Lisa Möst and Torsten Hothorn*

Conditional Transformation Models for Survivor Function Estimation

Abstract: In survival analysis, the estimation of patient-specific survivor functions that are conditional on a set of patient characteristics is of special interest. In general, knowledge of the conditional survival probabilities of a patient at all relevant time points allows better assessment of the patient's risk than summary statistics, such as median survival time. Nevertheless, standard methods for analysing survival data seldom estimate the survivor function directly. Therefore, we propose the application of conditional transformation models (CTMs) for the estimation of the conditional distribution function of survival times given a set of patient characteristics. We used the inverse probability of censoring weighting approach to account for right-censored observations. Our proposed modelling approach allows the prediction of patient-specific survivor functions. In addition, CTMs constitute a flexible model class that is able to deal with proportional as well as non-proportional hazards. The well-known Cox model is included in the class of CTMs as a special case. We investigated the performance of CTMs in survival data analysis in a simulation that included proportional and non-proportional hazard settings and different scenarios of explanatory variables. Furthermore, we re-analysed the survival times of patients suffering from chronic myelogenous leukaemia and studied the impact of the proportional hazards assumption on previously published results.

Keywords: component-wise boosting, conditional survivor function, Cox model, inverse probability of censoring weights, prediction of survival probabilities

DOI 10.1515/ijb-2014-0006

1 Introduction

The estimation of a patient's individual survival probabilities over time is a key aspect of survival analysis. Technically, we are interested in estimating the conditional survivor function, i.e. the probability of surviving up to a specific time point t, conditional on a set of patient-specific explanatory variables. However, common regression models for censored data seldom focus on the direct estimation of the conditional survivor function. Instead, the models concentrate either on the estimation of hazard functions or on summary statistics. In the omnipresent Cox proportional hazards model [1], the conditional hazard function is estimated by cleverly treating the baseline hazard function as a nuisance parameter. Only in a second step can the corresponding conditional survivor functions be derived from this model, for example, by using the Breslow estimator (e.g. [2]). Hence, if one is interested in the conditional survival probabilities, methods for the direct estimation of the conditional survivor function are required.

Moreover, assumptions associated with common modelling strategies for survival data are restrictive. For example, the Cox model is based on the assumption of proportional hazards, the proportional odds model assumes constant odds ratios over time and in the parametric accelerated failure time model, log-transformed responses imply survival times that are, e.g. log-normal-distributed or log-logistic distributed. Although remedies are available, such as stratified Cox models or time-varying effects [3–6], and although model diagnostics (e.g. based on Schoenfeld residuals or formal misspecification tests [7, 8]) and

***Corresponding author: Torsten Hothorn,** Institut für Sozial- und Präventivmedizin, Universität Zürich, Abteilung Biostatistik Hirschengraben 84, Zürich CH-8001, Switzerland, E-mail: Torsten.Hothorn@R-project.org
Lisa Möst, LMU München, Institut für Statistik, München, Germany, E-mail: lisa.moest@stat.uni-muenchen.de

particularly tests for the proportional hazards assumption (e.g. based on cumulative sums of martingale-based residuals or weighted residuals [9, 10]) help to detect unrealistic assumptions, models making less strong assumptions would be widely welcomed.

We suggest estimating the conditional distribution function of the survival times T given a set of patient characteristics $\boldsymbol{x}$ directly in terms of conditional transformation models (CTMs). CTMs have been presented recently in Hothorn et al. [11] and allow the direct and semiparametric estimation of the conditional distribution function $\mathbb{P}(T \leq t|\boldsymbol{X} = \boldsymbol{x})$ under rather weak assumptions. The general model class includes both the proportional odds model and the proportional hazards model as special cases. Nevertheless, the strict assumptions of proportional hazards or proportional odds are relaxed in CTMs. This is achieved by including interaction terms between the survival time and the explanatory variables. For example, the CTM framework allows for varying explanatory variable effects on the hazard function and hence is able to estimate non-proportional hazards as well. However, this advantage comes at the price of a more complex model, which is not easily communicated by simple parameter estimates or even p-values. Graphic approaches are needed to interpret the model, but we can always fall back on the classical approach when the more flexible model suggests that it is safe to assume proportional hazards. P-values or confidence intervals cannot be obtained based on large sample theory, but can be simulated using bootstrap approaches instead.

Transformation models play an important role in survival analysis. The one-to-one correspondences between the proportional hazards and proportional odds model to linear transformation models have already been established in Doksum and Gasko [12] and Cheng et al. [13]. Cheng et al. [14] extended the model class to semiparametric transformation models for failure times. Chen et al. [15] introduced a unified estimation procedure for the analysis of censored data using linear transformation models, and Zeng and Lin [16] proposed a class of semiparametric transformation models to characterize the effects of possibly time-varying covariates on the intensity functions of counting processes. For the estimation of the crude failure probabilities of a competing risk, conditional on explanatory variables, Fine [17] proposed a semiparametric transformation model. These approaches are based on generalized estimation equations. Our approach uses component-wise gradient boosting methodology for model fitting. This approach has the advantage that it incorporates variable selection and shrinkage of coefficient estimates into the model fitting process. These regularization techniques for regression models are necessary for the estimation of survival probabilities because patient characteristics are often highly correlated. Hence, prediction accuracy for the survival probabilities can usually be improved if only a subset of the available patient characteristics is incorporated into the prediction formula. Owing to the component-wise fitting procedure, the algorithm can deal with high-dimensional data. Variable selection in high-dimensional survival data has also been brought up by Lee et al. [18] and Van der Vaart and van der Laan [19]. Lu and Li [20] previously derived a component-wise boosting algorithm for the analysis of survival data in terms of nonlinear transformation models.

Fully nonparametric estimation of the conditional survivor function has also been considered in the past. Making no assumptions about the form of the survivor function can be advantageous over parametric or semiparametric approaches as the underlying assumptions may be violated. Furthermore, nonparametric approaches can be used to check whether one of the more restrictive parametric or semiparametric submodels provides a good fit to the data. The well-known product limit estimator introduced by Kaplan and Meier [21] enables nonparametric estimation of the unconditional survivor function. Dabrowska [22], Dabrowska [23], González Manteiga and Cadarso-Suarez [24] and Iglesias Pérez and González Manteiga [25] present generalizations of the product limit estimator by introducing kernel-based weights to estimate the *conditional* survivor function nonparametrically. In the light of counting process theory, McKeague and Utikal [26] propose a general counting process regression model for estimating conditional survivor functions, and Li and Doss [27] propose a class of estimators for the conditional survivor function based on a fully nonparametric model. The usage of local linear estimators for the conditional survivor function is suggested by Spierdijk [28].

In contrast to kernel-based methods, tree-based approaches and especially random forests can be used to estimate conditional distribution functions precisely without relying on strict model assumptions. For

right-censored data, Hothorn et al. [29] introduced a forest aggregation scheme that produces estimates of the conditional survivor function. The same scheme was used later by Meinshausen [30] for uncensored observations; an alternative forest variant (random survival forests) was introduced by Ishwaran et al. [31]. Conditional inference forests [32], based on an aggregation of conditional inference trees [33], use the aggregation scheme introduced by Hothorn et al. [29] and have been shown to perform akin to other forest variants for right-censored data [34] and were used as a completely nonparametric competitor for CTMs in our study here.

Another useful alternative to the Cox model or to linear transformation models is censored quantile regression (e.g. [35–41]). With this approach, the conditional quantiles of the survival times are modelled in terms of regression models. In contrast to our proposed CTM approach, not all conditional quantiles of the survival times are modelled simultaneously but are instead modelled separately. Hence, quantile crossing [42] is a potential problem of this procedure.

In order to illustrate CTMs for survival data, we checked the validity of the proportional hazards assumption in a re-analysis of a randomized clinical trial comparing busulfan (BUS), hydroxyurea (HU) and interferon-α (IFN-α) treatment of chronic myelogenous leukaemia. This trial has been analysed earlier using a Cox model [43–46]. Since the proportional hazards assumption is questionable for the different treatment groups, we re-analysed the data set using the CTM approach and allowed for non-proportional effects of the patient characteristics over time.

2 Conditional transformation models for survival data

In the following, T denotes a positive random variable describing the time from a well-defined starting point to an event of interest, e.g. death or recurrence of a disease. We consider N patients with survival times T_i, $i = 1, \ldots, N$, and a vector of patient characteristics $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$. As we do not assume that all patients experience the event of interest by the end of the study period and as some patients quit the study early, the event times sometimes are not actual event times but rather right-censored. The *observed* right-censored event times $\tilde{T}_i$ are defined by $\tilde{T}_i = \min(T_i, C_i), \quad i = 1, \ldots, N$, where C_i denotes the time under observation or censoring time. Furthermore, the event indicator $\delta_i = I(T_i \leq C_i)$ is 1 for observed event times and 0 for right-censored event times. A common assumption is that the survival time T and the vector of explanatory variables $\boldsymbol{X}$ are independent of the censoring time C.

The conditional survivor function S is defined as the conditional probability of being event free up to some time point t in terms of the conditional distribution function of the survival times given the explanatory variables $\boldsymbol{x}$:

$$S(t|\boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}(T > t|\boldsymbol{X} = \boldsymbol{x}) = 1 - \mathbb{P}(T \leq t|\boldsymbol{X} = \boldsymbol{x}). \tag{1}$$

When using CTMs, we aim at estimating the conditional distribution function of the survival times via

$$\mathbb{P}(T \leq t|\boldsymbol{X} = \boldsymbol{x}) = F(h(t|\boldsymbol{x})), \tag{2}$$

and the conditional survivor function can be calculated by the relationship given in eq. (1). Thereby, the conditional distribution function is modelled in terms of the monotone transformation function $h : \mathbb{R} \to \mathbb{R}$, which depends on the patient characteristics $\boldsymbol{x}$. F denotes an absolute continuous distribution function $F : \mathbb{R} \to [0, 1]$ with corresponding quantile function $Q = F^{-1}$. In CTMs, only the monotone transformation function h is estimated, whereas the link function F is chosen a priori.

To embed the well-known class of linear transformation models [12, 13] into CTMs exemplarily, we reconsider the formulation of the proportional hazards model in terms of a linear transformation model given in Doksum and Gasko [12]. The conditional distribution function of the survival times resulting from the Cox model can be written as

$$\mathbb{P}(T \leq t|\boldsymbol{X} = \boldsymbol{x}) = F(h_T(t) + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}), \tag{3}$$

where F denotes the distribution function of the minimum-extreme value distribution, and the transformation of the survival times $h_T(t)$ equals the logarithm of the cumulative baseline hazard. In linear transformation models, the influence of the explanatory variables is restricted to linear functions and, most importantly, the transformation function h is decomposed into a part depending only on the survival times $h_T(t)$ and a part depending only on the explanatory variables $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}$. This strict decomposition results in the proportional hazards assumption.

In CTMs, the proportional hazards assumption is relaxed by allowing for interactions between the survival times and the explanatory variables in terms of the conditional transformation function $h(t|\boldsymbol{x})$. Furthermore, we assume additivity on the scale of the transformation function and decompose the monotone transformation function h into J partial transformation functions, in which each $h_j : \mathbb{R} \to \mathbb{R}$ is conditional on $\boldsymbol{x}$:

$$\mathbb{P}(T \leq t|\boldsymbol{X} = \boldsymbol{x}) = F(h(t|\boldsymbol{x})) = F\left(\sum_{j=1}^{J} h_j(t|\boldsymbol{x})\right). \tag{4}$$

In analogy to the representation of the Cox model in eq. (3), we choose F to be the minimum-extreme value distribution function. In this way, we operate on the same scale of distribution functions in the CTM and the Cox model, and hence estimations from the two approaches are comparable. The CTM given in eq. (4) can be understood as a generalization of the proportional hazards model to more flexible non-proportional hazard functions, if F is the minimum-extreme value distribution function.

Since all interaction terms between the survival time and the explanatory variables are avoided in the Cox model (eq. (3)), the effects of the explanatory variables are estimated to be constant and are not allowed to vary over time. This assumption is relaxed in the more flexible model class of CTMs. Interaction terms between the survival time and the explanatory variables are established in terms of the partial transformation functions h_j that depend on the survival time *and* on the explanatory variables simultaneously (eq. (4)). Hence, the effects of the explanatory variables are allowed to vary over time, which usually results in non-proportional hazards. We not only estimate one single parameter for each explanatory variable as is done in the Cox model. Instead, separate partial transformation functions are defined for each explanatory variable, whereby a smooth parameter function over time is estimated for each group of a categorical explanatory variable. For continuous explanatory variables, a smooth parameter surface is estimated that depends on the survival time and on the continuous explanatory variable.

In comparison to alternative nonparametric approaches, the estimation of the conditional survivor function is not a black box procedure in CTMs. Although the model assumptions are weak in CTMs, a certain model structure is imposed by introducing additive partial transformation functions. The resulting effects of the explanatory variable over time can be interpreted and can be illustrated graphically. Hence, concerning model complexity, semiparametric CTMs can be placed in between the less flexible semiparametric linear transformation models (e.g. the Cox model) and more flexible nonparametric approaches.

If one is interested in better interpretable versions of CTMs, the model class of conditionally linear transformation models (CLTMs) introduced in Möst et al. [47] can be considered. In CLTMs, the conditional transformation function h is restricted to transformation functions that are linear in the response transformation. Due to this restriction, the explanatory variables are only allowed to influence the conditional mean and the conditional variance of the response transformation, whereas higher moments remain unaffected. The effects of the explanatory variables on the conditional mean and the conditional variance are nonlinear but can be interpreted in CLTMs. Further restrictions of the transformation function are conceivable. For example, if all interaction terms between the survival time and the explanatory variables are omitted and the effects of the explanatory variables have to be linear, the conditional transformation function of the Cox model (eq. (3)) results as a special case. The Cox model can even be further restricted by choosing special forms of the monotone response transformation $h_T(t)$. For example, the specification of $h_T(t) = \log(\lambda) + \nu \cdot \log(t)$ results in the Weibull model.

2.1 Estimating conditional transformation models for survival data

Hothorn et al. [11] explain thoroughly how CTMs are estimated by the minimization of the continuous ranked probability score (CRPS) (see [48]) using a component-wise boosting algorithm. The CRPS was chosen because it constitutes a proper scoring rule for distributional and probabilistic forecasts [11]. When we estimated CTMs for survival data, we also used a component-wise boosting algorithm to minimize an appropriate integrated loss function. First, we formulated the integrated loss function for uncensored observations, and then we extended the loss function to right-censored observations.

2.1.1 Integrated loss function for uncensored observations

In an uncensored survival data set-up, we observed the survival or event times $T_i,\ i = 1, \ldots, N$, for N patients under consideration. Furthermore, we considered a grid of time points $\{t_\iota : \iota = 1, \ldots, n\}$ ranging from the study's starting point $t_1 = 0$ to the study's end point t_n. Typical choices for the grid points $\{t_\iota : \iota = 1, \ldots, n\}$ are equally spaced grid points or a grid composed of all distinct survival and event times. Hence, we were able to observe the binary survival status $I(T_i \leq t_\iota)$ for each patient at each grid point; the status is 1 if the patient experienced the event by t_ι and is otherwise 0.

We aimed at estimating the conditional distribution function of the event times $\mathbb{P}(T \leq t_\iota | \boldsymbol{X} = \boldsymbol{x}) = F(h(t_\iota|\boldsymbol{x}))$ (see eq. (2)) in terms of the conditional transformation function h, where t_ι denotes some arbitrary time point in the study period. This estimation problem can be reformulated as estimating the probability $F(h(t_\iota|\boldsymbol{x}))$ for the binary event $T \leq t_\iota$ and is solved by minimizing an appropriate loss function. We chose the logarithmic score (or negative binomial log-likelihood) for measuring the loss between the binary event status $T_i \leq t_\iota$ and the corresponding probability $F(h(t_\iota|\boldsymbol{x}_i))$ for N patients at a specific time point t_ι:

$$\begin{aligned} \mathrm{LS}(t_\iota) = -\frac{1}{N} \sum_{i=1}^{N} \{ & I(T_i \leq t_\iota) \log(F(h(t_\iota|x_i))) \\ & + I(T_i > t_\iota) \log(1 - F(h(t_\iota|x_i)))\}. \end{aligned} \tag{5}$$

Alternatively, the Brier score or the absolute error loss can be chosen as an appropriate loss function [11, 48, 49].

Based on the logarithmic score for one specific time point t_ι (see eq. (5)), we defined the integrated logarithmic score over all time points, which allows estimation of the whole conditional distribution function $\mathbb{P}(T \leq t|\boldsymbol{X} = \boldsymbol{x})$ in one step:

$$\begin{aligned} \mathrm{ILS} = -\frac{1}{N} \sum_{i=1}^{N} \int_0^{t_n} \{ & I(T_i \leq t) \log(F(h(t|\boldsymbol{x}_i))) \\ & + I(T_i > t) \log(1 - F(h(t|x_i)))\}\, dW(t), \end{aligned} \tag{6}$$

where $W(t)$ denotes a weight function for the time points. By choosing the same weight $\frac{1}{n}$ for all time points $t_\iota,\ \iota = 1, \ldots, n$, we get the empirical version of eq. (6):

$$\begin{aligned} \widehat{\mathrm{ILS}} = -\frac{1}{N \cdot n} \sum_{i=1}^{N} \sum_{\iota=1}^{n} \{ & I(T_i \leq t_\iota) \log(F(h(t_\iota|\boldsymbol{x}_i))) \\ & + I(T_i > t_\iota) \log(1 - F(h(t_\iota|\boldsymbol{x}_i)))\}, \end{aligned} \tag{7}$$

which is used as the empirical loss function in the boosting algorithm. Of course, other weight functions $W(t)$ for the time points are conceivable.

When the conditional distribution function is estimated, the ultimate goal is to estimate the conditional transformation function h such that the empirical risk in eq. (7) is minimized. The minimization of the empirical risk is equivalent to the minimization of the loss between the true survival status at time point t_ι,

$I(T_i \leq t_\iota)$, and the corresponding estimated survival probability $F(\hat{h}(t_\iota|\boldsymbol{x}_i))$ for all time points and all patients. In other words, the survivor function for a specific patient $\hat{S}(t_\iota|\boldsymbol{x}_i) = 1 - F(\hat{h}(t_\iota|\boldsymbol{x}_i)), \quad \iota = 1, \ldots, n$, is estimated such that the survival probabilities fit the patient's true survival status best.

2.1.2 Integrated loss function for right-censored observations

In survival analysis, we often face right-censored survival times. We do not observe the true survival time T_i for the right-censored patients, and only the observed survival times $\tilde{T}_i = \min(T_i, C_i), \quad i = 1, \ldots, N$, are available. One way to account for right-censored observations in model estimation is the inverse probability of censoring weighting (IPCW) approach suggested by van der Laan and Robins [50] and used often in the past (e.g. see [51, 52]). For example, Robins and Finkelstein [53] present an IPCW version of the Kaplan–Meier estimator and the log-rank test to account for noncompliance and dependent censoring. Van der Laan and Robins [50] give an IPCW example for right-censored data with time-independent explanatory variables and censoring at random and suggest that the full data loss function (i.e. the integrated logarithmic score in our case) be weighted by the IPCWs:

$$\omega_{i\iota} = \frac{\Delta(t_\iota)}{\hat{K}(\min(T_i, t_\iota))}, \tag{8}$$

where $\Delta(t_\iota) = I(C_i > \min(T_i, t_\iota))$. $\hat{K}$ denotes the marginal Kaplan–Meier estimator of the censoring distribution, $\hat{K}(t) = \hat{\mathbb{P}}(\mathrm{T} > \mathrm{t})$, based on $(\tilde{T}_i, 1 - \delta_i), \ i = 1, \ldots, N$, hence on the observed survival times and the reverse censoring indicator, which is 1 for right-censored observations and 0 otherwise. Furthermore, the censoring time C_i is set to ∞ for uncensored observations.

To calculate the IPCWs for the integrated logarithmic score in eq. (7) based on eq. (8), we have to distinguish four different situations:

1. Uncensored observations ($\delta_i = 1$) that experience the event up to t_ι ($\tilde{T}_i \leq t_\iota$):

$$\omega_{i\iota} = \frac{I(\tilde{T}_i \leq t_\iota, \delta_i = 1) \cdot \overbrace{I(C_i > T_i)}^{=\Delta(t_\iota)=1}}{\hat{K}(T_i)} = \frac{1}{\hat{K}(T_i)} = \frac{1}{\hat{K}(\tilde{T}_i)}.$$

2. Uncensored observations ($\delta_i = 1$) that do not experience the event up to t_ι ($\tilde{T}_i > t_\iota$):

$$\omega_{i\iota} = \frac{I(\tilde{T}_i > t_\iota, \delta_i = 1) \cdot \overbrace{I(C_i > t_\iota)}^{=\Delta(t_\iota)=1}}{\hat{K}(t_\iota)} = \frac{1}{\hat{K}(t_\iota)}.$$

3. Right-censored observations ($\delta_i = 0$) that experience the censoring up to t_ι ($\tilde{T}_i \leq t_\iota$):

$$\omega_{i\iota} = \frac{I(\tilde{T}_i \leq t_\iota, \delta_i = 0) \cdot \overbrace{I(C_i > T_i)}^{=\Delta(t_\iota)=0}}{\hat{K}(\mathrm{NA})} = 0.$$

4. Right-censored observations ($\delta_i = 0$) that do not experience the censoring up to t_ι ($\tilde{T}_i > t_\iota$):

$$\omega_{i\iota} = \frac{I(\tilde{T}_i > t_\iota, \delta_i = 0) \cdot \overbrace{I(C_i > t_\iota)}^{=\Delta(t_\iota)=1}}{\hat{K}(t_\iota)} = \frac{1}{\hat{K}(t_\iota)}.$$

The resulting weighting scheme corresponds exactly to the weighting scheme given in Graf et al. [54], which results in a consistent estimator (see [51]). In short, the observations are weighted by the inverse probability of not being censored up to the event time (situation 1) or up to the specific time point under consideration (situations 2 and 4). The current survival status is unknown in situation 3; consequently, these observations

get zero weights. Thus, censored observations contribute to the model estimation process up to their censoring time point and those observations that have already been censored are accounted for in the IPCWs.

We extended the empirical logarithmic score for uncensored observations given in eq. (7) to right-censored observations by including the weighting scheme presented above. Hence, the empirical version of the integrated censored logarithmic score results in

$$\widehat{\mathrm{ILS}^C} = -\frac{1}{N \cdot n} \sum_{i=1}^{N} \sum_{\iota=1}^{n} \Bigg\{ I(\tilde{T}_i \leq t_\iota, \delta_i = 1) \log(F(h(t_\iota|\boldsymbol{x}_i))) \cdot \frac{1}{\hat{K}(\tilde{T}_i)} + I(\tilde{T}_i > t_\iota) \log(1 - F(h(t_\iota|\boldsymbol{x}_i))) \cdot \frac{1}{\hat{K}(t_\iota)} \Bigg\}, \tag{9}$$

which is used as empirical risk function in the boosting algorithm.

2.2 Boosting conditional transformation models for survival data

In CTMs, the conditional distribution function of uncensored responses is estimated using component-wise boosting with penalization (for a detailed description, see [11]). This algorithm has to be slightly modified for the estimation of right-censored survival data. Thereby, the empirical risk given in eq. (9) is minimized with respect to the transformation function h. Furthermore, the parametrization of the partial transformation functions $h_j,\ j = 1, \ldots, J$, (eq. (4)) has to be slightly adapted for survival data. In component-wise boosting algorithms, regularization is achieved by the application of penalized base learners. The overall model complexity is regulated by the number of boosting iterations M. For a thorough introduction to component-wise boosting with smooth base learners, see Bühlmann and Hothorn [55] and Schmid and Hothorn [56].

2.2.1 Parametrization of the partial transformation functions

Considering the parametrization of the partial transformation functions in Hothorn et al. [11], we defined for the jth partial transformation function:

$$h_j(t_\iota|\boldsymbol{x}) = \left(\boldsymbol{b}_j(\boldsymbol{x})^\top \otimes \boldsymbol{b}_T(t_\iota)^\top\right)\boldsymbol{\gamma}_j, \quad j = 1, \ldots, J, \tag{10}$$

where $\boldsymbol{b}_T : \mathbb{R} \to \mathbb{R}^{K_T}$ denotes the basis along the grid of time points $t_\iota,\ \iota = 1, \ldots, n$, and $\boldsymbol{b}_j : \chi \to \mathbb{R}^{K_j}$ is a basis for (a subset of) the explanatory variables $\boldsymbol{x}$. Both sets of basis functions are connected via a Kronecker product, whereby an interaction surface between the survival times and the explanatory variables is established. The vector $\boldsymbol{\gamma}_j \in \mathbb{R}^{K_j K_T}$ contains the basis coefficients for the established interaction surface. The basis $\boldsymbol{b}_T$ defines the functional form of the transformation of the survival times, and the functional form of $\boldsymbol{b}_j$ defines how the survival time transformation is influenced by the explanatory variables [11]. Hence, one usually chooses B-spline basis functions for $\boldsymbol{b}_T$, and depending on the desired flexibility or the measurement level of the explanatory variables, one chooses linear basis functions or B-spline basis functions for $\boldsymbol{b}_j$. In more detail, linear basis functions are chosen for $\boldsymbol{b}_j$ if $\boldsymbol{x}$ is univariate and categorical or if $\boldsymbol{x}$ is univariate and continuous, and a linear influence is assumed. B-spline basis functions are chosen for $\boldsymbol{b}_j$ if $\boldsymbol{x}$ is univariate and continuous, and the influence might be more flexible. Additionally, $\boldsymbol{b}_j$ might depend on more than one explanatory variable, and appropriate multivariate basis functions have to be considered. The partial transformation functions h_j are typically expected to be smooth in the first argument t and in the conditioning variable $\boldsymbol{x}$ because continuous distribution functions have to be smooth in the response variable. Moreover, we expect similar distribution functions for similar values of the explanatory variables. Therefore, appropriate penalty matrices $P_T \in \mathbb{R}^{K_T \times K_T}$ and $P_j \in \mathbb{R}^{K_j \times K_j}$ are imposed

on the basis functions defined in eq. (10). The penalty matrix for the Kronecker product of the basis functions is defined via $P_{Tj} = (\lambda_T P_j \otimes \mathbf{1}_{K_T} + \lambda_j \mathbf{1}_{K_j} \otimes P_T)$, where $\lambda_T \geq 0$ and $\lambda_j \geq 0$ denote smoothing parameters and $\mathbf{1}$ denotes the identity matrix.

As an example, we give the partial transformation function for the explanatory variable sex influencing the survival time transformation:

$$h_{\text{sex}}(t_\iota|\text{sex}) = \left(\boldsymbol{b}_{\text{sex}}^{\text{lin}}(\text{sex})^\top \otimes \boldsymbol{b}_T(t_\iota)^\top\right)\boldsymbol{\gamma}_{\text{sex}}.$$

Since the explanatory variable sex is binary, we chose linear basis functions for $\boldsymbol{b}_{\text{sex}}^{\text{lin}}(\text{sex})$, and furthermore, we chose B-spline basis functions for $\boldsymbol{b}_T$. No penalty term P_{sex} is specified for the linear basis $\boldsymbol{b}_{\text{sex}}^{\text{lin}}$ and a smoothness penalty term based on second-order differences P_T is defined for the B-spline basis $\boldsymbol{b}_T$. The resulting interaction surface for the explanatory variable sex and the survival time can also be understood as the separate estimation of a smooth survival time transformation for males and females. Hence, the difference in the survival probabilities of males and females is allowed to vary flexibly over time and is therefore able to display non-proportional hazards for the explanatory variable sex. For further details on parametrization and penalty specification, see Hothorn et al. [11].

2.2.2 Component-wise boosting algorithm for conditional transformation models for survival data

The component-wise boosting algorithm for right-censored survival data is only a slight modification of the algorithm presented in Hothorn et al. [11]:

(Init) Initialize the parameters $\boldsymbol{\gamma}_j^{[0]} \equiv 0$ for $j = 1, \ldots, J$, the step-size $\nu \in (0, 1)$ and the smoothing parameters $\lambda_j, j = 1, \ldots, J$. Define the grid $t_1 < \tilde{T}_{(1)} < \ldots < \tilde{T}_{(N)} \leq t_n$. Calculate the IPCWs $\omega_{i\iota}$ for each grid point ι and each observation i.
Set $m = 0$.

(Gradient) Compute the negative gradient of the censored log score:

$$\begin{aligned} U_{i\iota} &:= -\frac{\partial}{\partial h}\rho((\tilde{T}_i \leq t_\iota, \boldsymbol{x}_i), h)\Big|_{h=\hat{h}_{i\iota}^{[m]}} \\ &:= \left\{ I(\tilde{T}_i \leq t_\iota, \delta_i = 1)\, \frac{F^{|}(h(t_\iota|\boldsymbol{x}_i))}{F(h(t_\iota|\boldsymbol{x}_i))} \cdot \frac{1}{\hat{K}(\tilde{T}_i)} \right. \\ &\quad \left. - I(\tilde{T}_i > t_\iota)\, \frac{F^{|}(h(t_\iota|\boldsymbol{x}_i))}{1 - F(h(t_\iota|\boldsymbol{x}_i))} \cdot \frac{1}{\hat{K}(t_\iota)} \right\}\Bigg|_{h=\hat{h}_{i\iota}^{[m]}}, \end{aligned}$$

where $F^{|}(\cdot)$ denotes the density of the link function F, $\hat{K}(\cdot)$ denotes the marginal Kaplan–Meier estimator of the censoring distribution and

$$\hat{h}_{i\iota}^{[m]} = \sum_{j=1}^{J} \hat{h}_j^{[m]}(t_\iota|\boldsymbol{x}_i) = \sum_{j=1}^{J} (\boldsymbol{b}_j(\boldsymbol{x}_i)^\top \otimes \boldsymbol{b}_T(t_\iota)^\top)\boldsymbol{\gamma}_j^{[m]}.$$

Fit the base-learners for $j = 1, \ldots, J$:

$$\hat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta} \in \mathbb{R}^{K_j \cdot K_T}}{\arg\min} \sum_{i=1}^{N} \sum_{\iota=1}^{n} \omega_{i\iota} \left\{ U_{i\iota} - (\boldsymbol{b}_j(\boldsymbol{x}_i)^\top \otimes \boldsymbol{b}_T(t_\iota)^\top)\boldsymbol{\beta} \right\}^2 + \boldsymbol{\beta}^{\mathrm{T}} P_{Tj} \boldsymbol{\beta}$$

with penalty matrix P_{Tj}.
Select the base-learner

$$j^* = \underset{j=1,\ldots,J}{\arg\min} \sum_{i=1}^{N} \sum_{\iota=1}^{n} \omega_{i\iota} \left\{ U_{i\iota} - (\boldsymbol{b}_j(\boldsymbol{x}_i)^\top \otimes \boldsymbol{b}_T(t_\iota)^\top)\hat{\boldsymbol{\beta}}_j \right\}^2.$$

(Update) the parameters $\boldsymbol{\gamma}_{j^*}^{[m+1]} = \boldsymbol{\gamma}_{j^*}^{[m]} + \nu \cdot \hat{\boldsymbol{\beta}}_{j^*}$ and keep all other parameters fixed, *i.e.* $\boldsymbol{\gamma}_j^{[m+1]} = \boldsymbol{\gamma}_j^{[m]},\ j \neq j^*$. Iterate (Gradient) and (Update).
(Stop) if $m = M$. Output the model

$$\hat{\mathbb{P}}(T \leq t | \boldsymbol{X} = \boldsymbol{x}) = F(\hat{h}^{[M]}(t|\boldsymbol{x})) = F\left(\sum_{j=1}^{J} \hat{h}_j^{[M]}(t|\boldsymbol{x})\right)$$
$$= F\left(\sum_{j=1}^{J} \left(\boldsymbol{b}_j(\boldsymbol{x})^\top \otimes \boldsymbol{b}_{\mathrm{T}}(t)^\top\right)\boldsymbol{\gamma}_j^{[M]}\right)$$

as a function of arbitrary $t \in \mathbb{R}^+$ and arbitrary explanatory variables $\boldsymbol{x}$.

3 Simulation

3.1 Simulation study set-up

In the following simulations, we investigated the performance of CTMs in comparison to alternative semiparametric (ordinary Cox model and stratified Cox model) or nonparametric (Kaplan–Meier estimator; conditional random forests) modelling strategies in four different simulation settings with Weibull distributed survival times. We considered different scenarios of explanatory variables and proportional as well as non-proportional hazard settings. Since the handling of censored observations is an important issue, we considered different amounts of right-censored survival times. The censoring times were drawn independently from uniform distributions such that 5%, 10%, 25% and 50% right-censored observations resulted in each simulation setting.

The true hazard function and the corresponding true survivor function for Weibull distributed survival times are

$$\lambda(t) = \frac{c}{b^c} t^{c-1} \quad \text{and} \quad S(t) = \exp(-b^{-c} t^c), \tag{11}$$

where b and c denote the scale and shape parameter of the Weibull distribution, respectively. The choice of parameters b and c determines whether proportional hazards or non-proportional hazards result. The proportional hazards assumption is fulfilled if the explanatory variables influence only the scale parameter b and the shape parameter c is fixed. If the explanatory variables additionally influence the shape parameter c, the proportional hazards assumption is violated, which, e.g. results in crossing survivor functions.

3.1.1 Simulation 1

In the first simulation setting, we considered the simple data setting of two treatment groups G_1 and G_2, which differed with respect to their survival probabilities. The survival times were Weibull distributed with $b_1 = 1$ and $c_1 = 3$ for treatment group G_1 and $b_2 = 1.5$ and $c_2 = 3$ for treatment group G_2. Moreover, we included a non-informative continuous covariate x. Since the shape parameters were identical, the corresponding survivor functions followed the proportional hazards assumption (Figure 1). This could also be recognized by rewriting the conditional Weibull distribution in terms of the Cox linear transformation model (eq. (3)). The conditional Weibull distribution resulted from eq. (11) by inserting the scale parameter $b = \beta_G + \beta_x \cdot x$, where $\beta_G = 1$ for G_1 and $\beta_G = 1.5$ for G_2, and the shape parameter $c = \gamma_G + \gamma_x \cdot x$ with $\gamma_G = 3$ for both treatment groups. Since x was non-influential, $\beta_x = \gamma_x = 0$:

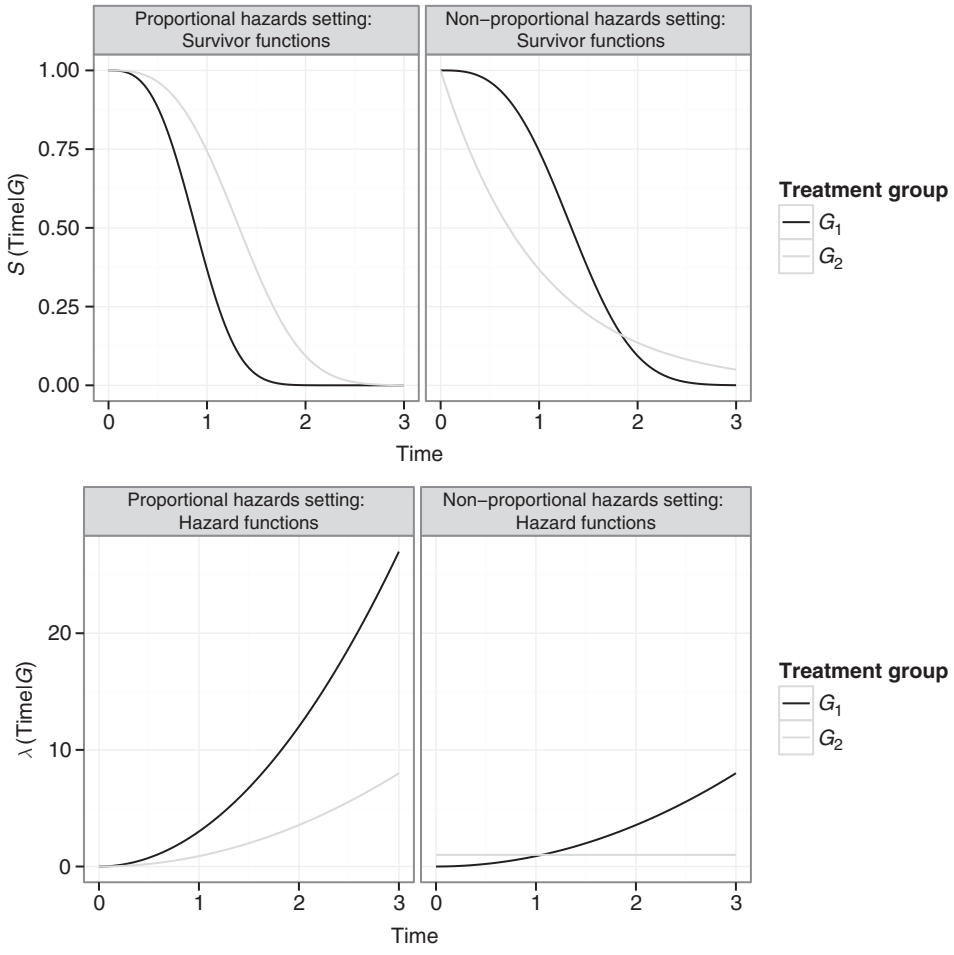

Figure 1: Simulation: True survivor and hazard functions for treatment groups G_1 and G_2 based on Weibull distributed survival times. Proportional hazards setting (simulation 1): $b_1 = 1$, $c_1 = 3$ (for G_1) and $b_2 = 1.5$, $c_2 = 3$ (for G_2); non-proportional hazards setting (simulation 2): $b_1 = 1.5$, $c_1 = 3$ and $b_2 = 1$, $c_2 = 1$.

$$\begin{aligned} 1 - S(t|G, x) &= 1 - \exp(-(\beta_G + \beta_x \cdot x)^{-\gamma_G - \gamma_x \cdot x} \cdot t^{\gamma_G + \gamma_x \cdot x}) \\ &\underset{\gamma_x = \beta_x = 0, \gamma_G = 3}{=} 1 - \exp(-\exp(-3 \cdot \log(\beta_G) + 3 \cdot \log(t))) \\ &= F(h_T(t) + \tilde{\beta}_G), \end{aligned}$$

where F denotes the minimum-extreme value distribution, $h_T(t) = 3 \cdot \log(t)$ and $\tilde{\beta}_G = -3 \cdot \log(\beta_G)$. This setting could be perfectly fitted using a Cox model as there was no interaction term between the treatment group G and the survival time t (i.e. the proportional hazards assumption was fulfilled), and G had a linear influence. We sampled $N_G = 200$ survival times T from the respective Weibull distribution for each treatment group and identical $N_G = 200$ independent x-values were chosen on an equidistant grid on $[0, 1]$ for the treatment groups.

3.1.2 Simulation 2

In analogy to simulation 1, the survival probabilities differed for treatment groups G_1 and G_2, and the continuous explanatory variable x was non-informative. The parameters of the Weibull distributed survival times were $b_1 = 1.5$ and $c_1 = 3$ for treatment group $G1$ and $b_2 = 1$ and $c_2 = 1$ for treatment group G_2. Since the scale and the shape parameters were treatment specific, the proportional hazards assumption was violated (Figure 1). Again, this could be clarified by writing the conditional Weibull distribution in terms of eq. (3):

$$\begin{aligned} 1 - S(t|G,x) &= 1 - \exp(-(\beta_G + \beta_x \cdot x)^{-\gamma_G - \gamma_x \cdot x} \cdot t^{\gamma_G + \gamma_x \cdot x}) \\ &\underset{\beta_x = \gamma_x = 0}{=} 1 - \exp(-\exp(-\gamma_G \cdot \log(\beta_G) + \gamma_G \cdot \log(t))), \end{aligned}$$

where $\beta_G = 1.5$ for G_1 and $\beta_G = 1$ for G_2, and $\gamma_G = 3$ for G_1 and $\gamma_G = 1$ for G_2. Since there is an interaction term between G and t, the proportional hazards assumption was violated. We sampled $N_G = 200$ survival times for each treatment group from the respective Weibull distributions. The independent and identical $N_G = 200$ x-values were chosen on an equidistant grid on $[0, 1]$ for the treatment groups.

3.1.3 Simulation 3

The survival times differed with respect to the treatment group G *and* with respect to the continuous explanatory variable x in simulation setting 3. The survival times were Weibull distributed with scale parameters $b_1 = \exp(1/4 + x)$ for treatment group G_1 and $b_2 = \exp(1 + x)$ for treatment group G_2. The shape parameters $c_1 = c_2 = 3$ were identical, which resulted in the proportional hazards assumption. Again, the connection to the Cox model could be established in terms of eq. 3. We inserted $b = \exp(\beta_G + \beta_x \cdot x)$ for the scale parameter, where $\beta_G = 0.25$ for G_1 and $\beta_G = 1$ for G_2, $\beta_x = 1$, and $c = 3$:

$$\begin{aligned} 1 - S(t|G,x) &= 1 - \exp(-\exp(-3 \cdot (\beta_G + x) + 3 \cdot \log(t))) \\ &= F(h_T(t) + \tilde{\boldsymbol{x}}^{\mathrm{T}} \tilde{\boldsymbol{\beta}}), \end{aligned}$$

where F denotes the minimum-extreme value distribution, $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_G \; \tilde{\beta}_x)^{\mathrm{T}}$ and $\tilde{\boldsymbol{x}} = (G \; x)$. More precisely, the parameters of the linear transformation model were $\tilde{\beta}_G = -0.75$ for G_1 and $\tilde{\beta}_G = -3$ for G_2, $\tilde{\beta}_x = -3$ and $h_T(t) = 3 \cdot \log(t)$. Hence, the simulation setting could be perfectly analysed using a Cox model as there were no interactions between the explanatory variables and the survival time, and G and x had a linear influence. First, we chose $N_G = 300$ x-values by defining an equidistant grid on $[0, 1]$ for the treatment groups. Afterwards, we sampled 300 survival times from the Weibull distributions with parameters b_1 and c for treatment group G_1 and 300 survival times from the Weibull distributions with parameters b_2 and c for treatment group G_2.

3.1.4 Simulation 4

In analogy to simulation 3, the survival probabilities were influenced by G and x. But this time, we chose a non-proportional hazards setting by keeping the scale parameter $b = \exp(0.5)$ fixed and letting the shape parameter to depend on the explanatory variables: $c_1 = 2 + x^2$ for treatment group G_1 and $c_2 = 2.5 + x^2$ for treatment group G_2. More general, the shape parameter c is $c = \gamma_G + \gamma_x(x)$ with $\gamma_G = 2$ for G_1 and $\gamma_G = 2.5$ for G_2 and a non-linear function in x, $\gamma_x(x) = x^2$. Hence, the shape parameters differed only slightly for the treatment groups and were mainly influenced non-linearly by x. Again, the conditional Weibull distribution of the survival times could be displayed as a CTM:

$$1 - S(t|G,x) = 1 - \exp\left(-\exp\left(-\frac{1}{2} \cdot \gamma_G - \frac{1}{2}x^2 + \gamma_G \cdot \log(t) + x^2 \cdot \log(t)\right)\right).$$

As there were interactions between the explanatory variables and the survival time, the proportional hazards assumption was violated. We first chose $N_G = 300$ x-values by defining an equidistant grid on $[0, 2]$. In this simulation setting, the x-values varied on $[0, 2]$ instead of $[0, 1]$, which resulted in a wider range of shape values c. Afterwards, 300 survival times were sampled from the Weibull distributions with parameters b and c_1 for treatment group G_1 and 300 survival times were sampled from the Weibull distributions with parameters b and c_2 for treatment group G_2.

3.2 Model estimation

We estimated the conditional survival curves for the treatment groups G_1 and G_2 and the continuous covariate x, $S(t|G,x)$, using a CTM, an ordinary Cox model and conditional random forests in all four simulations. In the Cox model, the hazard function was modelled via $\lambda(t|G,x) = \lambda_0(t)\exp(\beta_G \cdot G + \beta_x \cdot x)$, where $\lambda_0(t)$ denotes the baseline hazard. In the CTM, a partial transformation function for each explanatory variable was defined, $h(t|G,x) = h_G(t|G) + h_x(t|x)$, in which the influence of the explanatory variables was allowed to vary over time.

Separate Kaplan–Meier estimators can only be obtained for categorical explanatory variables. As x was non-influential in simulations 1 and 2, treatment-specific Kaplan–Meier estimates were additionally provided.

A non-proportional hazards setting was considered in simulations 2 and 4. Therefore, we additionally estimated a stratified Cox model with treatment-specific baseline hazard functions: $\lambda(t|G,x) = \lambda_G(t)\cdot \exp(\beta_x \cdot x)$.

The flexibility of CTMs can be restricted to the flexibility of a Cox model by considering a CLTM [47]. We avoided all interactions between the explanatory variables and the survival time and assumed linear influences for G and x in the corresponding conditional transformation function: $h(t|G,x) = h_G(1|G) + h_x(1|x) + h_T(t|1) = \beta_G \cdot G + \beta_x \cdot x + h_T(t)$. Hence, the CLTM and the Cox model could be considered as semiparametric alternatives, in which both models assumed proportional hazards.

3.2.1 Simulation 1

The conditional survivor functions were estimated using a CTM, a CLTM, an ordinary Cox model, the Kaplan–Meier estimator and conditional random forests. Thereby, the treatment-specific Kaplan–Meier estimator could be understood as a nonparametric alternative to conditional random forests, in which the Kaplan–Meier estimator was expected to perform better as the non-informative explanatory variable x was ignored.

3.2.2 Simulation 2

In this non-proportional hazards setting, the conditional survivor functions were additionally estimated using a stratified Cox model. As x was non-informative, treatment-specific Kaplan–Meier estimators were obtained as both a nonparametric and a predominant alternative to conditional random forests.

3.2.3 Simulation 3

We estimated the conditional survivor functions $S(t|G,x)$ using a CTM, a CLTM, a Cox model and conditional random forests. In general, the identification of the linear influence of x is difficult for conditional random forests, as the linear function has to be approximated by a step function.

3.2.4 Simulation 4

The conditional survivor functions were estimated using a CTM, a CLTM, an ordinary Cox model and a stratified Cox model and conditional random forests. Similar to simulation 3, conditional random forests had difficulties in identifying the non-linear influence of x, which had to be approximated by a step function.

3.3 Model evaluation

We aimed at evaluating the goodness of the CTM, the CLTM, the Cox model (both ordinary and stratified), the Kaplan–Meier estimator and conditional random forests for estimating the survivor functions of treatment groups G_1 and G_2 in all four simulation settings. Therefore, we used the out-of-sample uncensored log score (eq. (7)) and the mean absolute deviation (MAD) between the true and the estimated survivor functions as quality criteria.

For the evaluation, we drew 1,000 new observations for each treatment group. In simulations 1 and 2, we simply drew 1,000 new observations from the Weibull distributions with parameters b_1 and c_1, and b_2 and c_2, respectively. In simulations 3 and 4, we defined 1,000 x-values by determining an equidistant grid on $[0,1]$ and $[0,2]$, respectively. Afterwards, we drew a new Weibull distributed survival time depending on the shape and scale parameter induced by each x-value.

Based on these new observations, we calculated separate uncensored log scores for the two treatment groups. As an example, we describe the calculation of the uncensored log score for treatment group G_1: We compared the true survivor status $I(T_l \leq t_\iota)$ for each new survival time and corresponding x-value, $(T_l, x_l),\ l = 1, \ldots, 1,000$, for treatment group G_1 along a grid of time points t_ι with the corresponding estimated survival probabilities $\pi(t_\iota | G_1, x_l)$. Thereby, the estimated conditional survival probabilities $\pi(t_\iota | G_1, x_l)$ resulted from the CTM, the CLTM, the Cox model (ordinary or stratified), or conditional random forests. The survival probabilities $\pi(t_\iota | G_1)$ were only treatment specific for the Kaplan–Meier estimator in simulations 1 and 2. The grid of time points t_ι consisted of all new survival times $T_l,\ l = 1, \ldots, 1,000$. The uncensored log score for treatment group G_2 was calculated analogously.

In addition, we calculated the MAD of the estimated survival curves and the true Weibull distribution functions for each treatment group separately. Thereby, we also considered the grid of 1,000 x-values $x_l,\ l = 1, \ldots, 1,000$, and the grid consisting of the 1,000 new survival times for each treatment group, $t_\iota,\ \iota = 1, \ldots, 1,000$:

$$\mathrm{MAD}(G_k) = \frac{1}{1,000 \cdot 1,000} \sum_{l=1}^{1,000} \sum_{\iota=1}^{1,000} | p(t_\iota | G_k, x_l) - \pi(t_\iota | G_k, x_l) |, \tag{12}$$

where p denotes the true survival probabilities and π denotes the estimated survival probabilities. Furthermore, $k \in \{1, 2\}$ denotes the index for the two treatment groups and $l = 1, \ldots, 1,000$ is the index for the new observations for each treatment group. In the simulation settings 1 and 2, the true survival probabilities $p(t_\iota | G_k, x_l)$ reduced to $p(t_\iota | G_k)$ as x was non-informative. For reasons of interpretability, the MAD values and the uncensored log scores were multiplied by 100.

This procedure was repeated for $B = 100$ simulated data sets. We calculated mean values of the resulting 100 MADs or uncensored log scores for the different treatment groups and the different estimation techniques.

3.3.1 Simulation 1

All estimation approaches yielded similar results. The calculated mean MADs (Table 1; Figure 2) were small for all model approaches and indicated that the estimated survivor functions were in good accordance with the true Weibull survivor functions. Only for 50% censored observations did the MADs grow larger throughout. The Cox model, the CLTM and the Kaplan–Meier estimator performed slightly better because the Cox model and the CLTM profited from the proportional hazards assumption and the Kaplan–Meier estimator ignored the non-informative explanatory variable x. Nevertheless, the uncensored log score was the more interesting quality criterion, as it evaluates how well the estimation techniques are able to predict the survivor status of *new* observations. Again, all four estimation approaches yielded similar results (Table 2; Figure 3). All uncensored log scores grew larger with an increasing amount of censored observations.

Table 1: Simulation 1: Mean absolute deviations between true and estimated survival curves for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
	CTM	2.87	2.96	3.45	6.18
	CLTM	2.43	2.51	3.02	5.97
G_1	Cox	2.41	2.54	3.23	6.64
	Kaplan–Meier	2.41	2.51	3.21	6.63
	Cforest	2.59	2.67	3.34	6.74
	CTM	2.71	2.80	3.58	6.98
	CLTM	2.37	2.48	3.30	6.84
G_2	Cox	2.30	2.42	3.29	6.88
	Kaplan–Meier	2.28	2.38	3.24	6.71
	Cforest	2.50	2.57	3.33	6.64

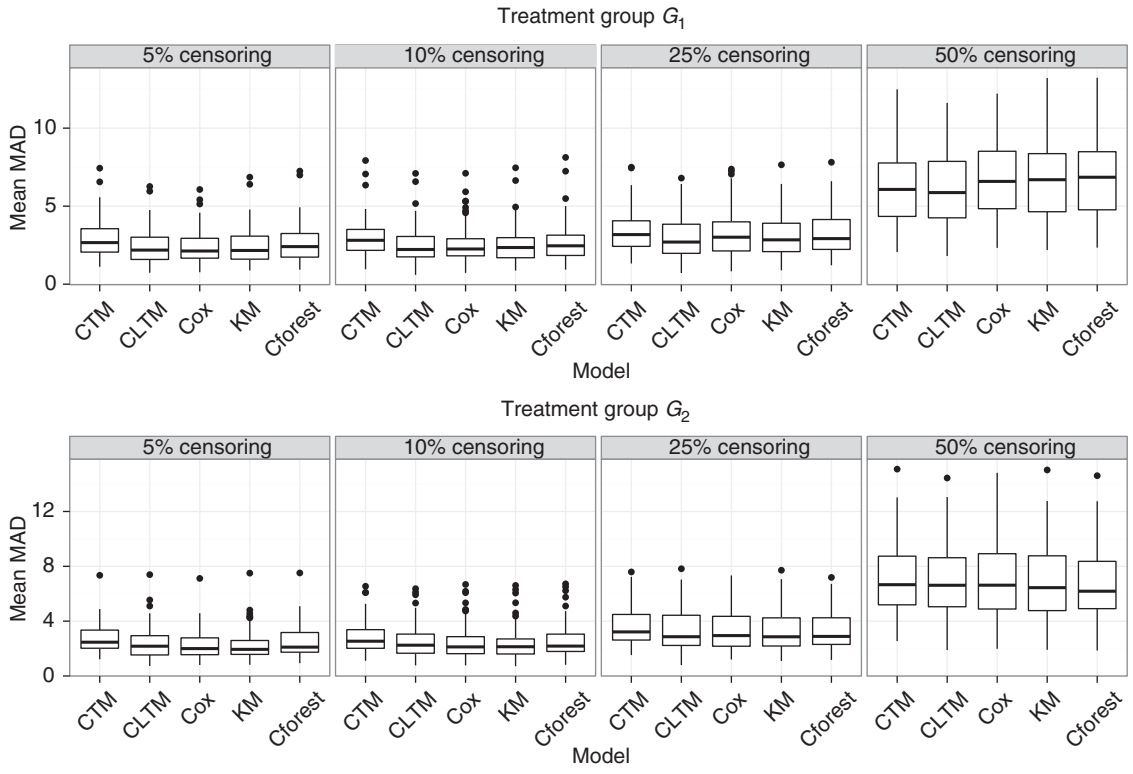

Figure 2: Simulation 1: Boxplot of the treatment-specific mean MAD values based on $B = 100$ simulations for the conditional transformation model (CTM), the conditionally linear transformation model (CLTM), the Cox model (Cox), the Kaplan–Meier estimator (KM) and conditional random forests (Cforest): 5%, 10%, 25% and 50% of right-censored observations were observed.

3.3.2 Simulation 2

The MADs of the CTM, the stratified Cox model, the Kaplan–Meier estimator and conditional random forests were similar throughout, whereas the ordinary Cox model and the CLTM clearly yielded higher MADs (Table 3; Figure 4). The only exception was the MADs for 50% censored observations, where all models had higher MADs. Moreover, the MADs for conditional random forests were most variable and the Kaplan–Meier

Table 2: Simulation 1: Out-of-sample uncensored log score based on 1,000 new observations for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
	CTM	50.52	50.58	50.82	52.27
	CLTM	50.39	50.45	50.68	52.02
G_1	Cox	50.40	50.47	50.78	52.41
	Kaplan–Meier	50.42	50.50	50.89	52.66
	Cforest	50.46	50.54	50.93	52.75
	CTM	50.54	50.63	50.97	52.79
	CLTM	50.47	50.58	50.89	52.70
G_2	Cox	50.39	50.48	50.79	52.52
	Kaplan–Meier	50.45	50.55	50.92	52.75
	Cforest	50.49	50.60	50.95	52.74

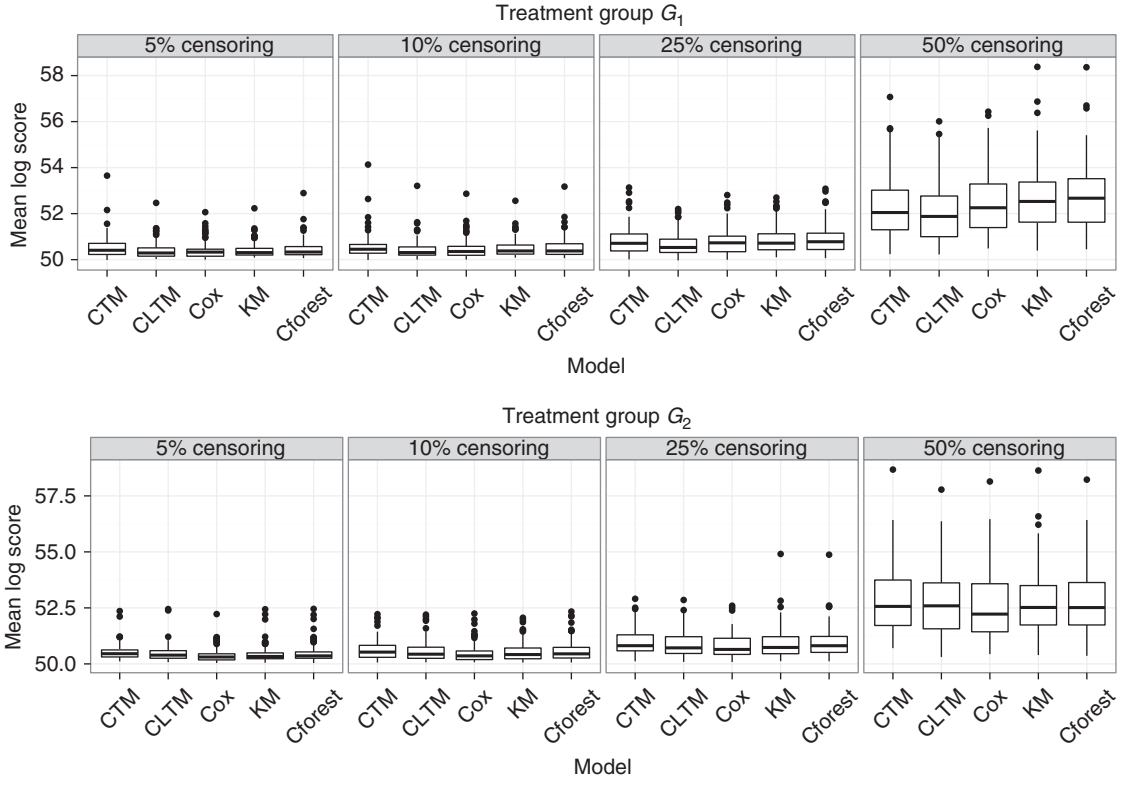

Figure 3: Simulation 1: Boxplot of the out-of-sample mean uncensored log scores based on 1,000 new observations for each treatment group and $B = 100$ simulations for the conditional transformation model (CTM), the conditionally linear transformation model (CLTM), the Cox model (Cox), the Kaplan–Meier estimator (KM) and conditional random forests (Cforest): 5%, 10%, 25% and 50% of right-censored observations were observed.

estimator performed better, as it profited from ignoring x. The uncensored log scores gave similar results (Table 4; Figure 5). Again, the log scores for the CTM, the stratified Cox model, the Kaplan–Meier estimator and conditional random forests were similar, whereas the ordinary Cox model and the CLTM clearly yielded higher values. One exception was the throughout larger uncensored log scores for 50% censored observations.

Table 3: Simulation 2: Mean absolute deviations between true and estimated survival curves for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
	CTM	2.60	2.74	3.31	6.34
	CLTM	8.14	8.08	8.27	10.26
G_1	Cox	9.02	8.56	7.45	7.31
	Kaplan–Meier	2.58	2.74	3.43	6.84
	Cforest	2.39	2.50	3.19	6.64
	Stratified Cox	4.16	4.07	4.19	5.85
	CTM	3.15	3.41	5.03	10.99
	CLTM	7.60	8.20	9.80	14.30
G_2	Cox	9.58	10.25	12.00	16.51
	Kaplan–Meier	2.55	2.88	4.83	11.15
	Cforest	2.34	2.64	4.63	10.96
	Stratified Cox	4.85	5.51	8.15	14.58

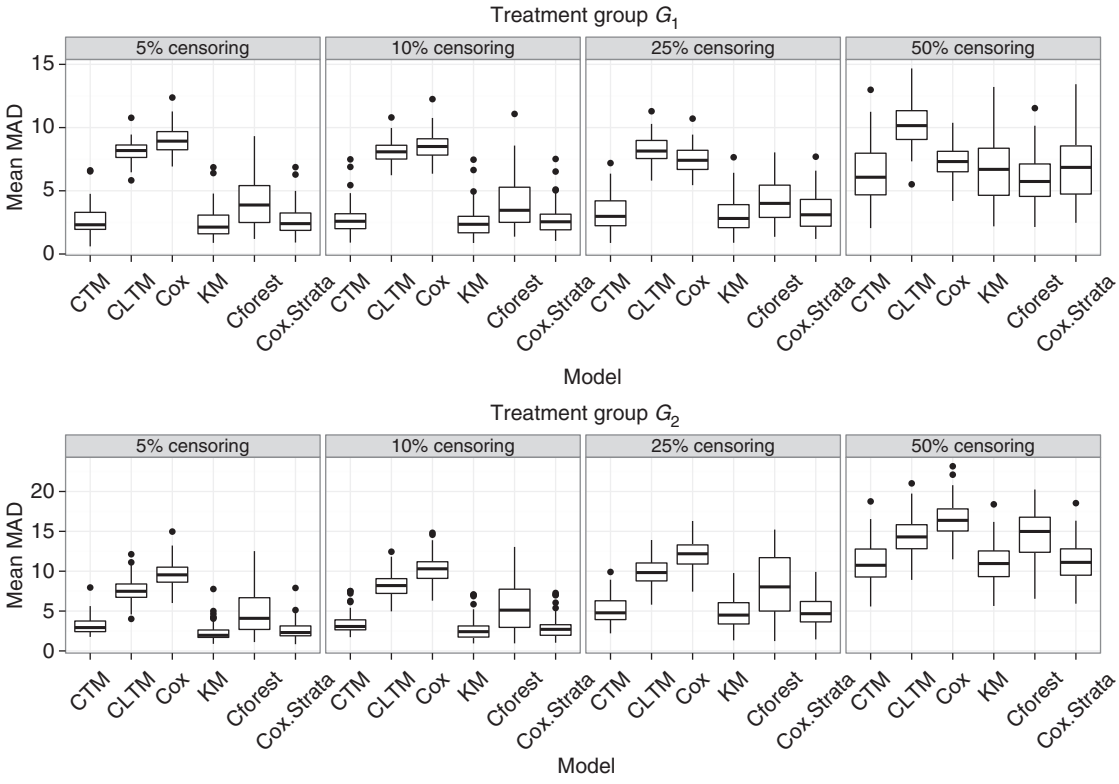

Figure 4: Simulation 2: Boxplot of the treatment-specific mean MAD values based on $B = 100$ simulations for the conditional transformation model (CTM), the conditionally linear transformation model (CLTM), the Cox model (Cox), the Kaplan–Meier estimator (KM), conditional random forests (Cforest), and the stratified Cox model (Cox.Strata): 5%, 10%, 25% and 50% of right-censored observations were observed.

3.3.3 Simulation 3

The Cox model and the CLTM approach performed almost equally well in the proportional hazards setting. The mean MADs (Table 5; Figure 6) and the out-of-sample uncensored log scores (Table 6; Figure 7) were

Table 4: Simulation 2: Out-of-sample uncensored log score based on 1,000 new observations for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
G_1	CTM	50.43	50.51	50.79	52.30
	CLTM	53.12	53.14	53.41	55.06
	Cox	53.07	52.87	52.46	52.87
	Kaplan–Meier	50.46	50.54	50.91	52.66
	Cforest	50.42	50.50	50.89	52.67
	Stratified Cox	50.89	50.90	51.06	52.03
G_2	CTM	50.75	50.92	51.65	55.46
	CLTM	53.47	53.85	55.13	60.09
	Cox	54.63	55.21	57.04	63.12
	Kaplan–Meier	50.52	50.70	51.53	55.45
	Cforest	50.48	50.66	51.47	55.37
	Stratified Cox	51.63	52.08	53.84	59.87

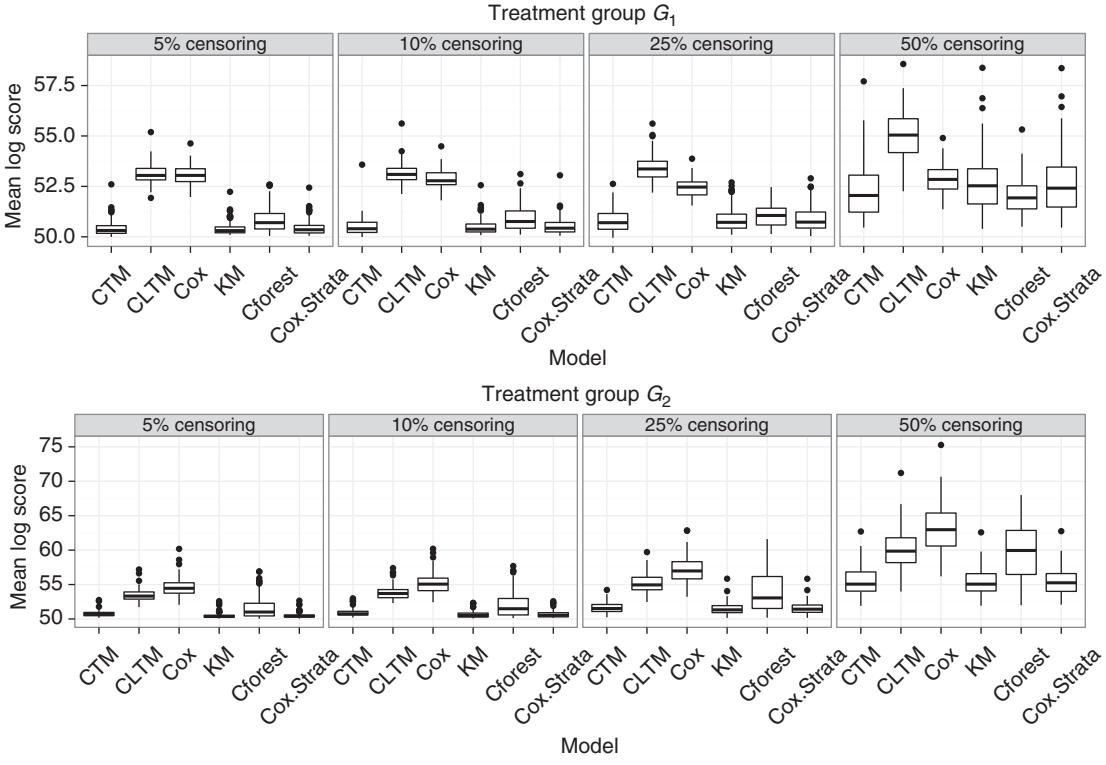

Figure 5: Simulation 2: Boxplot of the out-of-sample mean uncensored log scores based on 1,000 new observations for each treatment group and $B = 100$ simulations for the conditional transformation model (CTM), the conditionally linear transformation model (CLTM), the Cox model (Cox), the Kaplan–Meier estimator (KM), conditional random forests (Cforest), and the stratified Cox model (Cox.Strata): 5%, 10%, 25% and 50% of right-censored observations were observed.

similar for the Cox model and the CLTM, whereas the CTM was associated with slightly higher MADs and uncensored log scores. Conditional random forests performed worst because conditional random forests and the CTM were not able to profit from the proportional hazards assumption. Additionally, conditional random forests had difficulties in identifying the linear influence of x.

Table 5: Simulation 3: Mean absolute deviations between true and estimated survival curves for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
G_1	CTM	1.88	1.90	2.46	4.42
	CLTM	1.67	1.69	2.28	4.37
	Cox	1.60	1.67	2.44	5.17
	Cforest	4.04	4.42	5.52	8.51
G_2	CTM	1.97	2.06	2.71	5.07
	CLTM	1.74	1.78	2.43	4.81
	Cox	1.67	1.75	2.50	5.10
	Cforest	3.93	4.01	4.45	6.43

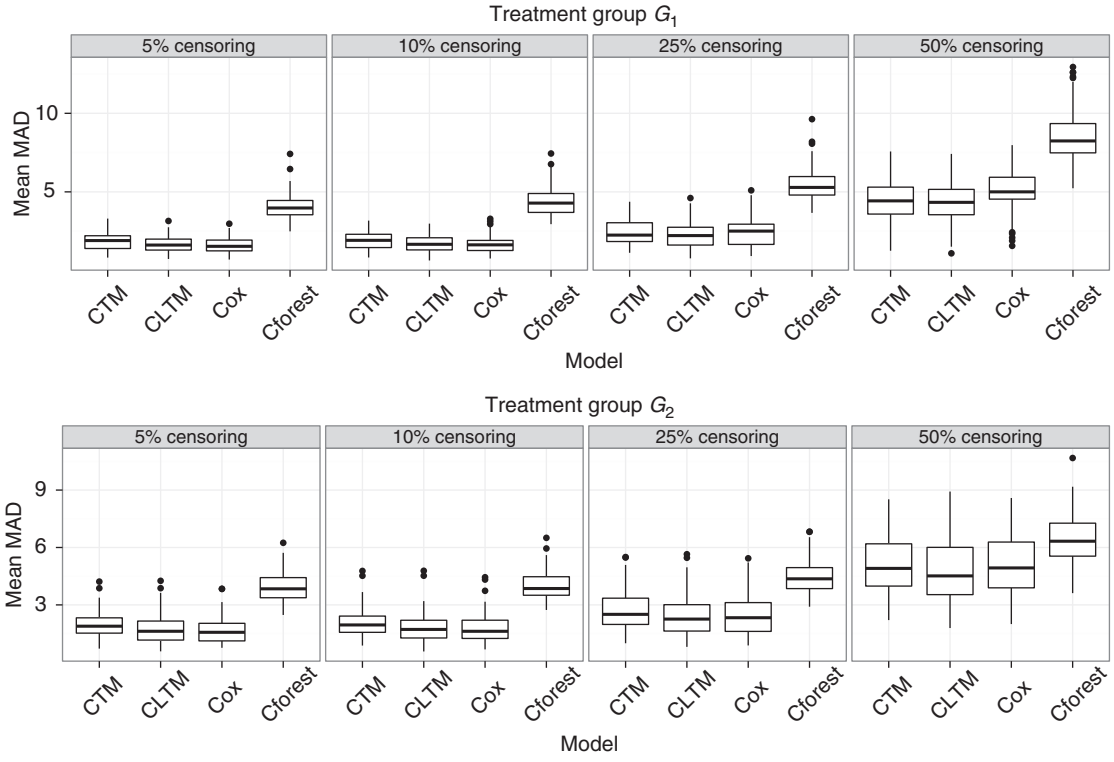

Figure 6: Simulation 3: Boxplot of the treatment-specific mean MAD values based on $B = 100$ simulations for the conditional transformation model (CTM), the conditionally linear transformation model (CLTM), the Cox model (Cox), and conditional random forests (Cforest): 5%, 10%, 25% and 50% of right-censored observations were observed.

3.3.4 Simulation 4

The CTM performed better than all alternative modelling approaches, as it showed lower MADs for all amounts of censoring than the CLTM, the Cox model, the stratified Cox model and conditional random forests (Table 7; Figure 8). Additionally, the CTM approach was associated with the smallest mean uncensored log scores (Table 8; Figure 9) because the CTM approach is the only approach that was able to account for the non-linear influence of x on the shape parameter of the Weibull distribution

Table 6: Simulation 3: Out-of-sample uncensored log score based on 1,000 new observations for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
G_1	CTM	38.96	38.99	39.21	40.08
	CLTM	38.90	38.92	39.13	39.98
	Cox	38.93	38.97	39.25	40.44
	Cforest	40.03	40.31	41.20	44.22
G_2	CTM	40.99	41.05	41.34	42.60
	CLTM	40.95	40.99	41.30	42.54
	Cox	40.86	40.91	41.20	42.40
	Cforest	41.74	41.82	42.17	44.00

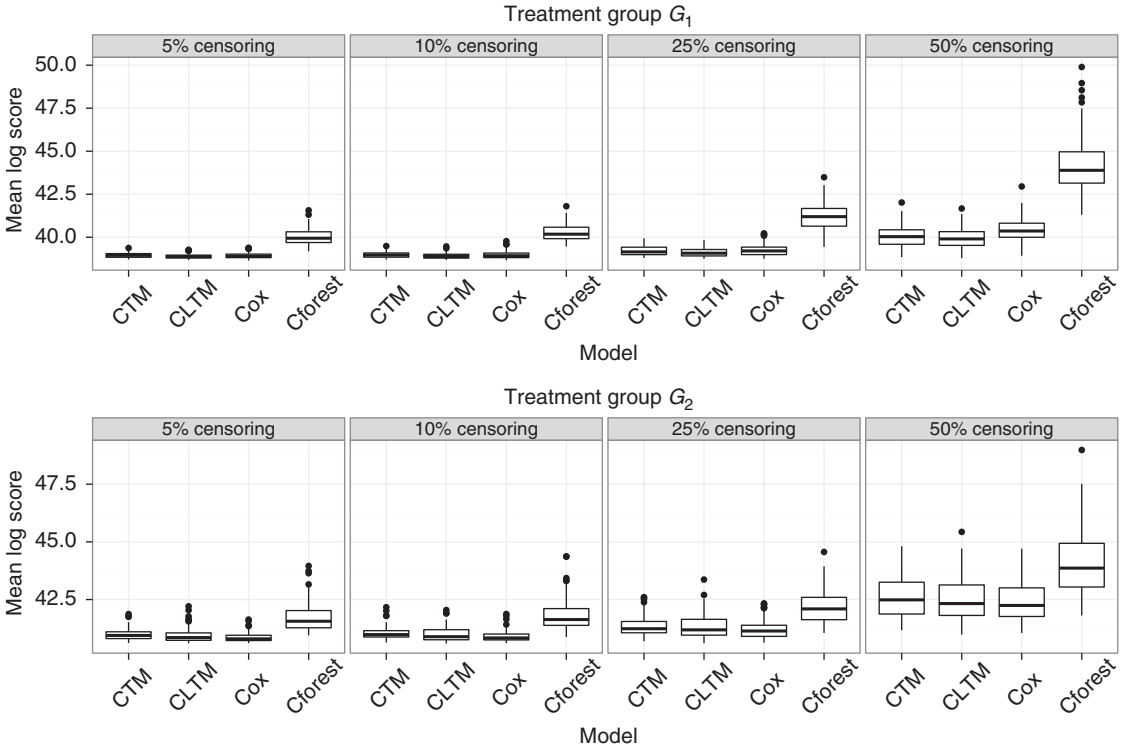

Figure 7: Simulation 3: Boxplot of the out-of-sample mean uncensored log scores based on 1,000 new observations for each treatment group and $B = 100$ simulations for the conditional transformation model (CTM), the conditionally linear transformation model (CLTM), the Cox model (Cox), and conditional random forests (Cforest): 5%, 10%, 25% and 50% of right-censored observations were observed.

adequately. The Cox model and the CLTM performed worse owing to the proportional hazards assumption. Since the non-proportionality of hazards was mainly induced by x, the stratified Cox model performed only slightly better than the ordinary Cox model in terms of the mean uncensored log score. Conditional random forests performed better than both Cox models and the CLTM as the approach can account for non-proportionality in G and x, but performed worse than the CTM owing to the non-linear influence of x on the shape parameter. Again, conditional random forests had difficulties in identifying the non-linear influence of x.

Table 7: Simulation 4: Mean absolute deviations between true and estimated survival curves for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
	CTM	2.48	2.55	3.36	6.67
	CLTM	5.09	5.12	5.60	7.79
G_1	Cox	5.75	5.82	6.43	8.71
	Cforest	4.34	4.42	4.92	7.14
	Stratified Cox	5.67	5.72	6.24	8.44
	CTM	2.38	2.46	3.26	5.78
	CLTM	4.47	4.51	4.96	6.83
G_2	Cox	5.27	5.30	5.67	7.29
	Cforest	3.82	3.85	4.29	6.26
	Stratified Cox	5.32	5.39	5.90	7.66

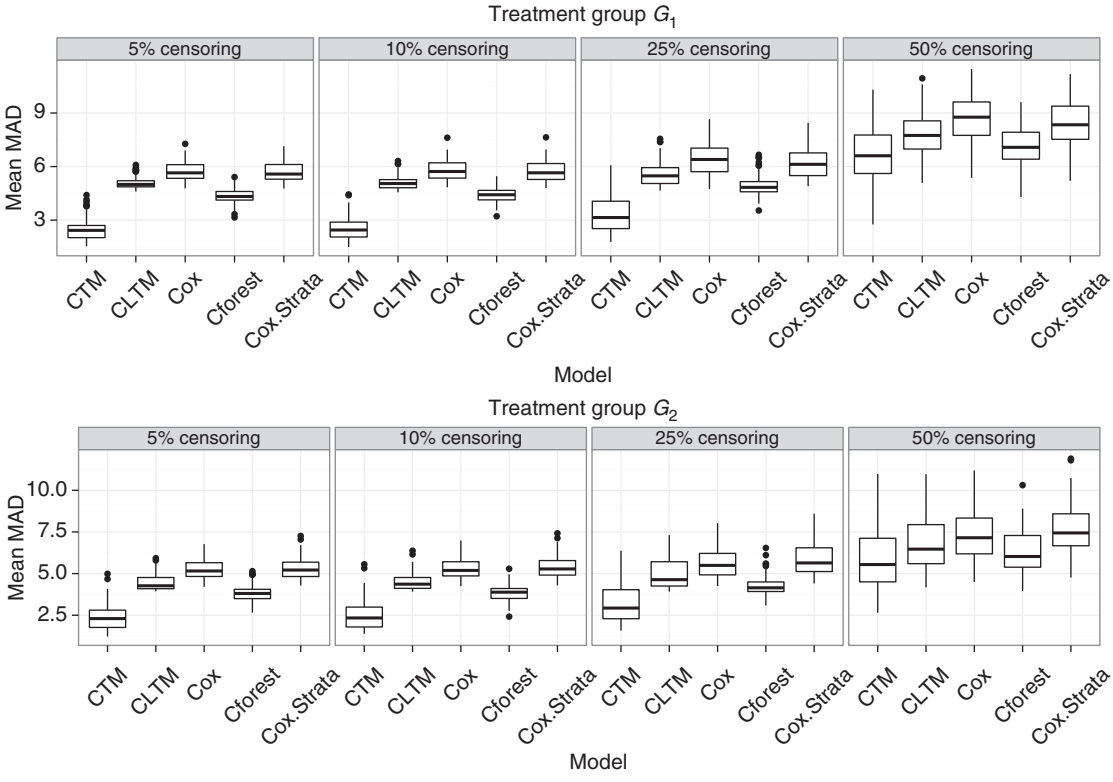

Figure 8: Simulation 4: Boxplot of the treatment-specific mean MAD values based on $B = 100$ simulations for the conditional transformation model (CTM), the conditionally linear transformation model (CLTM), the Cox model (Cox), conditional random forests (Cforest), and the stratified Cox model (Cox.Strata): 5%, 10%, 25% and 50% of right-censored observations were observed.

4 Chronic myelogenous leukaemia data

Curative bone marrow transplantation is feasible for only a minority of patients with chronic myelogenous leukaemia. Therefore, drug-based chemotherapy remains a treatment of central interest. The standard chemotherapy has long been with the cytostatic drugs BUS or HU. In a multicentre, randomized study,

Table 8: Simulation 4: Out-of-sample uncensored log score based on 1,000 new observations for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
G_1	CTM	49.23	49.28	49.64	51.29
	CLTM	50.59	50.64	51.03	52.66
	Cox	50.95	51.04	51.57	53.45
	Cforest	50.71	50.78	51.20	52.97
	Stratified Cox	50.01	50.10	50.47	52.02
G_2	CTM	49.23	49.28	49.55	50.66
	CLTM	50.21	50.25	50.46	51.46
	Cox	50.44	50.46	50.67	51.68
	Cforest	50.56	50.62	50.96	52.19
	Stratified Cox	49.78	49.81	50.01	51.04

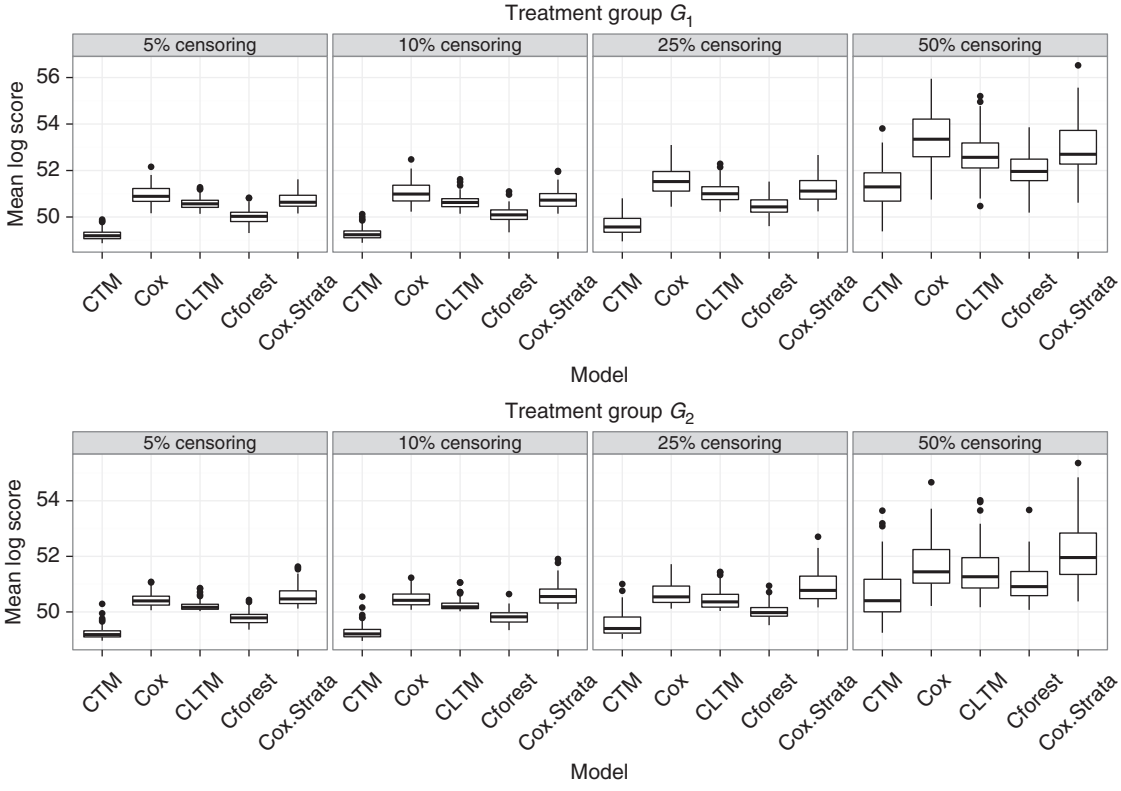

Figure 9: Simulation 4: Boxplot of the out-of-sample mean uncensored log scores based on 1,000 new observations for each treatment group and $B = 100$ simulations for the conditional transformation model (CTM), the conditionally linear transformation model (CLTM), the Cox model (Cox), conditional random forests (Cforest), and the stratified Cox model (Cox.Strata): 5%, 10%, 25% and 50% of right-censored observations were observed.

Hehlmann et al. [57] have shown that treatment with the drug IFN-α significantly prolongs survival compared to treatment with BUS, and survival times after treatment with IFN-α or HU were not significantly different. Within the scope of the study, 516 eligible patients were recruited in 57 study centres from 1983 to 1991. For 507 of the 516 patients, complete data on sex, age and a prognostic score distinguishing between low-, intermediate- and high-risk groups [58] are available. Of the 507 patients, 132 random patients were

treated with IFN-α, 182 were treated with BUS and 193 were treated with HU. Ninety patients were right-censored mainly due to bone marrow transplantation during the first chronic phase, and 417 patients died during the study period [59].

4.1 Model estimation

Herberich and Hothorn [59] analysed the treatment effects using a frailty Cox model [60] with Gaussian frailties for the 57 study centres. Furthermore, age, sex, treatment and risk group were included as linear predictors. In our re-analysis of the CML data set, the main goals were to check the validity of the proportional hazards assumption in Cox models, which have been used for analyses in the past (e.g. [59]). Moreover, we were interested in possible interactions between the explanatory variables treatment, risk group, sex and age. More precisely, we were interested in whether the superiority of the IFN-α treatment found in former studies (e.g. [57]) is present in all risk groups. Additionally, treatment effectiveness might differ between men and women and patients of different age in the different risk groups. Therefore, we fitted five models to the CML data, in which the proportional hazards assumption and the considered interaction terms differed. The random effect for the study centres was excluded in all models for the purpose of model comparison, as we found its variance to be negligibly small.

First, we estimated an ordinary Cox model with linear influences for the explanatory variables treatment, risk group, sex and age:

$$\lambda(t_\iota|\boldsymbol{x}) = \lambda_0(t_\iota) \cdot \exp(\beta_{\text{tr}} \cdot x_{\text{tr}} + \beta_{\text{risk}} \cdot x_{\text{risk}} + \beta_{\text{sex}} \cdot x_{\text{sex}} + \beta_{\text{age}} \cdot x_{\text{age}}),$$

where $\lambda_0(\cdot)$ denotes the baseline hazard function and proportional hazards were assumed.

To account for possible interactions between the categorical explanatory variables treatment, risk group and sex, we estimated an additional Cox model that included all two-time interactions between treatment, risk group and sex, and their three-time interaction. Again, all influences were assumed to be linear:

$$\begin{aligned}\lambda(t_\iota|\boldsymbol{x}) = \lambda_0(t_\iota) \cdot \exp(&\beta_{\text{tr}} \cdot x_{\text{tr}} + \beta_{\text{risk}} \cdot x_{\text{risk}} + \beta_{\text{sex}} \cdot x_{\text{sex}} + \beta_{\text{age}} \cdot x_{\text{age}} \\ &+ \beta_{\text{tr:risk}} \cdot x_{\text{tr:risk}} + \beta_{\text{tr:sex}} \cdot x_{\text{tr:sex}} + \beta_{\text{risk:sex}} \cdot x_{\text{risk:sex}} \\ &+ \beta_{\text{tr:risk:sex}} \cdot x_{\text{tr:risk:sex}}).\end{aligned}$$

Since the Cox model assumed proportional hazards for all patient characteristics, we alternatively used a CTM for data analysis. In the CTM, the proportional hazards assumption was relaxed by allowing for flexible influences of each explanatory variable over time:

$$h(t_\iota|\boldsymbol{x}) = h_{\text{tr}}(t_\iota|\text{tr}) + h_{\text{risk}}(t_\iota|\text{risk}) + h_{\text{sex}}(t_\iota|\text{sex}) + h_{\text{age}}(t_\iota|\text{age}).$$

We defined separate partial transformation functions for treatment, risk group, sex and age that were specified in terms of basis functions:

$$\begin{aligned}h_{\text{tr}}(t_\iota|\text{tr}) &= \left(\boldsymbol{b}_{\text{tr}}^{\text{lin}}(\text{tr})^{\text{T}} \otimes \boldsymbol{b}_T(t_\iota)^{\text{T}}\right)\boldsymbol{\gamma}_{\text{tr}}, \\ h_{\text{risk}}(t_\iota|\text{risk}) &= \left(\boldsymbol{b}_{\text{risk}}^{\text{lin}}(\text{risk})^{\text{T}} \otimes \boldsymbol{b}_T(t_\iota)^{\text{T}}\right)\boldsymbol{\gamma}_{\text{risk}}, \\ h_{\text{sex}}(t_\iota|\text{sex}) &= \left(\boldsymbol{b}_{\text{sex}}^{\text{lin}}(\text{sex})^{\text{T}} \otimes \boldsymbol{b}_T(t_\iota)^{\text{T}}\right)\boldsymbol{\gamma}_{\text{sex}} \text{ and} \\ h_{\text{age}}(t_\iota|\text{age}) &= \left(\boldsymbol{b}_{\text{age}}(\text{age})^{\text{T}} \otimes \boldsymbol{b}_T(t_\iota)^{\text{T}}\right)\boldsymbol{\gamma}_{\text{age}}.\end{aligned}$$

In other words, we fitted a separate function over time for each treatment, for each risk group and for each sex. For the age effect, we estimated a bivariate interaction surface depending on age and the survival time.

In analogy to the Cox model, the CTM was extended to include interaction terms. Nevertheless, we always consider interactions between the survival time and the explanatory variables in CTMs. Therefore, the three-time interaction term between treatment, risk group and sex cannot currently be considered in

CTMs. Furthermore, the number of two-time interactions should be restricted, which is why we chose to consider only the most interesting interaction between treatment and risk group:

$$h(t_\iota|\boldsymbol{x}) = h_{\text{tr:risk}}(t_\iota|\text{tr} : \text{risk}) + h_{\text{sex}}(t_\iota|\text{sex}) + h_{\text{age}}(t_\iota|\text{age}).$$

In contrast to the previous CTM, where three separate functions over time were estimated for each treatment and for each risk group, we estimated nine separate functions over time for all treatment–risk group combinations. By including the treatment–risk group interaction, we investigated whether different treatments should be considered depending on the specific risk group.

As a further comparative method, we analysed the CML data using conditional random forests. This nonparametric method is also able to relax the proportional hazards assumption and is able to consider interactions between the explanatory variables. More precisely, the method grew a survival tree by searching for significant split points in the explanatory variables treatment, risk group, sex and age. The estimated survival probabilities for the patients were obtained afterwards based on conditional Kaplan–Meier estimators for the observations in the final leaves. The bootstrap aggregation of conditional survival trees, which is performed when using conditional random forests as well, results in stable predictions of survival probabilities [29].

4.2 Model evaluation

In the previous section, we described the estimation of a Cox model, a Cox model with interactions, a CTM, a CTM with interactions and conditional random forests. The evaluation of the five different models served two main goals. First, we were interested in the validity of the proportional hazards assumption. The proportional hazards assumption could be checked by comparing the performance of the Cox models to the performance of CTMs and conditional random forests, as the proportional hazards assumption was relaxed in the latter two approaches. Moreover, we were interested in possible interactions between the explanatory variables treatment, risk group, sex and age. Thereby, all possible interactions could be considered in conditional random forests. All interactions between the *categorical* explanatory variables treatment, risk group and sex were considered in the Cox model with interactions. Owing to the higher flexibility, we only considered the treatment – risk group interaction in the CTM with interactions. Hence, the importance of interactions could be investigated by comparing the models with interactions to their counterparts without interactions, and by comparing the models with interactions among each other.

The model performance was quantified by calculating the out-of-sample censored log score given in eq. (9). For the evaluation we used the following procedure:

1. Generate $B = 100$ bootstrap samples by randomly sampling $n = 507$ observations with replacement from the patients in the CML data set. The resulting data sets are *estimation data sets*.
2. The corresponding *evaluation data sets* consist of all observations that have not been selected for the estimation data set.
3. For each bootstrap sample $b = 1, \ldots, 100$:
   (a) Estimate the five different models based on the estimation data set.
   (b) Predict the survival probabilities for the patients in the evaluation data set over a grid of time points.
   (c) Calculate the out-of-sample censored log score (eq. (9)) based on the predicted survival probabilities, the grid points over time and the IPCWs. Separate out-of-sample censored log scores result for the five model approaches.
4. Compare the $B = 100$ out-of-sample censored log scores for the five model approaches.

Two important characteristics of the above procedure have to be noted. The IPCWs and a grid of time points were needed when calculating the censored log score. Thereby, the grid of time points was fixed for all

bootstrap data sets and consisted of all event and censoring times of the CML patients. The IPCWs were calculated beforehand for *all* patients in the CML data set over the grid of time points defined above. When we selected the patients for the estimation and the evaluation data set, we also selected their respective IPCWs.

The boxplots of the out-of-sample censored log scores for the five different models revealed that the proportional hazards assumption is problematic for the CML data when only the main effects treatment, risk group, sex and age are considered (Figure 10). The CTM, the CTM with interactions and conditional random forests (i.e. the models that relax the proportional hazards assumption) showed lower out-of-sample censored log scores than the Cox model. Nevertheless, the non-proportional hazards of the main effects seem to be induced by disregarding interaction terms, as the Cox model with interaction terms performed as well as the CTM, which ignored all interaction terms but assumed non-proportional hazards. Nevertheless, the out-of-sample censored log scores for the CTM, the CTM with interactions and conditional random forests were similar. Hence, the inclusion of the treatment–risk group interaction in the CTM did not lead to model improvement. All interactions between explanatory variables could be considered in conditional random forests, but the model's predictive performance was not superior. Hence, the inclusion of interaction terms is unimportant for models that allow for non-proportional hazards.

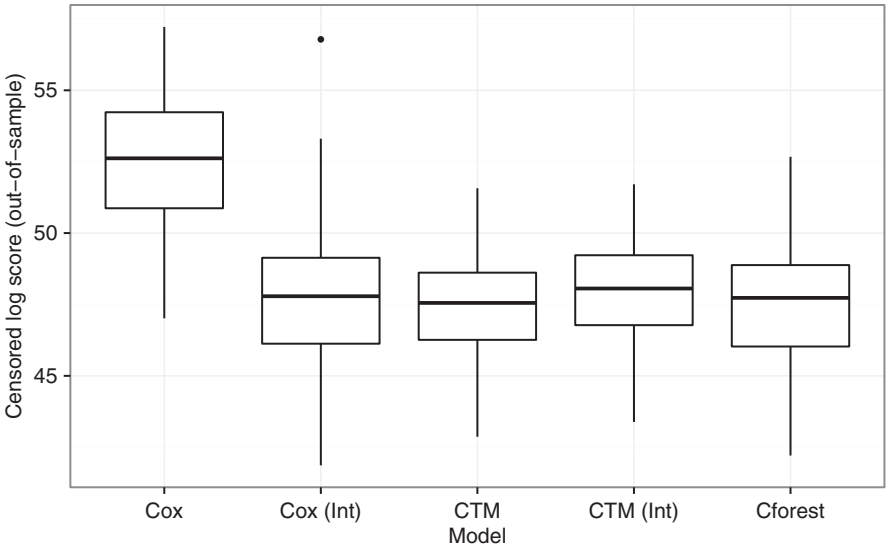

Figure 10: Out-of-sample censored log scores for the Cox model (Cox), the Cox model with interactions (Cox (Int)) the CTM (CTM), the CTM with interactions (CTM (Int)) and conditional random forests (Cforest) for 100 bootstrap evaluation data sets.

Through comparisons of models including CTMs, we found that the proportional hazards assumption is not violated for a Cox model with interactions. Hence, the best model to analyse the CML data is a Cox model that includes interaction terms between the categorical main effects.

5 Discussion

The direct estimation of the survivor function in survival data analysis is of special interest, as the reliable prediction of patient-specific survivor functions allows a better prognosis of the course of disease [61]. We propose the use of CTMs to directly estimate the conditional survivor function of the survival times given a set of patient characteristics.

The well-known Cox model is the regression model most commonly used in survival analysis [1]. One important restriction of the Cox model is the proportional hazards assumption. Of course, several strategies

deal with or identify non-proportional hazards for some of the explanatory variables. For example, if non-proportional hazards for a categorical variable are identified, the estimation of a stratified Cox model with separate baseline hazard functions for the subgroups is frequently used. Speculation about the validity of the proportional hazards assumption in the Cox model becomes superfluous when the CTM approach is used, because the proportional hazards assumption is relaxed and can be checked easily by graphic comparisons.

In our simulation, we investigated the performance of the CTM in cases of proportional hazards and non-proportional hazards and compared the performance to that of the CLTM, the (ordinary or stratified) Cox model, the Kaplan–Meier estimator and conditional random forests. We measured the performance in terms of the correspondence of true and estimated survival probabilities for new observations. In the simulation settings with informative binary treatment group and non-informative continuous explanatory variable, the CTM was able to keep up with the alternative methods in the case of proportional hazards. In the case of non-proportional hazards, the CTM clearly outperformed the ordinary Cox model and the CLTM and delivered results equally as good as those of the stratified Cox model, the Kaplan–Meier estimator and conditional random forests. In the simulation settings with informative binary treatment group and informative continuous explanatory variable, the CTM performed almost as well as the ordinary Cox model and the CLTM in the proportional hazards setting. In the non-proportional hazards setting, the CTM outperformed all alternative models, as it is the only method that was able to consider non-proportionality induced non-linearly by a continuous explanatory variable. One further advantage of the CTM was that owing to the imposed smoothness penalty, smooth estimated survival curves resulted, which is more realistic than the step functions resulting from the Cox model, the Kaplan–Meier estimator and conditional random forests. Moreover, the results of the simulation study showed that the CTM can handle up to 50% of right-censored observations without heavier losses in the quality of the resulting estimates compared to the alternative approaches.

Furthermore, we used the CTM approach to analyse survival times of patients suffering from chronic myelogenous leukaemia to check the proportional hazards assumption that has been implied when using Cox models in the past. Furthermore, we were interested in the importance of interactions between the considered explanatory variables. Therefore, the out-of-sample performances of a Cox model, a Cox model with interactions, a CTM, a CTM with interactions and conditional random forests were compared. Our analysis revealed that the violated proportional hazards assumption for the main effects treatment, risk group, sex and age was mainly induced by ignoring important interactions between the main effects. Furthermore, we would like to stress that models were checked *without* an extensive analysis of residuals.

The handling of right-censored observations is a main topic in survival analysis. In CTMs, the IPCW approach has been used to account for right-censored observations. The integrated Brier score or log score for right-censored observations are well-established scoring rules for model assessment and comparison, but, to the best of our knowledge, they have not yet been used as risk functions for model estimation. In the IPCW approach, the observations are reweighted by the inverse probability of remaining uncensored up to a specific time point. In CTMs, this probability is calculated in terms of the marginal Kaplan–Meier estimator of the censoring distribution. Hence, the weights are calculated based on observed data and, more importantly, it is assumed that the censoring mechanism does not depend on any explanatory variables. Especially the dependency of the censoring distribution on (some of) the explanatory variables would be a worthwhile extension and needs further investigation [51]. Nevertheless, Hothorn et al. [11] showed the consistency of the conditional transformation function h in CTMs, which transfers to CTMs for survival data as we only adapted the weighting scheme to account for right censoring. Mackenzie [62] previously estimated survival curves with dependent left-truncated data using Cox's model and inverse probability weighting. Thus, it would be interesting whether and how the suggested approach extends to left-truncated or interval-censored data.

Basically, three main assumptions are made when estimating CTMs for survival data. First, by assuming that the transformation function h exists, we assume that there is a monotone transformation from the unknown survival time distribution to the link function F. Second, h is decomposed additively into partial

transformation functions, whereby additivity on the scale of the transformation function is assumed. Third, the event times and the right-censoring times are assumed to be independent, which is a strong but common assumption in survival data analysis. The data analyst should be aware of these model assumptions as they might be violated.

6 Software

All analyses were carried out in the R system of statistical computing [63]. CTMs were estimated using the R add-on package ctmDevel [64]. To compare the proposed CTMs for survival data with established models, we estimated Cox models using the R add-on package survival [65], calculated Kaplan–Meier estimators using the R add-on package prodlim [66] and estimated conditional random forests using the R add-on package party [52]. R code for reproducing the results of Section 3 (in ctmDevel/inst/empeval) and Section 4 (in ctmDevel/inst/applications) is publicly available in the ctm package from the R-forge repository (https://r-forge.r-project.org/projects/ctm).

Acknowledgements: The authors are grateful to the German CML Study Group (Head: Prof. Dr R. Hehlmann) and especially to Markus Pfirrmann for providing us with the chronic myelogenous leukaemia data. The authors thank Karen A. Brune for linguistically improving the manuscript.

Funding: Financial support by Deutsche Forschungsgemeinschaft (grant DFG HO 3242-4/1).

References

1. Cox DR. Regression models and life-tables. J R Stat Soc Ser BMET 1972;34:187–220.
2. Andersen PK, Christensen E, Fauerholdt L, Schlichting P. Measuring prognosis using the proportional hazards model. Scand J Stat 1983;10:49–52.
3. Sargent D. A flexible approach to time-varying coefficients in the Cox regression setting. Lifetime Data Anal 1997; 3:13–25.
4. Scheike T, Martinussen T. On estimation and tests of time-varying effects in the proportional hazards model. Scand J Stat 2004;31:51–62.
5. Tian L, Zucker D, Wei L. On the Cox model with time-varying regression coefficients. J Am Stat Assoc 2005;100:172–83.
6. Xu R, O'Quigley J. Estimating average regression effect under non-proportional hazards. Biostatistics 2000;1:423–39.
7. Ng'Andu NH. An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. Stat Med 1997;16:611–26.
8. Schoenfeld D. Partial residuals for the proportional hazards regression model. Biometrika 1982;69:239–41.
9. Grambsch P, Therneau T. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 1994;81:515–26.
10. Lin DY, Wei LJ, Ying Z. Checking the Cox model with cumulative sums of martingale-based residuals. Biometrika 1993;80:557–72.
11. Hothorn T, Kneib T, Bühlmann P. Conditional transformation models. J R Stat Soc Ser B 2014;76:3–27.
12. Doksum KA, Gasko M. On a correspondence between models in binary regression analysis and in survival analysis. Int Stat Rev 1990;58:243–52.
13. Cheng SC, Wei LJ, Ying Z. Analysis of transformation models with censored data. Biometrika 1995;82:835–45.
14. Cheng SC, Wei LJ, Ying Z. Predicting survival probabilities with semiparametric transformation models. J Am Stat Assoc 1997;92:227–35.
15. Chen K, Jin Z, Ying Z. Semiparametric analysis of transformation models with censored data. Biometrika 2002;89:659–68.
16. Zeng D, Lin DY. Efficient estimation of semiparametric transformation models for counting processes. Biometrika 2006;93:627–40.
17. Fine JP. Regression modeling of competing crude failure probabilities. Biostatistics 2001;2:85–97.
18. Lee KH, Chakraborty S, Sun J. Bayesian variable selection in semiparametric proportional hazards model for high dimensional survival data. Int J Biostat 2011;7, Article 21. DOI: 10.2202/1557-4679.1301

19. van der Vaart A, van der Laan MJ. Estimating a survival distribution with current status data and high-dimensional covariates. Int J Biostat 2006;2, Article 9. DOI: 10.2202/1557-4679.1014
20. Lu W, Li L. Boosting method for nonlinear transformation models with censored survival data. Biostatistics 2008;9:658–67.
21. Kaplan E, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457–81.
22. Dabrowska D. Non-parametric regression with censored survival time data. Scand J Stat 1987;14:181–97.
23. Dabrowska D. Uniform consistency of the kernel conditional Kaplan-Meier estimate. Ann Stat 1989;17:1157–67.
24. González Manteiga W, Cadarso-Suarez C. Asymptotic properties of a generalized Kaplan-Meier estimator with some applications. J Nonparametr Stat 1994;4:65–78.
25. Iglesias Pérez C, González Manteiga W. Strong representation of a generalized product-limit estimator for truncated and censored data with some applications. J Nonparametr Stat 1999;10:213–44.
26. McKeague I, Utikal K. Inference for a nonlinear counting process regression model. Ann Stat 1990;18:1172–87.
27. Li G, Doss H. An approach to nonparametric regression for life history data using local linear fitting. Ann Stat 1995;23: 787–823.
28. Spierdijk L. Nonparametric conditional hazard rate estimation: A local linear approach. Comput Stat Data Anal 2008;52:2419–34.
29. Hothorn T, Lausen B, Benner A, Radespiel-Tröger M. Bagging survival trees. Stat Med 2004;23:77–91.
30. Meinshausen N. Quantile regression forests. J Mach Learn Res 2006;7:983–99.
31. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat 2008;2:841–60.
32. Strobl C, Boulesteix A-L, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinf 2007;8:25.
33. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. J Comput Graph Stat 2006;15:651–74.
34. Mogensen UB, Ishwaran H, Gerds TA. Evaluating random forests for survival analysis using prediction error curves. J Stat Softw 2012;50:1–23.
35. Chernozhukov V, Hong H. Three-step censored quantile regression and extramarital affairs. J Am Stat Assoc 2002;97: 872–82.
36. Honoré B, Khan S, Powell J. Quantile regression under random censoring. J Econometrics 2002;109:67–105.
37. Peng L, Huang Y. Survival analysis with quantile regression models. J Am Stat Assoc 2008;103:637–49.
38. Portnoy S. Censored regression quantiles. J Am Stat Assoc 2003;98:1001–12.
39. Powell J. Censored regression quantiles. J Econometrics 1986;32:143–55.
40. Wang H, Wang L. Locally weighted censored quantile regression. J Am Stat Assoc 2009;104:1117–28.
41. Wey A, Wang L, Rudser K. Censored quantile regression with recursive partitioning-based weights. Biostatistics 2014;15:170–81.
42. Dette H, Volgushev S. Non-crossing non-parametric estimates of quantile curves. J R Stat Soc Ser B 2008;70:609–27.
43. Aalen O. Heterogeneity in survival analysis. Stat Med 1988;7:1121–37.
44. Clayton D, Cuzick J. Multivariate generalizations of the proportional hazards model. J R Stat Soc Ser A 1985;148:82–117.
45. McGilchrist C, Aisbett C. Regression with frailty in survival analysis. Biometrics 1991;47:461–6.
46. Vaida F, Xu R. Proportional hazards model with random effects. Stat Med 2000;19:3309–24.
47. Möst L, Schmid M, Faschingbauer F, Hothorn T. Predicting birth weight with conditionally linear transformation models. Stat Methods Med Res 2014. DOI: 10.1177/0962280214532745
48. Gneiting T, Raftery A. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 2007;102:359–78.
49. Schemper M, Henderson R. Predictive accuracy and explained variation in Cox regression. Biometrics 2000;56:249–55.
50. van der Laan MJ, Robins JM. Unified methods for censored longitudinal data and causality. New York, NY: Springer, 2003
51. Gerds T, Schumacher M. Consistent estimation of the expected brier score in general survival models with right-censored event times. Biometrical J 2006;48:1029–40.
52. Hothorn T, Bühlmann P, Dudoit S, Molinaro A, van der Laan MJ. Survival ensembles. Biostatistics 2006;7:355–73.
53. Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics 2000;56:779–88.
54. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. Stat Med 1999;18:2529–45.
55. Bühlmann P, Hothorn T. Boosting algorithms: Regularization, prediction and model fitting. Stat Sci 2007;22:477–505.
56. Schmid M, Hothorn T. Boosting additive models using component-wise P-splines as base-learners. Comput Stat Data Anal 2008;53:298–311.
57. Hehlmann R, Heimpel H, Hasford J, and Others. Randomized comparison of interferon-α with busulfan and hydroxyurea in chronic myelogenous leukemia. Blood 1994;84:4064–77.
58. Hasford J, Pfirrmann M, Hehlmann R, and Others. A new prognostic score for survival of patients with chronic myeloid leukemia treated with interferon alfa. J Natl Cancer Inst 1998;90:850–58.
59. Herberich E, Hothorn T. Dunnett-type inference in the frailty Cox model with covariates. Stat Med 2012;31:45–55.
60. McGilchrist CA, Aisbett CW. Regression with frailty in survival analysis. Biometrics 1991;47:461–66.

61. Mackillop WJ, Quirt CF. Measuring the accuracy of prognostic judgments in oncology. J Clin Epidemiol 1997;50:21–9.
62. Mackenzie T. Survival curve estimation with dependent left truncated data using Cox's model. Int J Biostat 2012;8, Article 29. DOI: 10.1515/1557-4679.1312
63. R Core Team. 2013. R: A language and environment for statistical computing. Available at: http://www.R-project.org/
64. Hothorn T. 2013. ctmDevel: Conditional transformation models. Available at: https://r-forge.r-project.org/projects/ctm, R package version 0.1-0. SVN revision 56.
65. Therneau TM. 2013. Survival analysis. Available at: http://CRAN.R-project.org/package=survival, R package version 2.37-4.
66. Gerds TA. 2013. prodlim: Product limit estimation for event history and survival analysis. Available at: http://CRAN.R-project.org/package=prodlim, R package version 1.3.7.