assess which variables have an influence on the prognosis of the patients. The purpose was not to make individual predictions for each patient but to identify influential prognostic factors and to understand how these act on the survival time.

### **5.1 Starting model**

Based on subject-matter knowledge, six of the covariates included in the original choice set should be definitely included in the final model. We therefore adapted the two-stage selection strategy from the previous section to account for this fact. In a first step, we derived a starting model, where only the modelling possibility had to be chosen for each of the fixed covariates. This is implemented by accepting the best-fitting modelling alternative in step [iii] even if it does not reduce the  $AIC_c$  of the previous model. Application of the modified procedure to our data, led to a model containing the effects presented in the upper part of Table 2 (in the order of inclusion). As one can see in the last two steps, we did not use a stopping criterion (e.g. increase of  $AIC_c$ ) for the selection of the starting model. This is because we just wanted to detect suitable modelling alternatives but do

not select variables. We will provide further discussion of the effects of variables in the starting model along the results of the final model in the next section.

## 5.2 Prognostic model

After deciding on the optimal starting model, we derived the final prognostic model by applying the original two-stage procedure to the choice set of remaining covariates. Note that the fixed covariates are of course not subject to deletion in the backward selection step [iv].

Table 2 shows the full choice set, the model alternative selected in the stepwise procedure, and the conditional AIC obtained in the corresponding inclusion step. The covariate set mainly represents variables indicating renal, pulmonary and cardio-circulatory function, and nature and severity of the underlying disease. While most covariate labels are self-explanatory, some of them require supplementary explanation: Apache II score is a measure for the severity of disease determined within the first 24 hours of admission and the Horowitz ratio ( $P_aO_2/F_iO_2$ ) describes the quality of lung function by referring the arterial partial oxygen pressure to the corresponding inspiratory oxygen concentration.

The effects of the selected smooth terms are plotted in Figure 1 and time-varying terms are shown in Figure 2. In the latter, one can clearly see the differences in the shapes of the log-baseline hazard rate for the variable “fungal infection” (present vs. absent). In particular, Figure 2 also reveals the ability of structured hazard regression models to address the problem of nonproportional hazards.

Table 3 shows the linear effects of the variables included in the prognostic model.

[Figure 1 about here.]

[Figure 2 about here.]

Relevant risk factors ( $p < 0.05$ ) for a shorter survival time were palliative operation for malignant disease, age, creatinine concentration at admission, year of therapy and operation for thoracic diseases. According to 95% pointwise confidence intervals of smooth and time-varying terms, also a high Apache II score on admission day, and the nature and localization of the infection (fungal infection, peritonitis) were associated with an increased mortality. Gender, haemoglobin concentration, the need for renal replacement

or catecholamine therapy, the Horowitz ratio or the primary diagnosis of a malignant disease were not associated with mortality.

Our results show that a clinically plausible prognostic model can be constructed based on the suggested algorithm. This model can help the physician in charge to judge the true relevance of various clinical variables for patient outcome, and to adjust his therapeutic concepts according to individual risk profiles.

## 6 Summary and Discussion

We have developed a sensible model building strategy for structured, Cox-type hazard regression models that combines selection of relevant covariates with the determination of the appropriate modelling alternative in a two-stage stepwise procedure. It therefore allows to investigate time-constant and time-varying effects in one step, making it particularly attractive for model building in the survival analysis context. The two-stage procedure can, however, also be applied in any other type of regression models where multiple modelling alternatives exist for a covariate. A typical example are generalised additive models, where linear vs. non-linear modelling of continuous covariates is of interest. Further modelling alternatives may include specific forms of interactions, such as surfaces or varying coefficient terms. Note, however, that in this case two components have to be treated simultaneously in the modelling alternatives. For example, in case of an interaction between two continuous covariates, modelling alternatives would include effects of only one single covariate, an additive combination of both covariates and a completely nonparametric interaction surface (to name only a few). In our application, we defined the modelling alternatives in cooperation with our collaborators from the Department of Surgery and, as a consequence, restricted our attention to models without interaction effects.

Employing penalised splines as a major modelling component allows for considerable flexibility in estimating semiparametric regression models. In particular, time-varying effects of qualitatively very different functional forms can be treated within one framework. In contrast to the artificial covariate approach, penalised splines avoid the need to specify a global functional form for time-varying effects. Fractional polynomials fall in between the purely parametric approach and penalised spline estimation and would be an alternative

possibility to define modelling alternatives in the two-stage stepwise procedure. This could for example be useful in the determination of time-varying effects where the penalised spline approach had a tendency towards simple linear effects.

One of the main benefits of the two-stage procedure lies in the fact that it ensures that any selected covariate is included in exactly one modelling alternative. Hence it yields interpretable models and avoids models that contain, for example, combinations of time-varying and flexible effects of the same covariate.

Model comparison is accomplished based on a criterion such as the conditional AIC, allowing for the comparison of all models fitted throughout the iterative model building process. In addition, the final model can be compared to alternatives derived with competing procedures or from subject-matter knowledge. Note that the proposed procedure leaves the choice of the appropriate criterion to the user. While our simulation supports that the conditional AIC seems to be reasonable criterion in the context of hazard regression models with time-varying effects, other choices may be deemed preferable in other contexts. Moreover, recent research has shown that the conditional AIC has some severe theoretical defects in simple linear mixed models (Greven and Kneib, 2008) that may make its use also questionable in hazard regression.

In our application, we employed the model building procedure to determine a suitable prognostic model for 90-day survival of patients with severe sepsis. Another area where the proposed strategy can be applied is the derivation of confounder models (Moubarak et al., 2008). In this case, the two-stage procedure can be employed to determine the confounder model from a set of covariates excluding the covariate of primary interest. After performing model selection, the effect of this covariate can be assessed by additional inclusion in the chosen confounder model.

## **A Details on Penalised Likelihood Estimation**

This appendix summarises details on likelihood based estimation of structured hazard regression models. Further details can be found in Kneib and Fahrmeir (2007).

In vector matrix notation, the predictor (8) can be written as

$$\eta = W\gamma + Z\nu.$$

Accordingly, the penalised log-likelihood is then given by

$$l_p = \delta' \eta - \mathbf{1}' \Lambda - \frac{1}{2} \nu' \Sigma^{-1} \nu$$

where  $\delta$  refers to the vector of censoring indicators,  $\Lambda$  contains the evaluations of the cumulative hazard rate

$$\Lambda(t_i) = \int_0^{t_i} \lambda_i(t) dt,$$

$\mathbf{1}$  is a vector of ones and  $\Sigma$  corresponds to the covariance matrix of the random effects  $\nu$ .

Taking first and second derivatives yields the score function

$$s = \begin{pmatrix} s_\gamma \\ s_\nu \end{pmatrix} = \begin{pmatrix} \frac{\partial l_p}{\partial \gamma} \\ \frac{\partial l_p}{\partial \nu} \end{pmatrix} = \begin{pmatrix} \delta' W - \mathbf{1}' \frac{\partial \Lambda}{\partial \gamma} \\ \delta' Z - \mathbf{1}' \frac{\partial \Lambda}{\partial \nu} - \Sigma^{-1} \nu \end{pmatrix}$$

and the Fisher information matrix

$$H = \begin{pmatrix} H_{\gamma\gamma} & H_{\gamma\nu} \\ H_{\nu\gamma} & H_{\nu\nu} \end{pmatrix} = \begin{pmatrix} \mathbf{1}' \frac{\partial^2 \Lambda}{\partial \gamma \partial \gamma'} & \mathbf{1}' \frac{\partial^2 \Lambda}{\partial \gamma \partial \nu'} \\ \mathbf{1}' \frac{\partial^2 \Lambda}{\partial \nu \partial \gamma'} & \mathbf{1}' \frac{\partial^2 \Lambda}{\partial \nu \partial \nu'} + \Sigma^{-1} \end{pmatrix}. \quad (13)$$

Both quantities rely on first and second derivatives of the cumulative hazard rate that are (for example) given by

$$\frac{\partial}{\partial \gamma_j} \Lambda(t_i) = \int_0^{t_i} w_{ij}(t) \lambda_i(t) dt$$

and

$$\frac{\partial}{\partial \gamma_j \partial \gamma_k} \Lambda(t_i) = \int_0^{t_i} w_{ij}(t) w_{ik}(t) \lambda_i(t) dt$$

and analogous expressions when random effects  $\nu$  are involved. All integrals have to be evaluated numerically, relying on the trapezoidal rule for simplicity. Score function and Fisher information form the basis for a Newton-Raphson algorithm for the determination of the regression coefficients (for given variance parameters).

The Fisher information matrix also forms the basis for determining the degrees of freedom for the estimated model that have been employed in the definition of the conditional AIC and the conditional BIC. Let

$$\tilde{H} = \begin{pmatrix} \mathbf{1}' \frac{\partial^2 \Lambda}{\partial \gamma \partial \gamma'} & \mathbf{1}' \frac{\partial^2 \Lambda}{\partial \gamma \partial \nu'} \\ \mathbf{1}' \frac{\partial^2 \Lambda}{\partial \nu \partial \gamma'} & \mathbf{1}' \frac{\partial^2 \Lambda}{\partial \nu \partial \nu'} \end{pmatrix}$$

denote the Fisher information derived from the unpenalised log-likelihood. Then, the effective degrees of freedom are defined as

$$\text{df} = \text{trace} \left( \tilde{H} H^{-1} \right),$$

see Gray (1992).

Estimation of the variance parameters is based on an approximate marginal likelihood procedure, where a Laplace type approximation to the marginal log-likelihood

$$l^{marg}(\Sigma) = \log \left[ \int L_p(\gamma, \nu, \Sigma) d\gamma d\nu \right]$$

is considered, yielding

$$l^{marg}(\Sigma) \approx -\frac{1}{2} \log |\Sigma| - \frac{1}{2} \log |H| - \frac{1}{2} \nu' \Sigma^{-1} \nu.$$

First and second derivatives of this approximate marginal log-likelihood can be derived based on matrix differentiation rules, again allowing to set up a Newton-Raphson type estimation procedure.

## References

- Abrahamowicz, M. and MacKenzie, T. A. (2007). Joint estimation of time-dependent and non-linear effects of continuous covariates on survival, *Statistics in Medicine* **26**: 392–408.
- Bender, R., Augustin, T. and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models, *Statistics in Medicine* **24**: 1713–1723.
- Cai, T., Hyndman, R. and Wand, M. (2002). Mixed model-based hazard estimation, *Journal of Computational and Graphical Statistics* **11**: 784–798.
- Cox, D. R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**: 187–220.
- Dierckx, P. (1993). *Curve and Surface Fitting with Splines*, Clarendon Press, Oxford.
- Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties, *Statistical Science* **11**: 89–121.
- Gerds, T. and Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times, *Biometrical Journal* **48**: 1029–1040.
- Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association* **102**: 359–378.

- Graf, E., Schmoor, C., Sauerbrei, W. and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data, *Statistics in Medicine* **18**: 2529–2545.
- 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.
- Greven, S. and Kneib, T. (2008). On the behavior of marginal and conditional akaike information criteria in linear mixed models, *Technical Report 179*, Johns Hopkins University, Dept. of Biostatistics.
- Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall / CRC , London.
- Hennerfeind, A., Brezger, A. and Fahrmeir, L. (2006). Geoadditve survival models, *Journal of the American Statistical Association* **101**: 1065–1075.
- Hurvich, C., Simonoff, J. and Tsai, C. (1998). Smoothing parameter selection in non-parametric regression using an improved Akaike information criterion, *Journal of the Royal Statistical Society, Series B* **60**: 271–293.
- Kahn, J., Kramer, A. and Rubenfeld, G. (2007). Transferring critically ill patients out of hospital improves the standardized mortality ratio: a simulation study, *Chest* **131**: 68–75.
- Kauermann, G. (2005). Penalised spline fitting in multivariable survival models with varying coefficients, *Comput. Statist. Data Anal.* **49**: 169–186.
- Kauermann, G., Xu, R. and Vaida, F. (2008). Stacked Laplace-EM algorithm for duration models with time-varying and random effects, *Computational Statistics & Data Analysis* **52**: 2514–2528.
- Kneib, T. and Fahrmeir, L. (2007). A mixed model approach for geoadditve hazard regression, *Scandinavian Journal of Statistics* **34**: 207–228.
- Moubarak, P., Zilker, S., Wolf, H., Hofner, B., Kneib, T., Küchenhoff, H., Jauch, K.-W. and Hartl, W. H. (2008). Activity-guided antithrombin III therapy in severe surgical sepsis: Efficacy and safety according to a retrospective data analysis, *Shock* **30**: 634–641.

- Ruppert, D., Wand, M. and Carroll, R. (2003). *Semiparametric regression*, Cambridge University Press.
- Sauerbrei, W. and Royston, P. (1999). Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials, *Journal of the Royal Statistical Society, Series A* **162**: 71–94.
- Sauerbrei, W., Royston, P. and Look, M. (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.
- Therneau, T. M. and Grambsch, P. M. (2000). *Modeling survival data: Extending the Cox model*, Springer, New York.
- Vaida, F. and Blanchard, S. (2005). Conditional akaike information for mixed-effects models, *Biometrika* **92**: 351–370.
- Wager, C., Vaida, F. and Kauermann, G. (2007). Model selection for penalized spline smoothing using akaike information criteria, *Australian and New Zealand Journal of Statistics* **49**: 173–190.
- Wand, M. P. (2003). Smoothing and mixed models, *Computational Statistics* **18**: 223–249.



Variable (stage 1)	Modelling Alternative (stage 2)	AIC <sub>c</sub> in step		
		1	2	3
Apache II score (continuous)	linear	3188.09	–	–
	smooth	<b>3186.21</b>	–	–
	time-varying	3188.37	–	–
palliative operation for malignant disease (categorical)	linear	3530.43	<b>3176.31</b>	–
	time-varying	3532.26	3177.98	–
age (continuous)	linear	3524.45	3178.18	<b>3168.55</b>
	smooth	3525.74	3178.37	3168.58
	time-varying	4073.94	3697.34	3685.98

Table 1: “Toy example” for the two-stage stepwise procedure (simplified version of the first three selection steps of the starting model from our application, cf. Sec. 5).

Step	Covariate	Type of Covariate	Modelling Alternative used	AIC <sub>c</sub>
S1	Apache II score	continuous	smooth term	3186.21
S2	palliative operation for malignant disease	categorical	linear term	3176.31
S3	age (in years)	continuous	linear term	3168.55
S4	treatment period (after 2002 = 1)	categorical	linear term	3159.74
S5	malignant primary diseases	categorical	linear term	3161.22
S6	sex (male = 1)	categorical	linear term	3163.17
1	Horowitz ratio	continuous	smooth term	3011.41
2	haemoglobin concentration (in g/dl)	continuous	smooth term	2970.52
3	fungal infection	categorical	time-varying term	2942.82
4	creatinine concentration (in mg/dl)	continuous	linear term	2922.19
5	need for catecholamine therapy	categorical	linear term	2913.80
6	peritonitis	categorical	time-varying term	2909.04
7	surgery for thoracic disease	categorical	linear term	2905.37
8	need for renal replacement therapy	categorical	linear term	2904.85
–	direct postoperative admission	categorical	time-varying term	2906.02
–	need for artificial ventilation	categorical	linear term	2906.43
–	systolic blood pressure (in mmHg)	continuous	linear term	2906.44
–	readmission	categorical	linear term	2906.55
–	pneumonia	categorical	linear term	2906.55
–	emergency admission	categorical	linear term	2906.84

Table 2: Variables in the starting model (S1 – S6); Variables in the choice set that were included (1 – 8) and variables in the choice set that were not included (–); Where applicable, step of inclusion, model alternative used at inclusion and conditional AIC at inclusion are given. For variables not included in the model the best-fitting alternative and corresponding AIC<sub>c</sub> in the last step of the procedure are given. Unless stated otherwise, categorical variables are dummy-coded with yes = 1.

Variable	$\beta$	Std. Dev.	p-value	95% conf. interval
const	-6.361	0.5239	< 0.001	-7.3882 -5.3342
palliative operation for malignant diseases	0.684	0.2021	0.001	0.2874 1.0798
age (in years)	0.018	0.0051	< 0.001	0.0081 0.0282
treatment period (after 2002 = 1)	-0.527	0.1465	< 0.001	-0.8139 -0.2393
malignant primary disease	-0.176	0.1478	0.233	-0.4660 0.1133
sex (male = 1)	-0.065	0.1370	0.636	-0.3333 0.2038
creatinine concentration (in mg/dl)	0.116	0.0498	0.020	0.0186 0.2140
need for catecholamine therapy	0.349	0.3282	0.288	-0.2947 0.9923
surgery for thoracic diseases	0.559	0.2086	0.008	0.1497 0.9677
need for renal replacement therapy	-0.522	0.3534	0.140	-1.2148 0.1708

Table 3: Linear effects for prognostic model presented in the order of inclusion. Unless stated otherwise, categorical variables are dummy-coded with yes = 1.

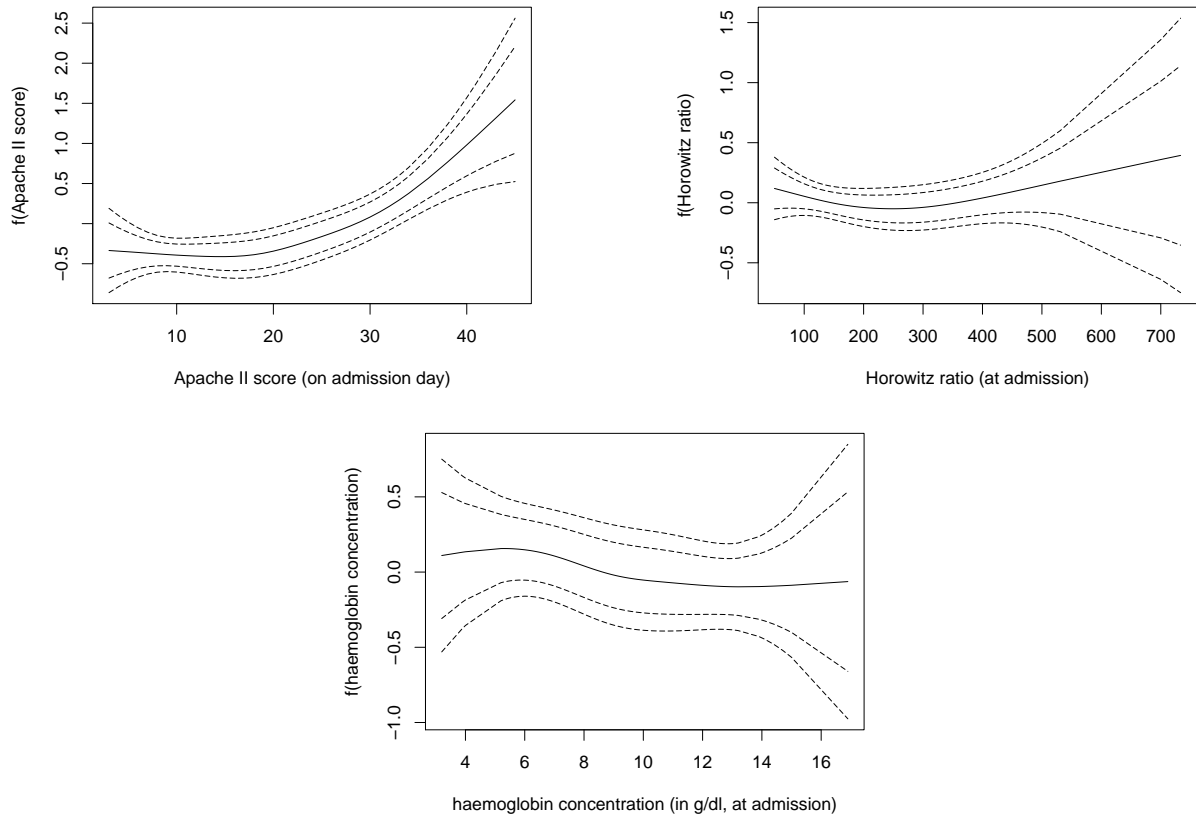


Figure 1: Smooth terms for “Apache II score” (pre-selected in the starting model), “Horowitz ratio”, and “haemoglobin concentration” in the prognostic model; Dashed lines are 80% and 95% point-wise confidence intervals.

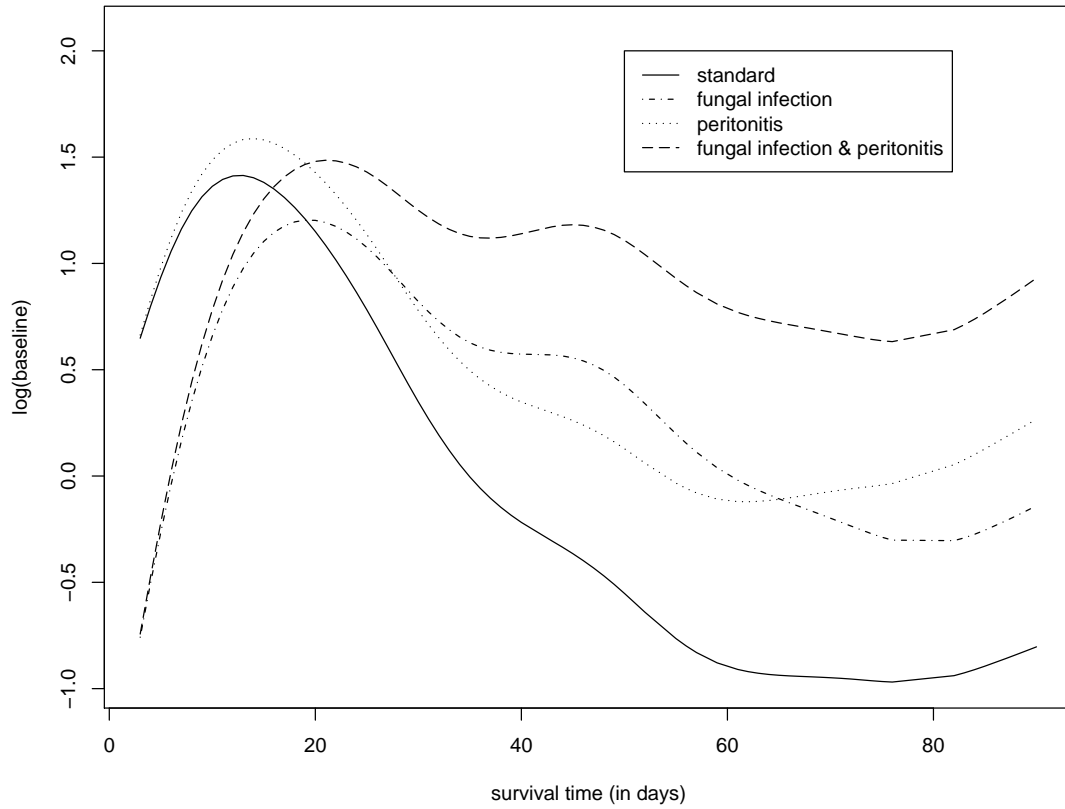


Figure 2: log(baseline hazard rate) in subgroups defined by fungal infection (present vs. absent) and peritonitis (present vs. absent).

AIC <sub>c</sub>				
	linear	smooth	time-varying	not selected
$x_1$ (true: linear)	<b>63</b>	9	28	0
$x_2$ (true: smooth)	0	<b>100</b>	0	0
$x_3$ (true: time-varying)	42	15	<b>43</b>	0
$x_4$	7	2	11	<b>80</b>
$x_5$	7	11	12	<b>70</b>

BIC <sub>c</sub>				
	linear	smooth	time-varying	not selected
$x_1$ (true: linear)	<b>98</b>	0	2	0
$x_2$ (true: smooth)	10	<b>75</b>	0	15
$x_3$ (true: time-varying)	88	0	<b>11</b>	1
$x_4$	0	0	0	<b>100</b>
$x_5$	0	0	0	<b>100</b>

Table 4: Selection frequencies of model components using AIC<sub>c</sub> (top) or BIC<sub>c</sub> (bottom) as inclusion criterion. Correctly selected model components are printed in bold face.

AIC <sub>c</sub>					
$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	
linear	smooth	time-varying	–	–	15
linear	smooth	linear	–	–	14
time-varying	smooth	time-varying	–	–	11
time-varying	smooth	linear	–	–	4
linear	smooth	time-varying	time-varying	–	4
linear	smooth	smooth	–	smooth	4
linear	smooth	smooth	–	–	4
linear	smooth	linear	–	time-varying	3
linear	smooth	linear	–	linear	3
smooth	smooth	time-varying	–	–	3
linear	smooth	time-varying	–	time-varying	3
linear	smooth	linear	–	smooth	3
time-varying	smooth	time-varying	linear	–	2
time-varying	smooth	linear	–	linear	2
time-varying	smooth	linear	linear	–	2
time-varying	smooth	time-varying	–	time-varying	2
linear	smooth	linear	time-varying	–	2
linear	smooth	linear	smooth	smooth	2
smooth	smooth	time-varying	linear	–	2
smooth	smooth	linear	–	–	2

BIC <sub>c</sub>					
$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	
linear	smooth	time-varying	–	–	9
linear	smooth	linear	–	–	65
linear	–	linear	–	–	13
linear	linear	linear	–	–	9

Table 5: Selection frequencies of models: Model selection based on AIC<sub>c</sub> (top) and BIC<sub>c</sub> (bottom); Only models with selection frequency > 1 are given. First row: True model.