Support vector methods for survival analysis: a comparison between ranking and regression approaches

V. Van Belle[1], K. Pelckmans[2], S. Van Huffel[1], and J.A.K. Suykens[1]

[1] Katholieke Universiteit Leuven, ESAT-SCD
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
{vanya.vanbelle,johan.suykens,sabine.vanhuffel}@esat.kuleuven.be
[2] Department of Information Technology,
University of Uppsala, SE-751 05 Uppsala, Sweden
kp@it.uu.se

Abstract. *Objective*: This paper presents empirical evidence for the use of machine learning based techniques for the estimation of non-linear transformation models for the analysis of survival data.
Methods: The task of estimating a transformation model based on a dataset reflecting survival of patients was embedded in a context of machine learning and Support Vector Machines (SVMs) before. The key idea is to rephrase the task as a ranking problem via the concordance index, a problem which can be solved efficiently in a context of structural risk minimization and convex optimization techniques. To establish the desired ranking, each observation is compared with his closest neighbor. A principal advantage of this approach is the use of non-linear kernels implementing automatically non-parametric effects of the covariates. The goal of this paper is then twofold: (i) we present additional empirical material supporting this technique based on 6 clinical cancer-related datasets, and discuss as well the relevance of the present technique over classical approaches for such data; and (ii) we investigate empirically how the SVM-based technique can be strengthened to include so-called 'regression-constraints', interpolating the method between ranking and point-prediction. A drawback of the ranking formulation of such models is that the ranking is established on an arbitrary choice of pairs of observations. It is therefore investigated whether the performance increases when dividing up the observations based on intervals on the event-times, a technique which can be implemented using dummy observations.
Results: We compare SVM-based survival models based on ranking constraints and SVM-based models based on regression constraints with models containing both ranking and regression constraints. The performance of the models was compared by means of three different measures: (i) the concordance index, measuring the model's discriminating ability; (ii) the logrank test statistic, indicating whether patients with a prognostic index lower than the median prognostic index have a significant different survival than patients with a prognostic index higher than the median; and (iii) the hazard rate after normalization to restrict the prognostic index between 0 and 1. The results on six clinical cancer survival datasets

show a significant improvement in performance for methods incorporating regression constraints compared with models only incorporating ranking constraints. In addition, our experiments do not show significant differences when comparing arbitrary ranking constraints with the use of dummy observations.
Conclusions: This work gives empirical evidence that SVM-based models using regression constraints perform significantly better than SVM-based models based on ranking constraints only. Secondly, it is shown that the estimation of the prognostic index is not improved when pinning the comparisons to dummy observations instead of arbitrarily choosing real observations to compare with. Our experiments did not reveal a significant difference between the SVM-based models and the proportional hazard model. However, the former models have the advantage that they are easily extendable towards non-linear models without the need to check non-linearities in the variables before modelling.

Keywords: failure time data, survival analysis, support vector machines, concordance index, censoring, prognostic models in medicine

1 Introduction

Survival studies arise in different areas. Although they are most well known to occur in medical and in particular in cancer studies, they also occur in economics (e.g. prediction of bankruptcy of factories), in mechanics (e.g. failure of airplanes, breakdown of engines, etc.), electronics (e.g. lifetime of electrical components), social sciences (e.g. estimating the time from marriage to divorce) and many other topics. Depending on the question at study one is interested in risk groups (which group of patients/components is more likely to experience the event?) or time predictions (before which time should the engine be replaced to decrease the risk of failure?).

The survival literature describes different models to answer these questions. Many common methods including the proportional hazard model (COX model) and logg-odds model can be phrased as a subset of the class of transformation models (TM)[1,2,3,4,5,6]. TMs are models which consist of a sequence of (i) assembling a prognostic index based on the covariates, and (ii) linking this prognostic index to the observed event times. One is in general only concerned with finding an explicit representation of the first stage. The standard COX model [7] for example avoids this second step by assuming that the hazard (the instantaneous risk to observe the event now, knowing that the event did not occur before) is proportional to an unspecified baseline hazard. Other models assume a fixed form for the transformation function h. The accelerated failure time model is one example which assumes that the transformation function $h(y)$, with y the outcome under study, equals the logarithmic function, and the proportional odds model takes $h(y) = \text{logit}(y)$ [6].

The approach that we take in the survival models based on support vector machines (SVM) [8] is different from the approaches taken in statistical models. Although also classifiable as a TM, SVM-based models do not assume a true

underlying function for which the parameters need to be estimated. Instead, as elaborated in [9], the empirical risk of misranking two instances with regard to their failure time, is minimized. The survival problem was therefore reformulated as a ranking problem. To reduce the computational load, a simplified version comparing each observation only with its closest neighbor instead of with all other observations, was proposed in [10]. A more theoretical framework was provided in [8]. We will refer to the survival model proposed in the latter work as MODEL 1. In this work, we question whether the inclusion of regression constraints can improve the performance. Therefore, the performance of MODEL 1 is compared with that of MODEL 2, including ranking and regression constraints. Secondly, we investigate whether inclusion of dummy observations to solve the arbitrary choice of the ranking constraints in the model, significantly improves the performance. The two models proposed above are compared with survival methods only including ranking constraints (see [9,10,11]) and only including regression constraints (see [12,13]). Table 1 gives an overview of the different models handled in this work, their constraints, the number of tuning parameters in case of a linear kernel and how the ranking constraints are defined.

This paper is organized as follows. Section 2 gives an overview of transformation models in survival analysis. Section 3 starts with a summary of existing SVM-based survival methods, followed by the introduction of two new models, proposed by the authors. Section 4 illustrates the 6 clinical datasets and performance measures, which are used in the experiments. In addition to the methods mentioned before, the experiments include the performance of the COX model for comparison.

The following notations are used throughout the text. $\mathcal{D}$ denotes the set of observation $\{x_i, y_i, \delta_i\}_{i=1}^n$, where x_i is a d-dimensional covariate vector, y_i is the corresponding survival time and δ_i denotes whether an event was observed ($\delta_i = 1$) or the observation was right censored (δ_i=0). For notational convenience, it is assumed that the observations in $\mathcal{D}$ are sorted such that for two observations $\{(x_i, y_i, \delta_i), (x_j, y_j, \delta_j)\}$ with $j < i$, it applies that $y_j < y_i$.

2 Transformation models

TMs describe the type of models in which a transformation of the outcome variable instead of the outcome variable itself is estimated as a function of the covariates. Initially, TMs were introduced in regression problems where the normality assumption on the distribution of the errors and the constant variance were not satisfied. A standard regression model for example tries to model the outcome y as a linear combination of the covariates:

$$y = w^T x + \epsilon\,, \tag{1}$$

where w is a coefficient vector and ϵ is the error variable. In cases where y is not normally distributed, the regression can be improved by transforming the dependent variable.

The general idea of transformation models can be summarized as follows [1]. Consider a model

$$h(y) = w^T x + \epsilon\,, \tag{2}$$

where w represents the regression parameters and ϵ is the error variable with density $f_\epsilon(\cdot)$. A TM specifies that some monotone function h of the response variable is linearly related to the regression variables (see equation (2)). Generalization towards non-linear functions $u(x) = w^T\varphi(x)$, with φ the feature map, of the covariates leads to

$$h(y) = u(x) + \epsilon\,. \tag{3}$$

TMs are also used in analyzing survival data. A large family of TMs in survival analysis are based on survival time distribution models, where one replaces the parameters of the model by a linear function of the covariates. The Weibull regression model and the accelerated failure time model are two examples of this type [14]. However, the most commonly used TM is the proportional hazard model (COX model) [7]. In this model the hazard at time t, defined as

$$\lambda(t|x) = \lim_{\Delta t \to 0} \frac{P\left(t \le y < t + \Delta t \middle| x, y \ge t\right)}{\Delta t}\,, \tag{4}$$

is modelled as the product of an undefined baseline hazard $\lambda_0(t)$ multiplied with the exponential of a linear combination of the covariates:

$$\lambda(x,t) = \lambda_0(t)\exp(w^T x)\,. \tag{5}$$

This leads to estimating the survival function $S(x,t) = P(y > t|x)$ as

$$S(x,t) = S_0(t)^{\exp(w^T x)}\,. \tag{6}$$

Taking $\ln(-\ln(\cdot))$ of both sides leads to

$$\ln(-\ln(S(x,t))) = \ln(-\ln(S_0(t))) - w^T x\,. \tag{7}$$

Written in the form of a TM this becomes

$$\epsilon = h(y) - u(x)\,. \tag{8}$$

The difference between this model and the parametric transformation models discussed previously is that in the previous models the transformation function $h(\cdot)$ was fixed, which is not the case in the COX model.

Although standard TMs provide a *linear* relation between the explanatory variables and a function of the expected value of the dependent variable, these methodologies can be generalized to *non-linear* relations. An example is the extension of the COX model with smoothing splines [15,16].

3 Kernel-based survival models

This Section starts with a brief discussion of existing survival models based on SVMs. In a second Subsection, two new methods are proposed. Since the outcome of this type of survival models can, in general, not be interpreted as a failure time, we will denote the outcome of the model as the prognostic index $u(x)$ instead of the prediction of the model. For the COX model this corresponds to $u(x) = w^T x$.

3.1 Existing svm-based survival models

The excellent performance of support vector machines for classification and regression led to the question whether this type of models can be extended towards other statistical problems. When analyzing survival data, one is interested in the time between a certain starting point and the occurrence of a predefined event. A first approach one could think of is to use regression models to model the time of recurrence. However, survival data typically contain datapoints with incomplete information, called censored data. Although different types of censoring exist [6], we will restrict our attention to right censoring in this work. Right censored observations are observations for which the exact time of event is not known, but where a lower bound on this time is known. Only incorporating observations for which the exact failure time is known would lead to underestimated survival times. Different proposals on how to model the survival problem by means of support vector machines are elaborated below.

Support vector regression for censored data In standard support vector regression [17] the prediction is defined as a linear combination of a transformation of the variables $\varphi(x)$, with $\varphi(\cdot)$ the feature map, plus a constant $b : \hat{y} = w^T\varphi(x) + b$. To obtain this prediction, the coefficients w and the constant b need to be calculated. In order to obtain good generalizability properties, the coefficients are kept small. SVMs are formulated as optimizations problems, where a loss function should be minimized subject to certain constraints. The constraints include that the prediction $\hat{y}$ should be larger than $y - \epsilon$, with $\epsilon > 0$. Similarly $-\hat{y}$ should be larger than $-y - \epsilon^*$, with $\epsilon^* > 0$. The loss function penalizes large values of ϵ and ϵ^* such that the resulting predictions $\hat{y}$ will be close to the observed values y. Formally, the problem is formulated as:

$$\begin{aligned} \min_{w,b,\epsilon,\epsilon^*} \quad & \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} (\epsilon_i + \epsilon_i^*) \\ \text{subject to} \quad & \begin{cases} w^T\varphi(x_i) + b \geq y_i - \epsilon_i, & \forall\, i = 1, \ldots, n \\ -w^T\varphi(x_i) - b \geq -y_i - \epsilon_i^*, & \forall\, i = 1, \ldots, n \\ \epsilon_i \geq 0, & \forall\, i = 1, \ldots, n \\ \epsilon_i^* \geq 0, & \forall\, i = 1, \ldots, n\,. \end{cases} \end{aligned} \tag{9}$$

The parameter γ is a strict positive regularization constant and ϵ and ϵ^* are slack variables allowing for errors in the predictions of the training data. The predicted outcome for a new point $x^\star$ is then found as:

$$\hat{y}(x^\star) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\varphi(x_i)^T\varphi(x^\star) + b\,, \tag{10}$$

with α_i and α_i^* the Lagrange multipliers. An advantage of SVM-based models, is that the feature map $\varphi(\cdot)$ does not need to be defined. According to Mercer's theorem [18], $\varphi(x_i)^T\varphi(x_j)$ can be written as

$$k(x_i, x_j) = \varphi(x_i)^T\varphi(x_j)\,, \tag{11}$$

provided that $k(\cdot,\cdot)$ is a positive definite kernel. Often used kernels are

- the linear kernel: $k(x,z) = x^Tz$,
- the polynomial kernel of degree a: $k(x,z) = (\tau + x^Tz)^a$, with $\tau \geq 0$
- the RBF kernel: $k(x,z) = \exp\left(-\frac{||x-z||_2^2}{\sigma^2}\right)$.

More recently, a kernel for clinical data [19] was proposed as an additive kernel $k(x,z) = \sum_{p=1}^{d} k_p(x,z)$, where $k_p(\cdot,\cdot)$ depends on the type of the variable p. For continuous and ordinal variables, $k_p(\cdot,\cdot)$ is defined as

$$k_p(x_p, z_p) = \frac{c - |x_p - z_p|}{c}\,, \tag{12}$$

with x_p the p^{th} covariate of observation x, $c = \max_p - \min_p$, with $\min_p$ and $\max_p$ the minimal and maximal value of the p^{th} covariate in the given training dataset $\mathcal{D}$. For categorical and binary data $k_p(\cdot,\cdot)$ is defined as

$$k_p(x_p, z_p) = \begin{cases} 1 \text{ if } x_p = z_p \\ 0 \text{ if } x_p \neq z_p\,. \end{cases} \tag{13}$$

The problem of using support vector regression for censored data lies in the uncertainty about the outcomes y. The earliest approaches to censored data using regression strategies either omitted the censored observations, resulting in underestimated failure times, or treated the censored observations as non-events, resulting in biased models. In order to use all the available information (censored observations give information on the interval wherein the event occurs) Shivaswamy [12] proposed a support vector regression approach to censored data (SVCR). For uncensored observations, the same constraints account as in the standard SVM regression model. For right censored observations, it is known that the failure did not occur until their censoring time. The first constraint in (9) is therefore still valid. However, the second constraint is too restrictive for

right censored observations. The support vector regression model for censored data as proposed by Shivaswamy can therefore be formulated as:

$$\boxed{\text{SVCR}}$$

$$\min_{w,b,\epsilon,\epsilon^*} \quad \frac{1}{2}w^T w + \gamma \sum_{i=1}^{n}(\epsilon_i + \epsilon_i^*)$$

$$\text{subject to} \begin{cases} w^T\varphi(x_i) + b \geq y_i - \epsilon_i, & \forall\ i = 1,\ldots,n \\ -\delta_i(w^T\varphi(x_i) + b) \geq -\delta_i y_i - \epsilon_i^*, & \forall\ i = 1,\ldots,n \\ \epsilon_i \geq 0, & \forall\ i = 1,\ldots,n \\ \epsilon_i^* \geq 0, & \forall\ i = 1,\ldots,n\,. \end{cases} \tag{14}$$

The prognostic index for a new point $x^\star$ is found as

$$\hat{u}(x^\star) = \sum_i (\alpha_i - \delta_i\alpha_i^*)\varphi(x_i)^T\varphi(x^\star) + b\,, \tag{15}$$

with α_i and α_i^* the Lagrange multipliers.

A second proposal to SVM regression for censored data was made by Khan [13] (SVMc). The difference between both models lies in the applied penalty or loss for incorrect prognostic indices (see Figure 1), which can be interpreted as predicted failure times for regression approaches. The SVCR method penalizes incorrect predictions the same whether the prediction was higher or lower than the observed failure time and penalizes incorrect predictions for right censored data only if the prediction is lower than the observed censoring time. The reasoning being that it is only known that the failure time is higher than the observed outcome. In addition, the penalty for wrong predictions is the same, whether it consists of a prediction for censored or observed failure times. On the contrary, SVMc applies different penalties for the four possible cases: (i) penalty γ for events with predicted survival less than the observed survival; (ii) penalty γ^* for events with predicted survival larger than the observed survival, with $\gamma^* > \gamma$; (iii) penalty γ_c for right censored data with predicted survival less than the observed censoring time; and (iv) penalty γ_c^* for right censored data with predicted survival larger than the observed censoring time, with $\gamma_c^* < \gamma_c$. Additionally, this model provides different $\epsilon-$losses for events and right censored data and for predictions higher and lower than the observed time. Therefore, the major drawback of the latter method is the large number of hyperparameters.

Support vector machines based on ranking constraints When not using regression models, survival problems are often translated into classification problems answering the question whether the patient survives a certain predefined time (e.g. surviving 5 years after surgery). However, to be able to include as much events as possible, this predefined time should be taken very late. On the contrary, one would like an early time, in order to retain as much patients as possible, since all patients censored before that time will be lost for the analysis.

A second problem in this approach is that the validity of the method decreases when more and more patients are censored earlier. To overcome the problems described above, it was proposed to formulate the survival problem as a ranking problem in [9,11] and a computational simplification was proposed in [10]. The idea behind formulating the survival problem as a ranking problem is that in clinical applications one is often interested in defining risks groups. One is not primarily interested in a prediction of the survival time, but in whether the patient has a high or low risk for the event to occur, such that appropriate treatment can be given. To obtain this goal, an SVM ranking method is used, similar to the rankSVM model for ranking or preference learning [20]. The method proposed by [9] and [11] involves a regularization as usual and a penalization for each comparable pair of datapoints for which the order in prognostic index differs from the observed order. A data pair $\{(x_i, y_i, \delta_i), (x_j, y_j, \delta_j)\}$ is said to be comparable whenever the order of their event times is known (e.g. two events, or an event and a right censored data point for which the censoring time of the latter is later than the failure time of the former). More formally, a comparability indicator for a pair of observations $\{(x_i, y_i, \delta_i), (x_j, y_j, \delta_j)\}$ is defined as

$$\text{comp}(i,j) = \begin{cases} 1 \quad \text{if} \quad \delta_i = 1 \;\&\; \delta_j = 1 \\ \qquad\quad \delta_i = 1 \;\&\; \delta_j = 0 \;\&\; y_i \leq y_j \\ 0 \text{ else.} \end{cases} \tag{16}$$

The model is then formulated as

RANKSVMC

$$\begin{aligned} \min_{w,\epsilon} \quad & \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} \sum_{\substack{j\,:\, y_i > y_j \\ \text{comp}(i,j)=1}} \epsilon_{ij} \\ \text{subject to} \quad & \begin{cases} w^T(\varphi(x_i) - \varphi(x_j)) \geq 1 - \epsilon_{ij}, \\ \qquad\qquad \forall\; i = 1, \ldots, n;\; \forall\; j : y_i > y_j \text{ and } \text{comp}(i,j) = 1 \\ \epsilon_{ij} \geq 0,\; \forall\; i = 1, \ldots, n;\; \forall\; j : y_i > y_j \text{ and } \text{comp}(i,j) = 1\,. \end{cases} \end{aligned} \tag{17}$$

Note that addition of a constant term b is not useful here since only differences in prognostic index are considered. The prognostic index for a new point $x^\star$ is found as

$$\hat{u}(x^\star) = \sum_{i=1}^{n} \sum_{\substack{j\,:\, y_i > y_j \\ \text{comp}(i,j)=1}} \alpha_{ij} \left(\varphi(x_i) - \varphi(x_j)\right)^T \varphi(x^\star)\,, \tag{18}$$

with α_{ij} the Lagrange multipliers.

A drawback of the method is that a quadratical programming (QP) problem of $\mathcal{O}(n^2)$ constraints needs to be solved. To overcome this problem, a computationally simplified approach solving a QP of $\mathcal{O}(n)$ constraints was proposed in [10]. The reduction was found by comparing each data point with its nearest

comparable neighbor instead of comparing with all comparable data points. We will refer to this simplified model as RANKSVMC.

An important result on how to relate standard statistical survival models like the proportional hazard model [7] and the accelerated failure time model [6] with SVM-based survival models, is found in transformation models [8]. In the latter work, a survival model was presented similar to the RANKSVMC model, the difference lying in the right hand side of the first set of constraints in equation (17), where 1 is replaced by $y_{(i)} - y_{\bar{j}(i)}$, where $\bar{j}(i)$ is defined as

$$\begin{aligned} \bar{j}(i) = \quad & \arg\max_{j} \quad j \\ & \text{subject to} \quad \begin{cases} \text{comp}(i,j) = 1 \\ y_j < y_i \,, \end{cases} \end{aligned} \tag{19}$$

for a given value of i. Remark that in the case without censoring one has $\text{comp}(i,j) = 1,\ \forall\ i = 1,\ldots,n; j \neq i$, such that $\bar{j}(i) := i - 1$. Since this will be the model that we will start from to propose new models, we will refer to this model as MODEL 1. This model is formulated as

$$\begin{aligned} & \boxed{\text{MODEL 1}} \\ & \min_{w,\epsilon} \quad \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} \epsilon_i \\ & \text{subject to} \quad \begin{cases} w^T(\varphi(x_i) - \varphi(x_{\bar{j}(i)})) \geq y_i - y_{\bar{j}(i)} - \epsilon_i, \forall\ i = 1,\ldots,n \\ \epsilon_i \geq 0, \qquad\qquad\qquad\qquad\qquad\qquad \forall\ i = 1,\ldots,n\,. \end{cases} \end{aligned} \tag{20}$$

The prognostic index for a new point $x^\star$ is found as

$$\hat{u}(x^\star) = \sum_{i=1}^{n} \alpha_i (\varphi(x_i) - \varphi(x_{\bar{j}(i)}))^T \varphi(x^\star)\,, \tag{21}$$

with α_i the Lagrange multipliers. We refer the interested reader to Appendix A for a deviation of the model.

3.2 New svm-based models

In this Section we propose two different TMs making use of SVM survival models. Firstly, the proposal to reduce the complexity of ranking formulation of [10] will be used. Secondly, it will be investigated whether the performance of the model can be improved by replacing this arbitrary choice of comparisons with unknown dummy observations. We will assume $u = w^T\varphi(x) + b$, unless stated otherwise.

In survival analysis the primary question to be answered is not to predict the failure time as accurately as possible, but to define risk profiles to the observations. This is exactly what is done in the COX model: a prognostic index $u(x)$ is estimated and the link function between this predictor and the true survival time is left unspecified. Kernel-based models for survival analysis also ignore $h(\cdot)$, but additionally let the estimation of the prognostic index free, resulting in more

flexible models. The core idea of survival SVM in [8] is that a prognostic index can be found by minimizing the empirical risk of misranking two observations (see Figure 2).

Instead of trying to estimate the survival time, a prognostic index is sought that estimates the order in which objects/persons relapse. More formally, the model becomes

$$\begin{aligned} \min_{w,\epsilon} \quad & \sum_{i=1}^{n} \epsilon_i \\ \text{subject to } & \begin{cases} u(x_i) - u(x_{\bar{j}(i)}) + \epsilon_i \geq \rho_i, & \forall\, i = 1, \ldots, n \\ \epsilon_i \geq 0, & \forall\, i = 1, \ldots, n\,, \end{cases} \end{aligned} \tag{22}$$

where ρ_i are given positive constants and $u = w^T \varphi(x)$.

Two more problems need to be solved: (i) how to define ρ_i and (ii) how to include regularization. It is seen that the RANKSVMC model is a special case of the above formulation taking $\rho_i = 1$ and regularizing using the maximal margin principle [17]. The methods that we propose here, go further along the work of [8], using the Lipschitz constant instead of the maximal margin for regularization purposes. Doing so solves the question as how to define ρ, since this methodology takes $\rho_i = y_i - y_{\bar{j}(i)}$.

Since MODEL 1 is built by minimizing the empirical risk on misrankings, the value of the prognostic model only has a relative importance. When one is interested in the prognostic model for which the prognostic index can be interpreted as times, one might include regression constraints as in [12,13]. MODEL 1 is therefore modified (MODEL 2) as described below:

MODEL 2

$$\begin{aligned} \min_{w,\epsilon,\xi,\xi^*,b} \quad & \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} \epsilon_i + \mu \sum_{i=1}^{n} (\xi_i + \xi_i^*) \\ \text{subject to } & \begin{cases} w^T(\varphi(x_i) - \varphi(x_{\bar{j}(i)})) \geq y_i - y_{\bar{j}(i)} - \epsilon_i, & \forall\, i = 1, \ldots, n \\ w^T \varphi(x_i) + b \geq y_i - \xi_i, & \forall\, i = 1, \ldots, n \\ -\delta_i (w^T \varphi(x_i) + b) \geq -\delta_i y_i - \xi_i^*, & \forall\, i = 1, \ldots, n \\ \epsilon_i \geq 0, & \forall\, i = 1, \ldots, n \\ \xi_i \geq 0, & \forall\, i = 1, \ldots, n \\ \xi_i^* \geq 0, & \forall\, i = 1, \ldots, n\,. \end{cases} \end{aligned} \tag{23}$$

The first set of constraints is the same as in MODEL 1 and is the ranking constraint optimizing the concordance index. The second set of constraints indicates that the prognostic index should be larger than the observed failure time. The third set indicates that the prognostic index u should be smaller than y. Multiplication with the censoring indicator δ_i makes sure that this constraint is not taken into account for right censored data. Taking the second and third set of constraints together, the value of the prognostic index u is targeted at the failure

time. However, for right censored data, the third constraint doesn't need to be fulfilled. The prognostic index for a new observation $x^\star$ is then found as

$$\hat{u}(x^\star) = \sum_{i=1}^{n} \left(\alpha_i(\varphi(x_i) - \varphi(x_{\bar{j}(i)}))^T + (\beta_i - \delta_i\beta_i^*)\varphi(x_i)^T\right)\varphi(x^\star) + b\,, \qquad (24)$$

with α_i, β_i and β_i^* the Lagrange multipliers. See Appendix B for more information.

Both MODEL 1 and MODEL 2 restrict the number of comparisons or ranking constraints to $\mathcal{O}(n)$ by comparing each observation with its closest (in observed time) comparable point. Since this choice is rather arbitrary, we question whether the performance of the model can be improved by generating k dummy observations with unknown utility v_l (also called threshold values) and known failure time z_l, with $l = 1, \ldots, k$. The z-values are spread uniformly over the observed time interval. The minimum z value equals zero since this is the lowest possible outcome, the largest z is higher than the highest observed outcome. All survival and censoring times lie therefore within the interval $[z_1, z_k]$. The inclusion of the thresholds is an elegant way to make an unarbitrary choice of comparisons. The difference between this approach and the approach of MODEL 2 is illustrated in Figure 3). Assume no censoring for now. In MODEL 2 a comparison will be made between each observation i and the observation with the highest failure time for which this failure time is smaller than y_i. In MODEL 3, these comparisons are made less arbitrary by comparing each observation with a dummy observation.

Modifying MODEL 2 by including these thresholds leads to the following problem formulation

MODEL 3

$$\min_{w,\epsilon,\epsilon^*,\xi,\xi^*,b,v} \frac{1}{2}w^T w + \gamma\sum_{i=1}^{n}(\epsilon_i + \epsilon_i^*) + \mu\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

$$\text{subject to } \begin{cases} w^T\varphi(x_i) + b - \ell_i^T v \geq y_i - \ell_i^T z - \epsilon_i, & \forall\, i = 1, \ldots, n \\ -\delta_i(w^T\varphi(x_i) + b - r_i^T(v - z) - y_i) \geq -\epsilon_i^*, & \forall\, i = 1, \ldots, n \\ w^T\varphi(x_i) + b \geq y_i - \xi_i, & \forall\, i = 1, \ldots, n \\ -\delta_i(w^T\varphi(x_i) + b) \geq -\delta_i y_i - \xi_i^*, & \forall\, i = 1, \ldots, n \\ \epsilon_i \geq 0, & \forall\, i = 1, \ldots, n \\ \epsilon_i^* \geq 0, & \forall\, i = 1, \ldots, n \\ \xi_i \geq 0, & \forall\, i = 1, \ldots, n \\ \xi_i^* \geq 0 & \forall\, i = 1, \ldots, n \\ v_l - v_{l-1} \geq 0, & \forall\, l = 1, \ldots, k\,. \end{cases} \qquad (25)$$

In the above problem formulation ℓ_i and r_i represent vectors of length k containing all zero values, except for the value $\bar{l} = \arg\max_l l$, subject to $z_l < y_i$ for which $\ell_i(\bar{l})$ equals 1 and $\bar{l} = \arg\min_l l$ subject to $z_l > y_i$ for which $r_i(\bar{l})$ equals 1. The first and second set of constraints are similar to the ones in equation (23), replacing the arbitrary comparisons with the nearest neighbor by comparisons with the nearest thresholds. The last set of equations makes sure that the

thresholds are increasing. The prognostic index $u(x^\star)$ for a new observation $x^\star$ can then be found as

$$\hat{u}(x^\star) = \sum_{i=1}^{n} (\alpha_i + \beta_i - \delta_i(\alpha_i^* + \beta_i^*))\varphi(x_i)^T \varphi(x^\star) + b\,, \tag{26}$$

where $\alpha_i, \alpha_i^*, \beta_i$ and β_i^* are the Lagrange multipliers. We refer to Appendix C for more details.

4 Data and performance measures

This Section describes the publicly available datasets which are used in the experiments of this paper. In addition, three different performance measures are discussed.

4.1 Datasets

We will use 5 publicly available datasets for 6 clinical experiments. The first dataset concerns the prediction of complete remission and death of 129 patients with leukemia [21]. Information on the treatment, sex, age, performance score (Karnofsky score), white blood cells, platelets, hemoglobin and whether the patient received a kidney transplant was given. In a first experiment (LD) the event at study is death, while in a second experiment (LCR) the event is complete remission. A second dataset concerns the veteran's administration lung cancer trial (VLC) [22,6]. Only 9 out of the 137 men with advanced inoperable lung cancer were alive at the end of the study. Patients were randomized into a standard or test chemotherapy. Information on the performance score, age, therapy, histology of the tumor, treatment and time between the diagnosis and the randomization was available. The third dataset concerns 506 patients with prostatic cancer (PC) [23]. The variables used for this experiment are: treatment, age, weight index, performance index, history of cardiovascular disease, size of the tumor, a combined index of stage of histologic grade and serum hemoglobin. The analysis was performed on the 483 patients with complete information. 125 (26%) patients died due to prostatic cancer during the study. All other patients were right censored at their date of last follow-up. The Mayo clinic lung cancer data (MLC) [24] contains information on 167 patients with advanced lung cancer. The dataset contains information on age, sex, two performance scores estimated by the physician and one estimated by the patient, the number of calories consumed and the weight loss in the last six months. During the study period, 72% of the patients died. A last dataset contains information on 720 breast cancer patients (the German breast cancer study) (GBSG) [25,26]. Information was available on hormonal therapy, menopausal status, age, grade, tumor size, number of positive lymph nodes, progesterone and estrogene receptors. The experiment is based on the 686 cases with complete information. During the study period, 299 patients experienced a breast cancer related event. All discussed datasets

can be found on the world-wide-web (`http://lib.stat.cmu.edu/datasets` (accessed: 29 August 2008) and the R-package `http://cran.r-project.org/web/packages/survival/index.html` (accessed: 20/11/2007)).

4.2 Performance measures

In order to compare the different models, three performance measures are used in the experiments:

Concordance index (c-index) The concordance index is the ratio of the number of concordant pairs and the number of comparable pairs:

$$\mathrm{c-index}(u) = \frac{\sum_{i=1}^{n} \sum_{j \neq i} \mathcal{I}\left[(u(x_i) - u(x_j))(y_i - y_j) \geq 0\right]}{\sum_{i=1}^{n} \sum_{j \neq i} \mathrm{comp}(i,j)}\,, \tag{27}$$

where $\mathcal{I}$ is the indicator function and $\mathrm{comp}(i,j)$ is defined as before.

Logrank χ^2 statistic Clinicians are often interested in the allocation of risk profiles to patients. In this context it is relevant to test how well the developed prognostic index is able to differentiate high and low risk patients. The patients are therefore divided into a low and high risk group according to whether their prognostic index is below or above the median value. The logrank test is used to test significant differences in survival between both groups. Instead of working with the resulting p-value, we will report on the χ^2-value of the test. The higher this value, the better the separating ability of the model.

Hazard rate estimated by a univariate COX model. All produced prognostic indices, after normalization to restrict its value between 0 and 1, will be fed to the COX model and the resulting hazard rate will be calculated. The higher the hazard rate, the better the index.

4.3 Experiments

This subsection compares models described in the literature (COX model, RANKSVMC, SVCR, MODEL 1) with the models proposed in this paper (MODEL 2 and MODEL 3) on 6 clinical datasets and 1 artificial experiment. 10-fold cross-validation with coupled simulated annealing [27] was used to tune the hyper-parameters. The concordance index was used as model selection criterion. All models were implemented in matlab[3], using the mosek[4] optimization toolbox for matlab.

Real data The two proposed models (MODEL 2 and MODEL 3) are compared with MODEL 1, RANKSVMC and SVCR on 50 randomizations between training (two thirds of the data) and test set (remaining third of the data) on 6 clinical

[3] http://www.mathworks.com/products/matlab/
[4] http://www.mosek.com/

datasets, using a linear and a clinical [19] kernel. Tables 2 and 5 summarizes the results. Statistically significant differences between MODEL 2 and the other models were calculated using the Wilcoxon rank sum test. The COX model is tested linearly in the covariates (PH_{linear}) and with penalized smoothing splines ($\text{PH}_{pspline}$) [15], for comparison. It is seen that both MODEL 1 and RANKSVMC are performing significantly less than the other models, independently of the used kernel. Addition of the regression constraints improves the performance significantly. However, the number of support vectors increases while doing so. No significant differences are seen between the arbitrary choice of comparisons and the threshold approach.

To investigate the difference in performance between models with and without regression constraints, 1000 bootstrap samples from one training set of the VLC dataset were taken. Using a linear kernel, the estimated weights for each covariate and the resulting concordance index on the test set were calculated. Figures 4 and 5 show the distribution of the estimated weights of the Karnofsky performance score and the concordance index, respectively. The model only taking ranking constraints into account results in high uncertainties about the weight estimates and therefore high differences in concordance index occur for slightly different training sets. Inclusion of regression constraints results in more stable estimates of weights and concordance index.

Artificial data In this last example, we investigate whether the improvement of MODEL 2 over MODEL 1 is dependent on the censoring percentage of the problem. In this experiment, the clinical setting is mimicked. Therefore, 100 datasets are constructed with 10 variables, 100 training and 500 test observations. All variables were normally distributed with zero mean and variance equal to 1. The correlations between the variables were zero except for the first and second covariate, the third and fourth, the fifth and sixth, and the seventh and eight, having a correlation coefficient of 0.7, 0.3, -0.7 and -0.3, respectively. The vector of regression coefficients w was taken to be $w = [-0.1, 1, 0.3, 0.4, 0.2, 0.03, 0.02, -0.01, -0.02, 0.01]^T$. The failure time was exponentially distributed with parameter equal to $w^T x$, with x the covariate vector. In a first example, the failure time was randomly censored (mimicking administrative censoring); in a second example the censoring time was sampled from the same distribution as the survival time $F_Y(x)$, the spread of the distribution was changed to create different censoring percentages ($G_Y(x) = aF_Y(x)$). Figure 6 illustrates the results. For both the random and informative censoring, the largest performance differences between both models are noted for low censoring percentages. However, MODEL 2 performs better than the MODEL 1 for all censoring percentages.

5 Conclusions

This work compared different support vector methods for survival analysis. Two questions were raised: (i) are ranking constraints sufficient to obtain good prog-

nostic indices or is the inclusion of extra regression constraints significantly improving the model's performance? and (ii) Can the estimation of the prognostic index be improved by pinning the comparisons to dummy observations instead of arbitrary choosing real observations to compare with. The experiments revealed a significant improvement when including regression constraints in addition to ranking constraints. Finally, the inclusion of the thresholds does not improve the performance. Since a model with comparisons with the closest neighbor is simpler, we prefer to use the latter model. Comparison of MODEL 2 with the COX model revealed no significant differences in performance. The advantage of MODEL 2 above COX model lies in the easy extension towards non-linear models without the need to check non-linearities in the variables before modelling.

Acknowledgments

This research is supported by Research Council KUL: GOA AMBioRICS, GOA MANET, CoE EF/05/006, IDO 05/010, IOF KP06/11, IOF SCORES4CHEM, several PhD, postdoc and fellow grants; Flemish Government: FWO: PhD and postdoc grants, G.0407.02, G.0360.05, G.0519.06, G.0321.06, G.0341.07 and projects G.0452.04, G.0499.04, G.0211.05, G.0226.06, G.0302.07; IWT: PhD Grants, McKnow-E, Eureka-Flite; Belgian Federal Science Policy Office: IUAP P6/04; EU: FP6-2002 LIFESCIHEALTH 503094, IST 2004-27214, FP6-MC-RTN 035801; Prodex-8 C90242; EU: ERNSI. V. Van Belle is supported by a grant from the IWT. Kristiaan Pelckmans is an associate professor ('Forskarassistent') at the University of Uppsala, Sweden. S. Van Huffel is a full professor and J.A.K. Suykens is a professor at the Katholieke Universiteit Leuven, Belgium.

References

1. Kalbfleisch J.D., Likelihood methods and nonparametric tests, Journal of the American Statistical Association 73 (361) (1978) 167–170.
2. Dabrowska D.M., Doksum K.A., Partial likelihood in transformation models with censored data, Scandinavian Journal of Statistics 15 (1) (1988) 1–23.
3. Doksum K., Gasko M., On a correspondence between models in binary regression and survival analysis., International Statistical Review 58 (1990) 243–252.
4. Cheng S.C., Wei L.J., Ying Z., Analysis of transformation models with censored data, Biometrika 82 (4) (1995) 835–845.
5. Cheng S.C., Wei L.J., Ying Z., Predicting Survival Probabilities with Semiparametric Transformation Models., Journal of the American Statistical Association 92 (437) (1997) 227–235.
6. Kalbfleisch J.D., Prentice R.L., The Statistical Analysis of Failure Time Data., Wiley series in probability and statistics, New York., 2002.
7. Cox D.R., Regression models and life-tables (with discussion), Journal of the Royal Statistical Society, Series B 34 (2) (1972) 187–220.
8. Van Belle V., Pelckmans K., Suykens J.A.K., Van Huffel S., Learning Transformation Models for Ranking and Survival Analysis, Tech. rep., 09-45, ESAT-SISTA, K.U.Leuven (Leuven, Belgium) (2009).

9. Van Belle V., Pelckmans K., Suykens J.A.K., Van Huffel S., Support Vector Machines for Survival Analysis., in: E. Ifeachor, A. Anastasiou (Eds.), Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED), 2007.
10. Van Belle V., Pelckmans K., Suykens J.A.K., Van Huffel S., Survival SVM: a Practical Scalable Algorithm., in: M. Verleysen (Ed.), Proceedings of the 16th European Symposium on Artificial Neural Networks (ESANN), d-side, Evere, 2008, pp. 89–94.
11. Evers L., Messow C.M., Sparse kernel methods for high-dimensional survival data, Bioinformatics 24 (14) (2008) 1632–1638.
12. Shivaswamy P.K., Chu W., Jansche M., A support vector approach to censored targets, in: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining (ICDM), IEEE Computer Society, California, 2007, pp. 655–660.
13. Khan F.M., Zubek V.B., Support vector regression for censored data (SVRc): a novel tool for survival analysis, in: F. Giannotti, D. Gunopulos, F. Turini, C. Zaniolo, N. Ramakrishnan, X. Wu (Eds.), Proceedings of the Eighth IEEE International Conference on Data Mining (ICDM), IEEE computer society, California, 2008.
14. Lawless J.F., Survival and event history analysis, Wiley reference series in biostatistics, Wiley, West Sussex, England, 2006, Ch. Parametric models in survival analysis, pp. 345–355.
15. Eilers P.H., Marx B.D., Flexible smoothing with B-splines and penalties, Statistical Science 11 (1996) 89–121.
16. Hurvich C.M., Simonoff J.S., Tsai C.L., Smoothing parameter selection in nonparametric regression using an improved akaike information criterion, Journal of the Royal Statistical Society: Series B 60 (1998) 271–293.
17. Vapnik V., Statistical Learning Theory., Wiley and Sons, New York, 1998.
18. M. J., Functions of positive and negative type and their connection with the theory of integral equations, Philosophical Transactions of the Royal Society A 209 (1909) 415–446.
19. Daemen A., De Moor B., Development of a kernel function for clinical data, in: Proceedings of the 31th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), IEEE, Piscataway, 2009, pp. 5913–5917.
20. Herbrich R., Graepel T., Obermayer K., Large margin rank boundaries for ordinal regression, Advances in Large Margin Classifiers (2000) 115–132.
21. Emerson S.S., Banks P.L.C., Case Studies in Biometry, Wiley-Interscience, New York, 1994, Ch. Interpretation of a leukemia trial stopped early, pp. 275–299.
22. Prentice R.L., A log gamma model and its maximum likelihood estimation, Biometrika 61 (3) (1974) 539–544.
23. Byar D., Green S., Prognostic variables for survival in a randomized comparison of treatments for prostatic cancer, Bulletin du Cancer 67 (1980) 477–490.
24. Therneau T.M. and Grambsch P.M., Modeling Survival Data: Extending the Cox Model, 2nd Edition, Springer-Verlag, New York, 2000.
25. Schumacher M., Basert G., Bojar H., Huebner K., Olschewski M., Sauerbrei W., Schmoor C., Beyerle C., Neumann R.L.A., Rauschecker H.F., Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients, Journal of Clinical Oncology 12.
26. Sauerbrei W., Royston P., Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials, Journal of the Royal Statistical Society Series A 162 (1) (1999) 71–94.

27. Xavier de Souza S., Suykens J.A.K., Vandewalle J., Bolle D., Coupled simulated annealing, IEEE Transactions on Systems, Man, and Cybernetics - Part B 40 (2) 320–335.
28. Boyd S., Vandenberghe L., Convex optimization., Cambridge University Press, Cambridge, 2004.

Table 1. Overview of different kernel-based survival methods and their properties, in case a linear kernel is used.

model	reference	ranking constraint	regression constraint	# tuning parameters	ranking constraints
RANKSVMC	[11,9,10]	✓		1	nearest neighbor
SVCR	[12]		✓	1	
SVRC	[13]		✓	4	
MODEL 1	[8]	✓		1	nearest neighbor
MODEL 2		✓	✓	2	nearest neighbor
MODEL 3		✓	✓	2	dummy observations

Table 2. Comparison of the different survival SVM models and the COX model on 6 cancer datasets with clinical variables, using a linear kernel. The median performance on 50 randomizations between training and test set, together with the number of support vectors (# sv) are summarized. Statistical significant differences between MODEL 2 (indicated in grey) and the other models was tested using the Wilcoxon rank sum test. Use of the regression constraints improves the performance significantly. No significant differences are noted with the COX model. However, the SVM-based models have the advantage that one does not have to check the linearity and proportional hazard assumption.

dataset	method	c-index	logrank χ^2	hazard rate	# sv
VLC	MODEL 1	0.61±0.07[3]	1.38±2.19[3]	3.97±4.29[3]	84.5±7.75[3]
	MODEL 2	**0.69±0.03**	5.27±5.35	10.50±13.19	91.0±0.70
	MODEL 3	**0.69±0.03**	4.73±4.50	9.00±6.28	91.0±12.84
	RANKSVMC	0.62±0.08[3]	3.02±3.19[3]	5.03±6.53[3]	**78.0±6.44**[3]
	SVCR	**0.69±0.04**	**5.75±4.37**	9.50±9.93	89.0±1.21[3]
	PH_{linear}	0.68±0.03	5.45±3.66	**10.61±6.47**	
LCR	MODEL 1	0.55±0.07[3]	1.08±2.23[1]	2.4±5.36[3]	**56.0±9.97**[3]
	MODEL 2	**0.60±0.05**	1.87±2.51	**5.06±9.65**	86.0±0.93
	MODEL 3	**0.60±0.05**	**2.04±2.18**	4.77±3.61	84.5±12.14
	RANKSVMC	0.51±0.07[3]	0.51±2.81[3]	1.43±5.60[3]	72.5±6.55[3]
	SVCR	0.59±0.04	2.02±2.73	4.46±2.90	80.0±3.25[3]
	PH_{linear}	**0.60±0.04**	1.60±2.06	4.05±4.80	
LD	MODEL 1	0.62±0.07[3]	2.19±2.92[3]	4.45±6.53[3]	83.5±8.35[3]
	MODEL 2	0.68±0.04	7.53±5.18	13.8±15.21	86.0±2.21
	MODEL 3	0.63±0.10	3.07±4.16	7.68±11.07	86.0±12.24
	RANKSVMC	0.62±0.08[3]	2.55±3.57[3]	4.68±15.96[3]	**80.0±4.24**[3]
	SVCR	**0.69±0.05**	6.70±3.95	14.21±12.7	81.0±5.10[3]
	PH_{linear}	**0.69±0.05**[1]	**9.73±8.31**[2]	**22.21±32.09**[1]	
MLC	MODEL 1	0.6±0.05[2]	2.27±3.77	3.22±3.43	**88.5±14.22**[3]
	MODEL 2	**0.62±0.05**	**3.32±3.41**	4.12±5.44	110.5±5.88
	MODEL 3	**0.62±0.04**	2.72±3.09	3.64±4.66	111.0±7.69
	RANKSVMC	0.59±0.05[1]	3.06±2.91	3.20±6.79	99.0±4.45[3]
	SVCR	0.61±0.04	2.27±2.86	3.48±4.77	95.0±6.51[3]
	PH_{linear}	0.61±0.04[1]	2.89±2.68	**6.04±5.55**	
PC	MODEL 1	0.75±0.04	10.05±5.94	148.86±340.89	**64.0±73.74**[3]
	MODEL 2	0.76±0.03	11.86±5.69	231.79±296.90	322.0±4.95
	MODEL 3	**0.77±0.03**	**12.90±6.07**	**269.10±474.44**	322.0±4.48
	RANKSVMC	0.75±0.03	12.56±5.37	166.10±340.56	255.5±13.54[3]
	SVCR	0.76±0.03	10.15±5.64	198.45±479.37	137.5±28.72[2]
	PH_{linear}	**0.77±0.03**	11.39±5.68	266.26±478.54	
BC	MODEL 1	0.64±0.04[3]	12.03±9.57[1]	36.02±1657.30[2]	399.5±30.67[3]
	MODEL 2	**0.67±0.03**	16.45±8.06	70.14±2876.42	457.0±12.46
	MODEL 3	**0.67±0.03**	17.74±7.00	**112.65±4499.18**	457.0±40.06[2]
	RANKSVMC	0.62±0.10[3]	9.99±9.24[3]	21.36±306.01[3]	399.5±59.32[3]
	SVCR	**0.67±0.10**	**20.17±8.74**	67.29±861.97	**366.0±72.09**[3]
	PH_{linear}	**0.67±0.03**	18.11±8.76	102.78±3182.89	

[1], $p < 0.05$;[2], $p < 0.01$;[3], $p < 0.001$ (Wilcoxon rank sum test)

Table 3. Comparison of the different survival SVM models on 6 cancer datasets with clinical variables, using a clinical kernel. The median performance on 50 randomizations between training and test set, together with the number of support vectors (# sv) are summarized. Statistical significant differences between MODEL 2 (indicated in grey) and the other models was tested using the Wilcoxon rank sum test. Use of the regression constraints improves the performance significantly. No significant differences are noted with the COX model. However, the SVM-based models have the advantage that one does not have to check the linearity and proportional hazard assumption.

dataset	method	c-index	logrank χ^2	hazard rate	# sv
VLC	MODEL 1	0.58±0.06[3]	2.01±3.72[3]	3.26±4.70[3]	91.0±3.51
	MODEL 2	**0.69±0.03**	6.27±5.11	9.78±7.46	91.0±1.10
	MODEL 3	**0.69±0.03**	6.85±5.00	9.85±7.64	90.5±1.16
	RANKSVMC	0.57±0.07[3]	1.60±3.83[3]	2.70±5.36[3]	**77.0±6.53**[3]
	SVCR	**0.69±0.04**	**9.88±6.05**[1]	**11.12±9.75**	**90.0±0.96**[3]
	$PH_{pspline}$	0.67±0.09[2]	4.52±6.62	8.40±19.14	
LCR	MODEL 1	0.54±0.06[3]	1.04±1.63	1.98±2.85[2]	80.0±8.24[3]
	MODEL 2	0.59±0.04	1.08±1.90	3.56±3.50	83.0±4.72
	MODEL 3	0.58±0.05	1.17±2.08	2.56±3.43	**76.5±6.05**[3]
	RANKSVMC	0.52±0.06[3]	0.37±2.17	1.38±2.88[3]	78.0±7.22[3]
	SVCR	**0.60±0.04**	**1.94±1.72**[1]	**4.04±3.25**	80.5±3.67[3]
	$PH_{pspline}$	0.58±0.05	1.39±1.98	3.70±5.83	
LD	MODEL 1	0.66±0.07[2]	3.60±4.35[3]	5.97±12.37[3]	86.0±3.68
	MODEL 2	**0.69±0.05**	8.50±7.48	17.23±23.33	86.0±2.03
	MODEL 3	**0.69±0.05**	6.05±6.95	17.47±21.59	86.0±2.24
	RANKSVMC	0.65±0.08[3]	2.68±6.60[3]	5.12±8.95[3]	**76.0±5.04**[3]
	SVCR	**0.69±0.04**	**10.7±6.51**	13.1±33.54[3]	**82.0±4.13**[3]
	$PH_{pspline}$	0.67±0.05	4.89±5.45[1]	**18.35±62.26**	
MLC	MODEL 1	0.61±0.05	3.13±3.74	3.48±7.48	106.5±6.25
	MODEL 2	0.61±0.05	2.53±2.76	3.28±3.83	111.0±8.55
	MODEL 3	0.61±0.05	2.77±2.55	2.96±3.44	111.0±9.75
	RANKSVMC	0.61±0.05	2.55±3.70	3.30±7.42	**95.0±4.90**[3]
	SVCR	**0.62±0.04**	**3.70±3.22**[1]	**4.58±5.30**[1]	96.0±7.84[3]
	$PH_{pspline}$	0.57±0.04[3]	2.04±2.27	3.08±3.46	
PC	MODEL 1	0.76±0.04[2]	11.22±5.71[1]	108.39±107.79[1]	**88.5±28.21**[3]
	MODEL 2	**0.78±0.03**	**14.56±6.47**	148.08±137.34	112.0±29.54
	MODEL 3	**0.78±0.03**	14.52±6.86	144.72±163.35	89.5±8.47[3]
	RANKSVMC	0.76±0.03	11.89±6.00	111.26±116.90	264.0±11.38
	SVCR	**0.78±0.03**	13.15±6.60	150.66±165.24	160.0±62.93[1]
	$PH_{pspline}$	0.77±0.03	12.76±5.89	**286.6±615.05**[3]	
BC	MODEL 1	0.63±0.04[3]	9.34±8.83[3]	10.10±20.91[3]	443.0±21.24[3]
	MODEL 2	**0.68±0.03**	**22.62±9.09**	24.23±13.19	457.0±5.60
	MODEL 3	**0.68±0.03**	22.42±8.51	24.20±13.30	456.5±22.93[3]
	RANKSVMC	0.62±0.10[3]	8.93±8.83[3]	11.20±15.23[3]	**404.0±58.26**[3]
	SVCR	**0.68±0.10**	19.51±7.89[1]	18.99±12.02	453.0±75.70[3]
	$PH_{pspline}$	0.67±0.03	17.36±8.24[1]	**98.93±4052.05**[3]	

[1], $p < 0.05$;[2], $p < 0.01$;[3], $p < 0.001$ (Wilcoxon rank sum test)

Figure captions

Figure 1 Loss functions as defined by SVCR (a-b) [12] and SVRc (c-d) [13]. The loss functions for events are denoted in (a) and (c). The loss functions for right censored data points are illustrated in (b) and (d). Both methods differ in the way the loss function is calculated. A major disadvantage of the SVRc method is the necessity to define 4 hyperparameters.

Figure 2 Illustration of the search for a prognostic index which minimizes the empirical risk of misranking two observations. The upper part shows the ranking of 7 observations according to the survival time and according to two prognostic models $u(x)$. The lower part shows the calculation of the concordance index. Assuming that all survival times are known exactly, all pairs can be compared. A pair is concordant for a certain model whenever the ranking in time equals the ranking in model (no crosses in the upper part of the Figure). (a) The first model results in a prognostic index which misranks 3 out of 21 possible pairs of datapoint. (b) The second model on the contrary misranks 10 pairs, resulting in a much lower concordance index.

Figure 3 Illustration of the difference between MODEL 2 (a) and MODEL 3 (b). In MODEL 2 each observation i is compared (indicated by the arrows) with one observation with a failure time the closest to, but smaller than, the failure time of observation i (y_i). Since this is a rather arbitrary choice, MODEL 3 includes dummy observations with failure times distributed equidistantly over the time axis (indicated by the dashed vertical lines). Each observation i is then compared with 2 dummy observations: (i) the dummy observation with the highest failure time lower than y_i (indicated by the arrows); and (ii) the dummy observation with the lowest failure time higher than y_i (not shown).

Figure 4 Distribution of the estimated weights (using a linear kernel) of the Karnofsky score on 1000 bootstraps of one particular training set of the VLC dataset: (a) MODEL 1, (b) MODEL 2. MODEL 1 has a wide distribution without a clear estimate (estimates with different signs). Addition of the regression constraints in MODEL 2 results in an undoubted positive relation between the Karnofsky score and the survival time (weight>0).

Figure 5 Distribution of the concordance index on test on 1000 bootstraps of one particular training set of the VLC dataset: (a) MODEL 1, (b) MODEL 2. MODEL 1 has a wide distribution from which one can not decide whether the estimated prognostic index is related or inversely related to the failure time. Addition of the regression constraints in MODEL 2 leads to a narrowing of this distribution. This results in an clear positive relation between the prognostic index and the survival time (c-index>0.5)

Figure 6 Comparison between models MODEL 1 (solid line) and MODEL 2 (dashed line) for random (a) and informative (b) censoring in a 10-dimensional artificial example. All variables were normally distributed with zero mean and variance equal to 1. All variables were uncorrelated except for 4 pairs of variables with correlations of 0.7, 0.3, -0.7 and -0.3. The fixed coefficient vector w was taken. The failure time was sampled from an exponential dis-

tribution with a parameter equal to $w^T x$, x being the variable vector. The percentage of censored observations was varied from 0 to 100%. The largest performance differences are noted for low censoring percentages. However, MODEL 2 seems to perform better for all censoring percentages, both for random and informative censoring.

Appendix A. MODEL 1

The methods proposed in this paper are based on the survival model proposed in [8]. We give the deviation of the model in this first appendix. The survival model was written as a ranking problem with arbitrary ranking constraints as

$$\boxed{\text{MODEL 1}}$$
$$\begin{aligned} \min_{w,\epsilon} \quad & \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} \epsilon_i \\ \text{subject to } & \begin{cases} w^T(\varphi(x_i) - \varphi(x_{\bar{j}(i)})) \geq y_i - y_{\bar{j}(i)} - \epsilon_i, \forall\, i = 1, \ldots, n \\ \epsilon_i \geq 0, \qquad\qquad \forall\, i = 1, \ldots, n\,. \end{cases} \end{aligned} \tag{28}$$

Formulating the Lagrangian of (28)

$$\begin{aligned} \mathcal{L}(w, \epsilon; \alpha, \beta) = \frac{1}{2} w^T w + \gamma \sum_{i=1}^{n} \epsilon_i - \sum_{i=1}^{n} \alpha_i w^T(\varphi(x_i) - \varphi(x_{\bar{j}(i)})) \\ - \sum_{i=1}^{n} \alpha_i \left(-y_i + y_{\bar{j}(i)} + \epsilon_i\right) - \sum_{i=1}^{n} \beta_i \epsilon_i\,, \end{aligned} \tag{29}$$

and solving the Karush-Kuhn-Tucker (KKT) [28] equations

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{n} \alpha_i \left(\varphi(x_i) - \varphi(x_{\bar{j}(i)})\right) & \\ \frac{\partial \mathcal{L}}{\partial \epsilon_i} = 0 \rightarrow \gamma = \alpha_i + \beta_i & \forall\, i = 1, \ldots, n \\ \alpha_i \geq 0 & \forall\, i = 1, \ldots, n \\ \beta_i \geq 0 & \forall\, i = 1, \ldots, n \\ \sum_{i=1}^{n} \alpha_i \left(w^T(\varphi(x_i) - \varphi(x_{\bar{j}(i)})) - y_i + y_{\bar{j}(i)} + \epsilon_i\right) = 0 & \\ \sum_{i=1}^{n} \beta_i \epsilon_i = 0\,, & \end{cases} \tag{30}$$

leads, after elimination of the primal variables and β_i to the dual

$$\begin{aligned} \min_{\alpha_i} \quad & \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k (\varphi(x_i) - \varphi(x_{\bar{j}(i)}))^T (\varphi(x_k) - \varphi(x_{\bar{l}(k)})) \\ & - \sum_{i=1}^{n} \alpha_i (y_i - y_{\bar{j}(i)}) \\ \text{subject to } & 0 \geq \alpha_i \geq \gamma, \forall\, i = 1, \ldots, n\,. \end{aligned} \tag{31}$$

The prognostic index $u(x^\star)$ for a new observation $x^\star$ can then be found as

$$\hat{u}(x^\star) = \sum_{i=1}^{n} \alpha_i \left(\varphi(x_i) - \varphi(x_{\bar{j}(i)})\right)^T \varphi(x^\star)\,. \tag{32}$$

Appendix B. MODEL 2

This and the following Appendix contain the derivations from the newly proposed models. For every model, we start from the model formulation as given in

Section 3, describe the Lagrangian and derive the KKT conditions for optimality. MODEL 2 was defined as

MODEL 2

$$\min_{w,\epsilon,\xi,\xi^*,b} \quad \frac{1}{2}w^T w + \gamma \sum_{i=1}^{n} \epsilon_i + \mu \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

$$\text{subject to } \begin{cases} w^T(\varphi(x_i) - \varphi(x_{\bar{j}(i)})) \geq y_i - y_{\bar{j}(i)} - \epsilon_i, & \forall\ i = 1, \ldots, n \\ w^T\varphi(x_i) + b \geq y_i - \xi_i, & \forall\ i = 1, \ldots, n \\ -\delta_i(w^T\varphi(x_i) + b) \geq -\delta_i y_i - \xi_i^*, & \forall\ i = 1, \ldots, n \\ \epsilon_i \geq 0, & \forall\ i = 1, \ldots, n \\ \xi_i \geq 0, & \forall\ i = 1, \ldots, n \\ \xi_i^* \geq 0, & \forall\ i = 1, \ldots, n\,. \end{cases} \tag{33}$$

The Lagrangian becomes

$$\begin{aligned} \mathcal{L}(w, \epsilon, b; \alpha, \beta) = {} & \frac{1}{2}w^T w + \gamma \sum_{i=1}^{n} \epsilon_i + \mu \sum_{i=1}^{n} (\xi_i + \xi_i^*) \\ & - \sum_{i=1}^{n} \alpha_i \left(w^T(\varphi(x_i) - \varphi(x_{\bar{j}(i)})) - y_i + y_{\bar{j}(i)} + \epsilon_i \right) \\ & - \sum_{i=1}^{n} \beta_i \left(w^T\varphi(x_i) + b - y_i + \xi_i \right) \\ & - \sum_{i=1}^{n} \beta_i^* \left(-\delta_i(w^T\varphi(x_i) + b - y_i) + \xi_i^* \right) \\ & - \sum_{i=1}^{n} \eta_i \epsilon_i - \sum_{i=1}^{n} \nu_i \xi_i - \sum_{i=1}^{n} \nu_i^* \xi_i^*\,. \end{aligned} \tag{34}$$

The KKT conditions for optimality are

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{n} \alpha_i \left(\varphi(x_i) - \varphi(x_{\bar{j}(i)}) \right) + \sum_{i=1}^{n} (\beta_i - \delta_i \beta_i^*) \varphi(x_i) \\ \frac{\partial \mathcal{L}}{\partial b} = 0 \rightarrow \sum_{i=1}^{n} (-\beta_i + \delta_i \beta_i^*) = 0 \\ \frac{\partial \mathcal{L}}{\partial \epsilon_i} = 0 \rightarrow \gamma = \alpha_i + \eta_i & \forall\ i = 1, \ldots, n \\ \frac{\partial \mathcal{L}}{\partial \xi_i} = 0 \rightarrow \mu = \beta_i + \nu_i & \forall\ i = 1, \ldots, n \\ \frac{\partial \mathcal{L}}{\partial \epsilon_i} = 0 \rightarrow \gamma = \beta_i^* + \nu_i^* & \forall\ i = 1, \ldots, n \\ \alpha_i, \beta_i, \beta_i^*, \eta_i, \xi_i, \xi_i^* \geq 0 & \forall\ i = 1, \ldots, n \\ \sum_{i=1}^{n} \alpha_i \left(w^T(\varphi(x_i) - \varphi(x_{\bar{j}(i)})) - y_i + y_{\bar{j}(i)} + \epsilon_i \right) = 0 \\ \sum_{i=1}^{n} \beta_i \epsilon_i = 0 \\ \sum_{i=1}^{n} \beta_i \left(w^T\varphi(x_i) + b - y_i + \xi_i \right) = 0 \\ \sum_{i=1}^{n} \beta_i^* \left(-\delta_i(w^T\varphi(x_i) + b - y_i) + \xi_i^* \right) = 0 \\ \sum_{i=1}^{n} \eta_i \epsilon_i = 0 \\ \sum_{i=1}^{n} \nu_i \xi_i = 0 \\ \sum_{i=1}^{n} \nu_i^* \xi_i^* = 0\,. \end{cases} \tag{35}$$

After elimination of all unknowns except α_i, β_i and β_i^* the solution is found as

$$\begin{aligned}\min_{\alpha_i,\beta,\beta^*} \quad & \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n}\alpha_i\alpha_k(\varphi(x_i)-\varphi(x_{\bar{j}(i)}))^T(\varphi(x_k)-\varphi(x_{\bar{l}(k)})) \\ & +\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n}\left(\beta_i\beta_k+\delta_i\delta_k\beta_i^*\beta_k^*\right)\varphi(x_i)^T\varphi(x_k)-\sum_{i=1}\alpha_i(y_i-y_{i-1}) \\ & +\sum_{i=1}^{n}\sum_{k=1}^{n}\left(\beta_i-\delta_k\beta_k^*\right)\alpha_k\varphi(x_i)^T(\varphi(x_k)-\varphi(x_{\bar{l}(k)})) \\ & -\sum_{i=1}^{n}\sum_{k=1}^{n}\beta_i^*\delta_i\beta_i\varphi(x_i)^T\varphi(x_k)-\sum_{i=1}^{n}\beta_i y_i+\sum_{i=1}^{n}\delta_i\beta_i^* y_i \\ \text{subject to} \quad & \begin{cases} 0\geq\alpha_i\geq\gamma,\ \forall\ i=1,\ldots,n \\ 0\geq\beta_i\geq\mu,\ \forall\ i=1,\ldots,n \\ 0\geq\beta_i^*\geq\mu,\ \forall\ i=1,\ldots,n\,. \end{cases}\end{aligned} \tag{36}$$

The prognostic index $u(x^\star)$ for a new observation $x^\star$ can then be found as

$$\hat{u}(x^\star)=\sum_{i=1}^{n}\left(\alpha_i(\varphi(x_i)-\varphi(x_{\bar{j}(i)}))+(\beta_i-\delta_i\beta_i^*)\varphi(x_i)\right)^T\varphi_p(x^\star)+b\,. \tag{37}$$

Appendix C. MODEL 3

The model formulation when including dummy observations to compare with within MODEL 2 was as follows:

MODEL 3

$$\begin{aligned}\min_{w,\epsilon,\epsilon^*,\xi,\xi^*,b,v} \quad & \frac{1}{2}w^Tw+\gamma\sum_{i=1}^{n}(\epsilon_i+\epsilon_i^*)+\mu\sum_{i=1}^{n}(\xi_i+\xi_i^*) \\ \text{subject to} \quad & \begin{cases} w^T\varphi(x_i)+b-\ell_i^Tv\geq y_i-\ell_i^Tz-\epsilon_i, & \forall\ i=1,\ldots,n \\ -\delta_i(w^T\varphi(x_i)+b-r_i^T(v-z)-y_i)\geq-\epsilon_i^*, & \forall\ i=1,\ldots,n \\ w^T\varphi(x_i)+b\geq y_i-\xi_i, & \forall\ i=1,\ldots,n \\ -\delta_i(w^T\varphi(x_i)+b)\geq-\delta_iy_i-\xi_i^*, & \forall\ i=1,\ldots,n \\ \epsilon_i\geq0, & \forall\ i=1,\ldots,n \\ \epsilon_i^*\geq0, & \forall\ i=1,\ldots,n \\ \xi_i\geq0, & \forall\ i=1,\ldots,n \\ \xi_i^*\geq0 & \forall\ i=1,\ldots,n \\ v_l-v_{l-1}\geq0, & \forall\ l=1,\ldots,k\,. \end{cases}\end{aligned} \tag{38}$$

The Lagrangian becomes

$$\begin{aligned}
&\mathcal{L}(w,\epsilon,\epsilon^*,\xi,\xi^*,v,b;\alpha,\alpha^*,\beta,\beta^*,\nu,\nu^*,\eta,\eta^*,\theta) = \tfrac{1}{2}w^T w + \gamma\textstyle\sum_{i=1}^n(\epsilon_i+\epsilon_i^*)\\
&+\mu\sum_{i=1}^n(\xi_i+\xi_i^*) - \sum_{i=1}^n \alpha_i\left(w^T\varphi(x_i)+b-\ell_i^T(v-z)-y_i+\epsilon_i\right)\\
&-\sum_{i=1}^n \alpha_i^*\left(-\delta_i\left(w^T\varphi(x_i)+b-r_i(v-z)-y_i\right)+\epsilon_i^*\right)\\
&-\sum_{i=1}^n \beta_i\left(w^T\varphi(x_i)+b-y_i+\xi_i\right) - \sum_{i=1}^n \beta_i^*\left(-\delta_i\left(w^T\varphi(x_i)+b-y_i\right)+\xi_i^*\right)\\
&-\sum_{l=1}^k \theta_l\left(v_l-v_{l-1}\right) - \sum_{i=1}^n \eta_i\epsilon_i - \sum_{i=1}^n \eta_i^*\epsilon_i^* - \sum_{i=1}^n \nu_i\xi_i - \sum_{i=1}^n \nu_i^*\xi_i^*\,.
\end{aligned} \tag{39}$$

The KKT conditions for optimality are

$$\begin{cases}
\frac{\partial\mathcal{L}}{\partial w} = 0 \rightarrow w = \sum_{i=1}^n(\alpha_i+\beta_i-\delta_i(\alpha_i^*+\beta_i^*))\varphi(x_i)\\
\frac{\partial\mathcal{L}}{\partial b} = 0 \rightarrow \sum_{i=1}^n(\alpha_i+\beta_i-\delta_i(\alpha_i^*+\beta_i^*))\\
\frac{\partial\mathcal{L}}{\partial\epsilon_i} = 0 \rightarrow \gamma = \alpha_i+\eta_i & \forall\ i=1,\ldots,n\\
\frac{\partial\mathcal{L}}{\partial\epsilon_i^*} = 0 \rightarrow \gamma = \alpha_i^*+\eta_i^* & \forall\ i=1,\ldots,n\\
\frac{\partial\mathcal{L}}{\partial\xi_i} = 0 \rightarrow \mu = \beta_i+\nu_i & \forall\ i=1,\ldots,n\\
\frac{\partial\mathcal{L}}{\partial\xi_i^*} = 0 \rightarrow \mu = \beta_i^*+\nu_i^* & \forall\ i=1,\ldots,n\\
\frac{\partial\mathcal{L}}{\partial v_l} = 0 \rightarrow \sum_{i=1}^n(\alpha_i\ell_{i,l}-\delta_i\alpha_i^*r_{i,l}-\theta_l+\theta_{l-1}) = 0 & \forall\ l=1,\ldots,k\\
\alpha_i,\alpha_i^*,\beta_i,\beta_i^*,\eta_i,\eta_i^*,\xi_i,\xi_i^* \geq 0 & \forall\ i=1,\ldots,n\\
\theta_l \geq 0 & \forall\ l=1,\ldots,k\\
\sum_{i=1}^n \alpha_i\left(w^T\varphi(x_i)+b-\ell_i^T(v-z)-y_i+\epsilon_i\right) = 0\\
\sum_{i=1}^n \alpha_i^*\left(-\delta_i(w^T\varphi(x_i)+b-r_i^T(v-z)-y_i)+\epsilon_i^*\right) = 0\\
\sum_{i=1}^n \beta_i\left(w^T\varphi(x_i)+b-y_i+\xi_i\right) = 0\\
\sum_{i=1}^n \beta_i^*\left(-\delta_i w^T\varphi(x_i)-b+\delta_i y_i+\xi_i^*\right) = 0\\
\sum_{i=1}^n \eta_i\epsilon_i = 0\\
\sum_{i=1}^n \eta_i^*\epsilon_i^* = 0\\
\sum_{i=1}^n \nu_i\xi_i = 0\\
\sum_{i=1}^n \nu_i^*\xi_i^* = 0\\
\sum_{l=1}^k \theta_l(v_l-v_{l-1}) = 0\,,
\end{cases} \tag{40}$$

with $\ell_{i,l}$ and $r_{i,l}$ the l^{th} element of the vectors ℓ_i and r_i respectively. After elimination of some variables, the solution is found as

$$\begin{aligned}\min_{\alpha,\alpha^*,\beta,\beta^*,\theta} \; & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\alpha_i\alpha_j+\beta_i\beta_j\right)\varphi(x_i)^T\varphi(x_j) \\ & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\delta_j\left(\alpha_i^*\alpha_j^*+\beta_i^*\beta_j^*\right)\varphi(x_i)^T\varphi(x_j) \\ & +\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\alpha_i\beta_j-\delta_j((\alpha_i+\beta_i)(\alpha_j^*+\beta_j^*))\right)\varphi(x_i)^T\varphi(x_j) \\ & -\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\alpha_i^*\beta_j^*\varphi(x_i)^T\varphi(x_j)-\sum_{i=1}^{n}\left(\alpha_i-\delta_i\alpha_i^*\right)y_i \\ & +\sum_{i=1}^{n}(\ell_i-r_i)^Tz-\sum_{i=1}^{n}\beta_iy_i+\sum_{i=1}^{n}\delta_i\beta_i^*y_i\end{aligned} \tag{41}$$

$$\text{subject to}\begin{cases}0\geq\alpha_i\geq\gamma, & \forall\; i=1,\ldots,n\\ 0\geq\alpha_i^*\geq\gamma, & \forall\; i=1,\ldots,n\\ 0\geq\beta_i\geq\mu, & \forall\; i=1,\ldots,n\\ 0\geq\beta_i^*\geq\mu, & \forall\; i=1,\ldots,n\\ \sum_{i=1}^{n}\left(\alpha_i\ell_{i,l}-\delta_i\alpha_i^*r_{i,l}\right)-\theta_l+\theta_{l-1} & \forall\; l=1,\ldots,k\,.\end{cases}$$

The prognostic index $u(x^\star)$ for a new observation $x^\star$ can then be found as

$$\hat{u}(x^\star)=\sum_{i=1}^{n}(\alpha_i+\beta_i-\delta_i(\alpha_i^*+\beta_i^*))\varphi(x_i)^T\varphi(x^\star)+b\,. \tag{42}$$

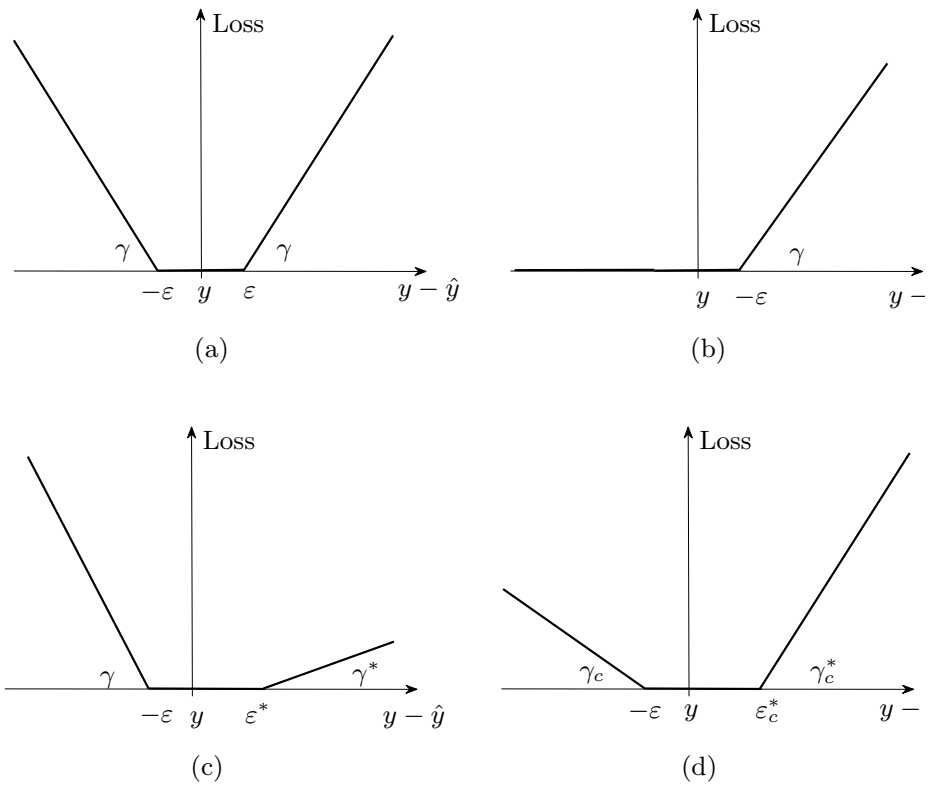
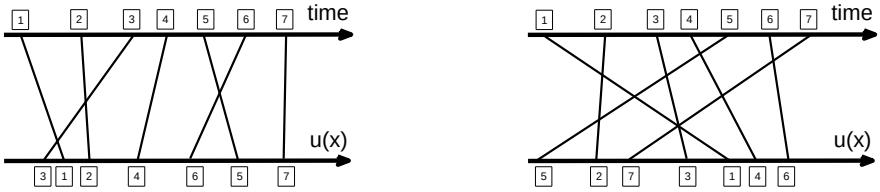


Fig. 1. Van Belle *et al.*



concordance	rank 1 2 3 4 5 6	ALL
2	1	1
3	0 0	0
4	1 1 1	3
5	1 1 1 1	4
6	1 1 1 1 0	4
7	1 1 1 1 1 1	6
		c-index
ALL	5 4 4 3 1 1	**18/21**

(a) first model

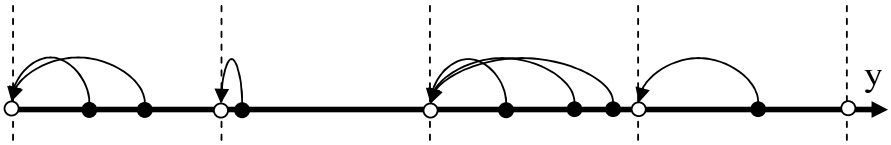
concordance	rank 1 2 3 4 5 6	ALL
2	0	0
3	0 1	1
4	1 1 1	3
5	0 0 0 0	0
6	1 1 1 1 1	1
7	0 1 0 0 1 0	2
		c-index
ALL	2 4 2 1 2 0	**11/21**

(b) second model

Fig. 2. Van Belle *et al.*



(a) MODEL 2



(b) MODEL 3

Fig. 3. Van Belle *et al.*

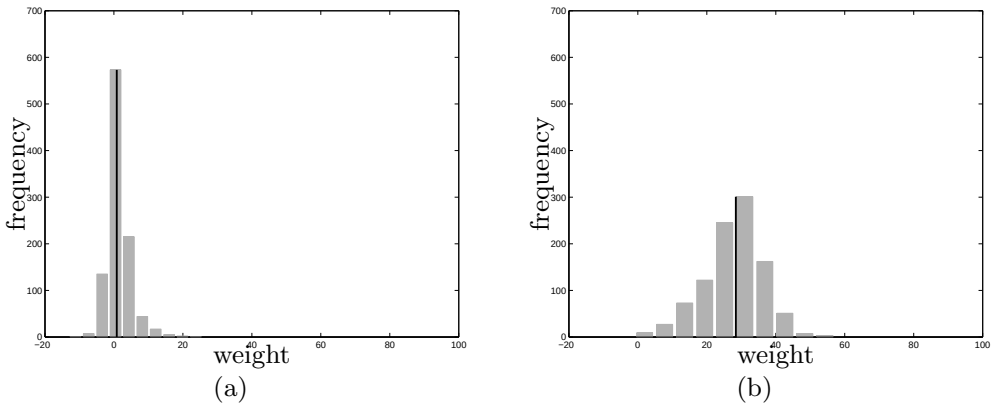


Fig. 4. Van Belle *et al.*

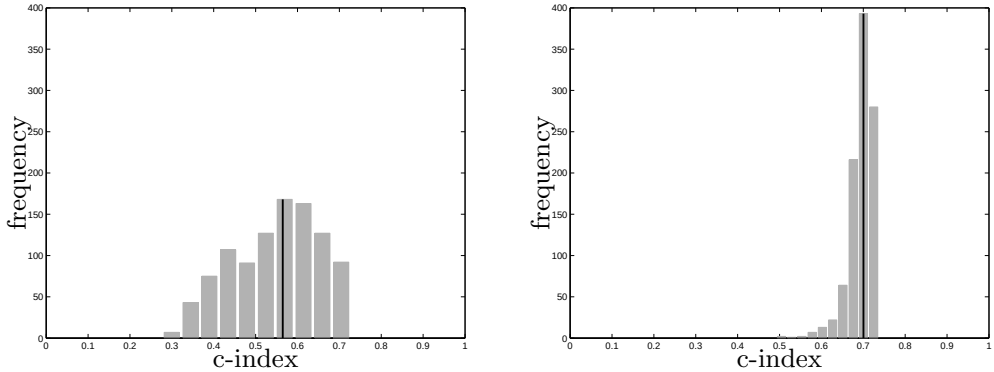


Fig. 5. Van Belle *et al.*

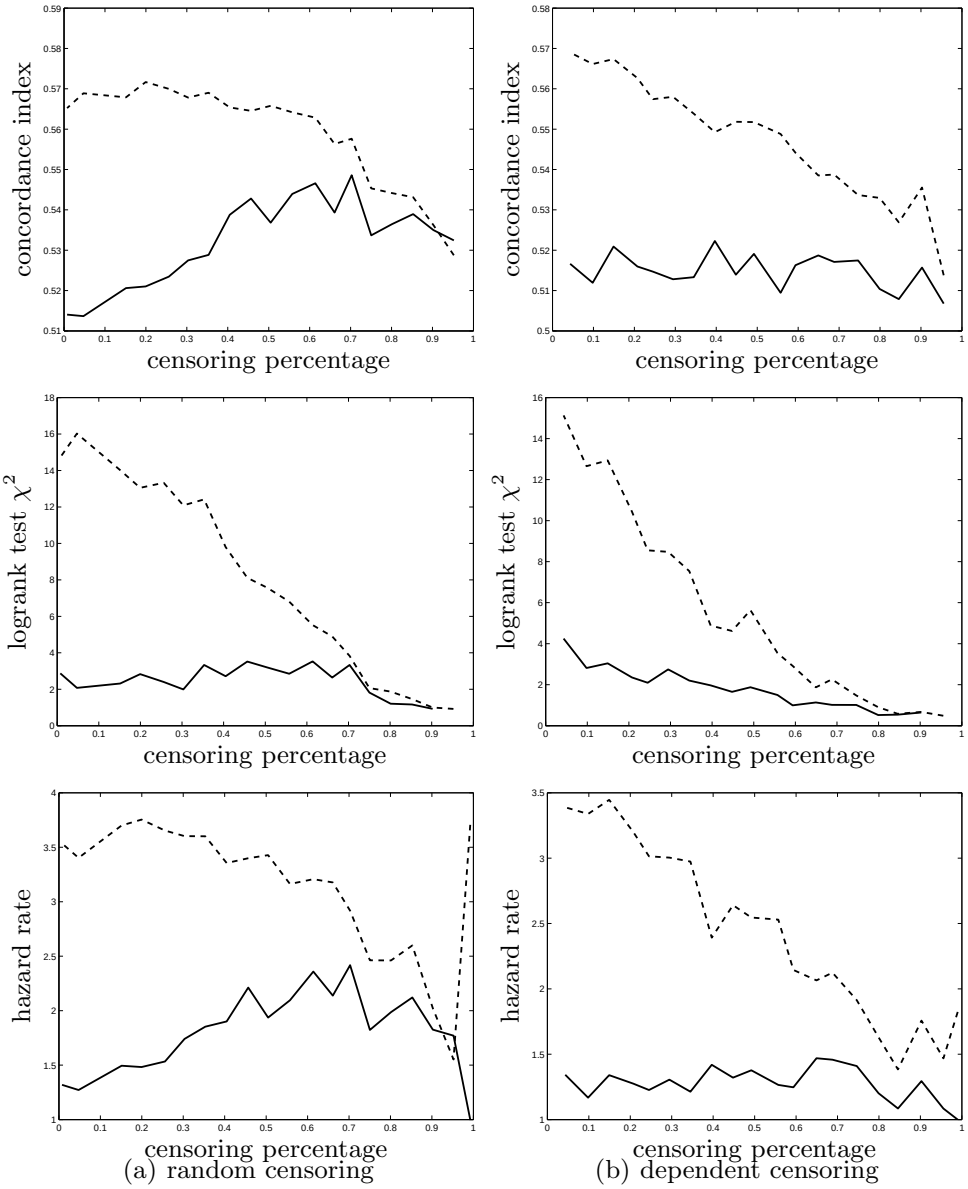


Fig. 6. Van Belle *et al.*