Analysis of sarcoma and non-sarcoma clinical data with statistical methods and machine learning techniques

Kantidakis, G.

8

Statistical models versus machine learning for competing risks: development and validation of prognostic models

This chapter is based on a joint work with H. Putter, S. Litière, and M. Fiocco.

Abstract

Background: In health research, several chronic diseases are susceptible to competing risks (CRs). Initially, statistical models (SM) were developed to estimate the cumulative incidence of an event in the presence of CRs. As recently there is a growing interest in applying machine learning (ML) for clinical prediction, these techniques have also been extended to model CRs but literature is limited. Here, our aim is to investigate the potential role of ML versus SM for CRs within non-complex data (small/medium sample size, low dimensional setting).

Methods: A dataset with 3826 retrospectively collected patients with extremity soft-tissue sarcoma (eSTS) and nine predictors is used to evaluate model-predictive performance in terms of discrimination and calibration. Two SM (cause-specific Cox, Fine-Gray) and three ML techniques are compared for CRs in a simple clinical setting. ML models include an original partial logistic artificial neural network for CRs (PLANNCR original), a PLANNCR with novel specifications in terms of architecture (PLANNCR extended), and a random survival forest for CRs (RSFCR). The clinical endpoint is the time in years between surgery and disease progression (event of interest) or death (competing event). Time points of interest are 2, 5, and 10 years.

Results: Based on the original eSTS data, 100 bootstrapped training datasets are drawn. Performance of the final models is assessed on validation data (left out samples) by employing as measures the Brier score and the Area Under the Curve (AUC) with CRs. Miscalibration (absolute accuracy error) is also estimated. Results show that the ML models are able to reach a comparable performance versus the SM at 2, 5, and 10 years regarding both Brier score and AUC (95% confidence intervals overlapped). However, the SM are frequently better calibrated.

Conclusions: Overall, ML techniques are less practical as they require substantial implementation time (data preprocessing, hyperparameter tuning, computational intensity), whereas regression methods can perform well without the additional workload of model training. As such, for non-complex real life survival data, these techniques should only be applied complementary to SM as exploratory tools of model's performance. More attention to model calibration is urgently needed.

8.1 Introduction

Survival analysis (also referred as time-to-event analysis) is used to estimate the lifespan of a particular population under study. Frequently, survival data are right censored; time to event is not observed for all patients due to follow-up interruption before experiencing the event of interest or time limitations (study termination). Competing risks (CRs) occur frequently in clinical applications of survival data [1–4]. In this type of data an individual may fail from one of several causes. A CR is an event whose occurrence precludes the occurrence of an event of interest (for instance death may preclude the occurrence of disease relapse) [5, 6]. In health research, CRs are unlikely to be independent as the biology suggests at least some dependence between events. In several chronic diseases attributable to aging and frailty such as cancer, chronic heart failure, or dementia, study populations are susceptible to CRs [7].

The most popular non-parametric approach to estimate survival in the presence of right censored time-to-event data is the Kaplan-Meier's methodology (KM) [8]. However, in the presence of CRs, this methodology overestimates the probability of failure which might lead to over-treatment of patients [1, 5, 9]. Different statistical models (SM) have been developed to estimate the cumulative incidence (absolute risk) of an event in the presence of CRs such as the cause-specific Cox model [10], and the Fine-Gray sub-distribution hazards regression model [11]. The former is a natural extension of the standard proportional hazards Cox model for the CRs setting where a Cox model is applied for each cause-specific hazard. The latter models the effect of covariates directly on the cumulative incidence function (CIF) over time reporting on the sub-distribution hazard ratio [9].

Nowadays, there is a growing interest in applying machine learning (ML) for prediction (diagnosis or prognosis) of clinical outcomes [12, 13] which has sparked a debate regarding the added value of ML techniques versus SM in the medical field. Criticism is attributed to ML prediction models. Despite no assumptions about the data structure are made, and being able to naturally incorporate interactions between predictive features, they are prone to overfitting of the training data and they lack extensive assessment of predictive accuracy (i.e., absence of calibration curves) [14, 15]. On the other hand, traditional regression methods are considered straightforward to use and harder to overfit. That being said, they do make certain (usually strong) assumptions such as the proportional hazards over time for the Cox model, and require manual pre-specification of interaction terms.

Amongst ML techniques, artificial neural networks have been a common choice in healthcare. This trend is pertinent with the collection of large and complex patient information in electronic health records, and the rise of computational power [16]. Over the years, neural networks and other ML techniques have been developed for survival data. Wang *et al.* in 2019 provide a comprehensive survey of conventional and modern approaches for right-censored time-to-event data [17]. The authors describe several ML techniques and suggest that neural networks are well-suited to predict survival and estimate disease risk.

A common approach in the literature is the partial logistic artificial neural network (PLANN) of Biganzoli *et al.* (1998) [18]. For the purpose of implementation, time is specified in discrete non-overlapping time intervals which are added as an input feature in a longitudinally transformed feed-forward network with logistic activation, and entropy error function. The output layer estimates smoothed discrete hazards for each time interval. PLANN was extended by Lisboa *et al.* (2003) under a Bayesian regularisation framework which performs automatic relevance determination (PLANN-ARD) [19]. Recently, Kantidakis *et al.* in 2020 proposed extensions of PLANN in terms of architecture i.e., new hyperparameters, new activation functions, and time interval specification as multiple input features [20]. Next to survival neural networks (SNNs), another well-known ML technique for clinical prediction of survival data is random survival forests (RSF, Ishwaran *et al.* 2008) [21]. RSF adapt Breiman's random forest method by using a collection of survival trees [22].

ML approaches have also been employed for CRs, but the literature is limited. The PLANNCR approach was developed by Biganzoli *et al.* in 2006 for the joint modelling of discrete cause-specific hazards [23]. This extends

PLANN by using the time (in discrete time intervals) as an input feature in a longitudinally transformed network with multinomial error function and logistic - softmax activation functions for the hidden and the output layer (multiple output nodes), respectively. Later, Lisboa *et al.* (2009) implemented PLANNCR under a Bayesian regularisation framework (PLANNCR-ARD) [24]. Ishwaran *et al.* extended RSF for CRs (RSFCR) in 2014 to estimate the CIF of competing events [25].

For this work, a dataset with small/medium sample size and limited number of predictive features (low-dimensional setting) is analysed. This concerns a retrospectively collected cohort of 3826 patients with high-grade extremity soft-tissue sarcomas (eSTS) treated surgically with curative intent. Nine prognostic factors are used to develop and validate several clinical prediction models with CRs for ML techniques and SM. The clinical endpoint of the study is defined as the time in years between surgery and disease progression (as local recurrence or distant metastasis; event of interest) of eSTS, where death is a competing event. Time points of interest are 2, 5, and 10 years (5-year horizon is of major clinical interest). Analyses were performed in R programming language version 4.1.2 [26].

The aims of this manuscript can be summarised as: (i) examination of extensions of PLANNCR method (PLANNCR extended) for the development and validation of prognostic clinical prediction models with competing events, (ii) systematic evaluation of model-predictive performance for ML techniques (PLANNCR original, PLANNCR extended, RSFCR) and SM (cause-specific Cox, Fine-Gray) regarding discrimination and calibration, (iii) investigation of the potential role of ML in contrast to conventional regression methods for CRs in non-complex eSTS data (small/medium sample size, low dimensional setting), (iv) practical utility of the methods for prediction.

The paper is organized as follows. In Section ”Methods”, the eSTS data is presented. Further sections discuss basic concepts for CRs, the SM and the ML techniques, model training, and how the predictive performance was assessed. Section ”Results” describes PLANNCR extended tuned with two measures, and compares the predictive performance of all methods in terms of discrimination and calibration. The manuscript ends with a ”Discussion” about findings, limitations, and future perspectives of this work.

8.2 Methods

This section is divided into several subsections where the methodology used for this work is presented to the reader. To begin with, the clinical data is described. Next, the SM and the ML techniques are discussed. Afterwards, it is presented how the models were trained, and which performance measures were used to evaluate their predictive ability. More technical details are provided in the supplementary material.

8.2.1 Dataset

Extremity soft-tissue sarcomas (eSTS) constitute a wide variety of histological subtypes with different sizes and grades that affect patients of any age group. Treatment protocols may differ between institutes and countries. Hence, important differences can be observed in the clinical course and prognosis of patients [27]. Over the years, several prognostic prediction models have been developed for overall survival and local recurrence [28–30].

For this project, a retrospectively collected cohort of 3826 patients with eSTS was used [29]. The dataset contained pseudo-anonymised patients from Leiden University Medical Center (Leiden, the Netherlands), Royal Orthopaedic Hospital (Birmingham and Stanmore, UK), Netherlands Cancer Institute (Amsterdam, the Netherlands), Mount Sinai Hospital (Toronto, Canada), the Norwegian Radium Hospital (Oslo, Norway), Aarhus University Hospital (Aarhus, Denmark), Skåne University Hospital (Lund, Sweden), Medical University Graz (Graz, Austria), Royal Marsden Hospital (London, UK), Daniel den Hoed (Rotterdam, the Netherlands), Radboud University Medical Center (Nijmegen, the Netherlands), University Medical Center Groningen (Groningen, the Netherlands), Haukeland University Hospital (Bergen, Norway), Helios Klinikum Berlin-Buch (Berlin, Germany), MedUni Vienna (Vienna, Austria), Vienna General Hospital (Vienna, Austria). In addition, eSTS patients from EORTC 62931 randomised controlled trial were included [31]. Data from the centers was collected between January 2000 and December 2014. Patients from the EORTC trial were recruited between February 1995 and December 2003.

Characteristics	**Total** (N = 3826)
Gender (%)	
Female	1713 (44.77%)
Male	2113 (55.23%)
Mean age in years (sd)	59.40 (18.04)
Mean tumor size in cm (sd)	8.97 (5.69)
Surgical margin (%)	
R_0	3310 (86.51%)
R_{1-2}	516 (13.49%)
Adjuvant chemotherapy (%)	
No	3350 (87.56%)
Yes	476 (12.44%)
Tumor grade (%)	
II	656 (17.15%)
III	3170 (82.85%)
Histological subtype (%)	
Myxofibrosarcoma	771 (20.15%)
Synovial sarcoma	450 (11.76%)
MFH/UPS/NOS	1330 (34.76%)
Leiomyosarcoma	385 (10.06%)
Liposarcoma	421 (11.00%)
Other	469 (12.26%)
Tumor depth (%)	
Superficial	1014 (26.50%)
Deep	2812 (73.50%)
Radiotherapy (%)	
No	1341 (35.05%)
Neoadjuvant	521 (13.62%)
Adjuvant	1964 (51.13%)

Table 8.1: **Patient demographics.** sd, standard deviation; R_0, negative margin; R_{1-2}, positive margin with tumor cells in the inked surface of the resection margin; MFH/UPS/NOS, alignant fibrous histiocytoma / undifferentiated pleomorphic sarcoma / (pleomorphic) soft tissue sarcomas not-otherwise-specified; histology ”Other”, angiosarcoma, clear cell sarcoma, conventional fibrosarcoma, epithelioid sarcoma, giant cell sarcoma, malignant granular cell tumor, malignant peripheral nerve sheath tumor, rhabdomyosarcoma (adult form), spindle cell sarcoma, unclassified soft tissue sarcoma and undifferentiated sarcoma.

Patients were selected from the sarcoma registry of each hospital based on histological diagnosis. Those initially treated without curative intent, showed local recurrence or distant metastasis at baseline, had Kaposi’s sarcoma or rhabdomyosarcoma (pediatric form), tumor was present in their abdomen, thorax, head or neck, or were treated with isolated limp perfusion as neoadjuvant treatment were excluded from the collection.

The dataset contained nine prognostic factors. Seven were categorical; *gender* (female or male), *surgical margin* (R_0 for negative or R_{1-2} for positive with tumor cells in the inked surface of the resection margin), *adjuvant chemotherapy* (no or yes), *tumor grade* (II or III), *tumor depth* in relation to investing fascia (superficial or deep), *radiotherapy* (no, neoadjuvant or adjuvant), *histological subtype* (myxofibrosarcoma, synovial sarcoma, malignant fibrous histiocytoma / undifferentiated pleomorphic sarcoma / (pleomorphic) soft tissue sarcomas not-otherwise-specified, leiomyosarcoma, liposarcoma or other). Two were continuous; *age* at baseline (in years) and *tumor size* by the largest diameter measured at pathological examination (in centimetres).

Median follow-up survival time is 5.98 years estimated by reverse Kaplan-Meier (25% quartile: 3.94 years, 75% quartile: 8.80 years, range: 0.01 to 16.85 years) [8]. The endpoint of interest is defined as the time in years between surgery and disease progression (local recurrence or distant metastasis) of eSTS, with death as competing event; 1773 patients were alive/censored at the end of follow-up (46.34%), 1554 had disease progression (40.62%), and 499 died without local recurrence/distant metastasis (13.04%).

The dataset contained 3.70% missing data overall for the nine variables, with 2514 complete cases (65.71%). More specifically, there were missing values (0.97-11%) for all variables; 11.00% for *tumor depth* (421/3826), 8.21% for *histological subtype* (314/3826), 7.40% for *surgical margin* (283/3826), 4.36% for *adjuvant chemotherapy* (167/3826), 4.05% for *tumor size* (155/3826), 3.53% for *gender* (135/3826), 2.61% for *radiotherapy* (100/3826), 1.99% for *tumor grade* (76/3826), and 0.97% for *age* (37/3826), in decreasing order, respectively.

A simple imputation was used to avoid discarding observations from nearly complete records. The `missForest` algorithm was applied to reconstruct any missing values, which is the most exhaustive/accurate random forest algorithm for missing data [32]. This is a nonparametric imputation method that does not make any a priory assumptions regarding the data structure. A random forest with 1000 trees (for model stability) was built for each variable with missing information, testing all possible variable combinations as responses. Table 8.1 provides patient demographics of the final dataset (demographics of the original dataset are provided in table S1 of Supplementary file 1).

8.2.2 Basic concepts for competing risks

Typically for survival data, if several types of events occur, a model describing progression for each of the CRs is needed. The observable data is represented by the time of failure T, the cause of failure D ($D \in 1, \cdots, k$, $k \geq 1$; here $k = 2$), and a covariate vector **Z**. Usually there is one type of event that is of interest (i.e., disease progression as local recurrence or distant metastasis) whereas the other events could prevent it from occurring (here competing event is death).

Following Putter *et al.* (2007) [1], a fundamental concept in modelling CRs is the cause-specific hazard function which denotes the hazard of failing from a given cause in the presence of CRs:

$$\lambda_k(t) = \lim_{\Delta t \to 0} \frac{Prob(t \leq T < t + \Delta t, D = k | T \geq t)}{\Delta t}. \tag{8.1}$$

Then, the cumulative cause-specific hazard can be specified as

$$\Lambda_k(t) = \int_0^t \lambda_k(s) ds \tag{8.2}$$

and the survival function (probability of not having failed from any cause at time t) can be written as

$$S(t) = \exp\Big(- \sum_{j=1}^{k} \Lambda_j(t) \Big). \tag{8.3}$$

The cumulative incidence function (CIF) of cause k is defined as $I_k(t) = Prob(T \leq t, D = k)$, the probability of failing from cause k before time t. This can be linked to the cause-specific hazards through the expression:

$$I_k(t) = \int_0^t \lambda_k(s) S(s) ds. \tag{8.4}$$

This is also called the subdistribution function based on the fact that the cumulative probability to fail from cause k cannot reach one, and therefore, it is not a proper probability distribution.

8.2.3 Regression models for competing risks

Cause-specific Cox model

Regression on cause-specific hazards is an extension of the popular Cox proportional hazards model for CRs [10, 33]. The cause-specific hazard of cause k of a subject with covariate vector $\mathbf{Z}$ is modelled as

$$\lambda_k(t|\mathbf{Z}) = \lambda_{k,0}(t)\,\text{exp}\,(\boldsymbol{\beta}_k^T\mathbf{Z}), \tag{8.5}$$

where $\lambda_{k,0}(t)$ is the cause-specific hazard, and the vector $\boldsymbol{\beta}_k$ represents the effects of covariates on cause k. Patients who move to another state other than k are censored at their transition time.

Fine and Gray model

In 1999, Fine and Gray introduced a subdistribution hazards model, which can directly regress on CIF [11]:

$$\tilde{\lambda}_k(t) = -\frac{d\,\text{log}(1 - I_k(t))}{dt}. \tag{8.6}$$

For the cause-specific Cox model, the risk set (number of patients at risk) decreases at each time point where there is a failure of another cause. On the other hand, for Fine and Gray's model, individuals who fail from another cause remain in the risk set. The subdistribution hazards are then modelled assuming proportional hazards:

$$\tilde{\lambda}_k(t|\mathbf{Z}) = \tilde{\lambda}_{k,0}(t)\,\text{exp}\,(\boldsymbol{\beta}_k^T\mathbf{Z}). \tag{8.7}$$

Similar to the standard Cox model, the partial likelihood approach is used to estimate the parameters.

8.2.4 Machine learning techniques for competing risks

Random survival forests

Random survival forests for competing risks (RSFCR) [25] are an extension of the RSF framework [21, 22] for CRs with right censored data proposed by Ishwaran *et al* in 2014. It is a fully non-parametric ensemble tree approach for the estimation of the CIF for competing events (CIF and cause-specific hazard function are related as shown in equation (8.4)). RSFCR can directly model non-linear effects and interactions to perform accurate prediction without making any prior assumptions about the underlying data.

The algorithm of RSFCR is based on recursive binary partitioning while injecting randomness in two ways: (a) drawing B bootstrap samples from the learning data, and (b) growing a single CRs tree for each bootstrap sample by randomly selecting a subset of candidate variables at each node (region of the tree). A CR splitting rule is maximised to split each parent node into daughter nodes using the selected variables. The authors propose two splitting rules: either an event-specific or a combination of event-specific splitting rules across the k events. Here, the event-specific splitting rule was applied because disease progression was of major interest (weighted log-rank splitting, technical details in [25]). Then each tree is grown to full size under the constraint that terminal nodes (the ends of each tree) should have at least one unique case. In the terminal nodes, the Kaplan-Meier [8] and the Aalen-Johansen [34] methodologies are used to estimate the event-free survival function and the cause-specific

CIF, respectively. Finally, the ensemble estimates are calculated averaging each estimator over the B grown trees. More technical details are provided in Supplementary file 2.

Partial logistic artificial neural networks

In 2006, Biganzoli *et al.* extended the partial logistic artificial neural network to competing risks (PLANNCR) for the joint modelling of discrete cause-specific hazards [18, 23]. PLANNCR is a feed-forward network comprised of a group of units called nodes (or neurons) in each layer. It has an input layer that picks up the signals and passes them to a single hidden layer after the application of an activation (also called transformation) function. An activation function modulates the degree of non-linearity transferred from the input features to the hidden layer. Connections between the artificial neurons of different layers are called edges - each having a weight. Weights are adjusted through training increasing or decreasing the strength of each connection [35]. Signals are transmitted towards the output layer, which provides a smoothed estimation of discrete conditional event probabilities (in multiple output nodes; each for an event), with another activation function.

For the purpose of implementation, survival times are discretized into a set of $l = 1, \cdots, L$ disjoint intervals $A_l = (\tau_{l-1}, \tau_l]$, where $0 = \tau_0 < \tau_1 < \cdots < \tau_L$ is a set of pre-defined time points (usually years). For the l^{th} interval, observed times are grouped on a single point τ_l. Data has to be transformed into a longitudinal format where the time variable (in intervals) is added as part of the input features next to the prognostic features. Subjects are repeated for the number of intervals observed on the training data, and for all time intervals on the test data. PLANNCR can model non-linear, non-proportional, and non-additive effects between the prognostic factors on the cause-specific hazards. Here, without loss of generality, each subject was repeated for 1 up to 11 time intervals denoting years since surgery. The last interval included survival times longer than 10 years (subsequent intervals were not of interest).

In the CRs model, the response vector has $R+1$ variables, with $r = 1, \cdots, R$ the possible causes of interest (here $R = 2$). Let $\mathbf{z}_k = (\tau_l, \mathbf{x}_k)$ be defined by two components: the covariate vector $\mathbf{x}_k$ ($k = 1, 2, \cdots, p$) and the time interval τ_l. The joint dependence of the discrete cause-specific hazards is modelled as:

$$\eta_{lr}(\mathbf{z}_k, \boldsymbol{\beta}) = \beta_0 + \sum_{h=1}^{H} \beta_r^a \alpha_h(\beta_{0h} + \boldsymbol{\beta}_h^T \mathbf{z}_k) \tag{8.8}$$

where $h = 1, \cdots, H$ nodes in the hidden layer, $\boldsymbol{\beta}$ the vector of estimated weights for the input-hidden $(\beta_{01}, \cdots, \beta_{0H}, \beta_1, \cdots, \beta_H)$, hidden-output layers $(\beta_0, \beta_1^a, \cdots, \beta_R^a)$, and α_h the `sigmoid` (logistic) activation function for the hidden layer $\alpha_h(\mathbf{z}_k, \boldsymbol{\beta}_h) = \frac{\exp(\beta_{0h} + \boldsymbol{\beta}_h^T \mathbf{z}_k)}{1 + \exp(\beta_{0h} + \boldsymbol{\beta}_h^T \mathbf{z}_k)}$.

Activation function for the output layer is the `softmax` providing the discrete cause-specific hazards:

$$\tilde{h}_{lr}(\mathbf{z}_k, \boldsymbol{\beta}) = \frac{\exp\left(\eta_{lr}(\mathbf{z}_k, \boldsymbol{\beta})\right)}{\sum_{r=1}^{R+1} \exp\left(\eta_{lr}(\mathbf{z}_k, \boldsymbol{\beta})\right)}, \tag{8.9}$$

for $l = 1, \cdots, L$ intervals, and $r = 1, \cdots, R$ causes of interest. Since PLANNCR has a different output node for each CR ($1 + R$ output nodes in total), it is an extension of standard neural networks for multiple classification resorting to the multinomial likelihood. For the rest of this paper, this will be called PLANNCR original [23].

Similar extensions to the specification of the PLANNCR are provided as in Kantidakis *et al.* (PLANN extended, 2020) [20]. More specifically, PLANNCR extended is tuned investigating two new activation functions for the hidden layer: (1) the rectified linear unit (ReLU) a common activation function, $\alpha_h(\mathbf{z}_k, \boldsymbol{\beta}_h) = \max(0, \beta_{0h} + \boldsymbol{\beta}_h^T \mathbf{z}_k)$, and (2) the hyperbolic tangent (tanh), $\alpha_h(\mathbf{z}_k, \boldsymbol{\beta}_h) = \frac{1 - \exp(-2(\beta_{0h} + \boldsymbol{\beta}_h^T \mathbf{z}_k))}{1 + \exp(-2(\beta_{0h} + \boldsymbol{\beta}_h^T \mathbf{z}_k))}$. Note that the activation function for the output layer is necessarily the `softmax` to provide smoothed discrete hazard estimation. New hyperparameters are specified in a state-of-the-art R library [36]. In contrast with Kantidakis *et al.* (2020), the L non-overlapping intervals are specified in one time variable (instead of L separate variables) to not inflate the number of input features. Moreover, networks with two hidden layers are not tested here due to the danger for overfitting (small-medium sample size, small number of predictors). More technical details for PLANNCR original and PLANNCR extended are provided in Supplementary file 2.

8.2.5 Model training

Figure 8.1 shows how model training was performed. Based on the original eSTS data, 100 bootstrapped training datasets were drawn with 3826 patients each (sampling with replacement, $\approx 63.2\%$ of the original data). These datasets were randomly split into two complementary parts to tune the hyperparameters of the ML models using grid search ($\frac{3}{4}$ to train the models and $\frac{1}{4}$ to test their performance, same parts for all methods). Performance of the final models was assessed on the validation data, which were the left out samples (out-of-bag, $\approx 36.8\%$ of the data). Out-of-bag error estimates are almost identical to N-fold cross-validation [37]. For the standard regression approaches, models were built on each complete training dataset (consisted of 3826 patients) using the nine covariates. Their predictive performance was evaluated on the respective validation dataset. Complex functional form dependencies (non-linear, non-additive, time-dependent effects) were not investigated. All analyses were performed in R programming language version 4.1.2 [26]. Packages used in the implementation and tuning parameters for the ML techniques are provided in Supplementary file 2.

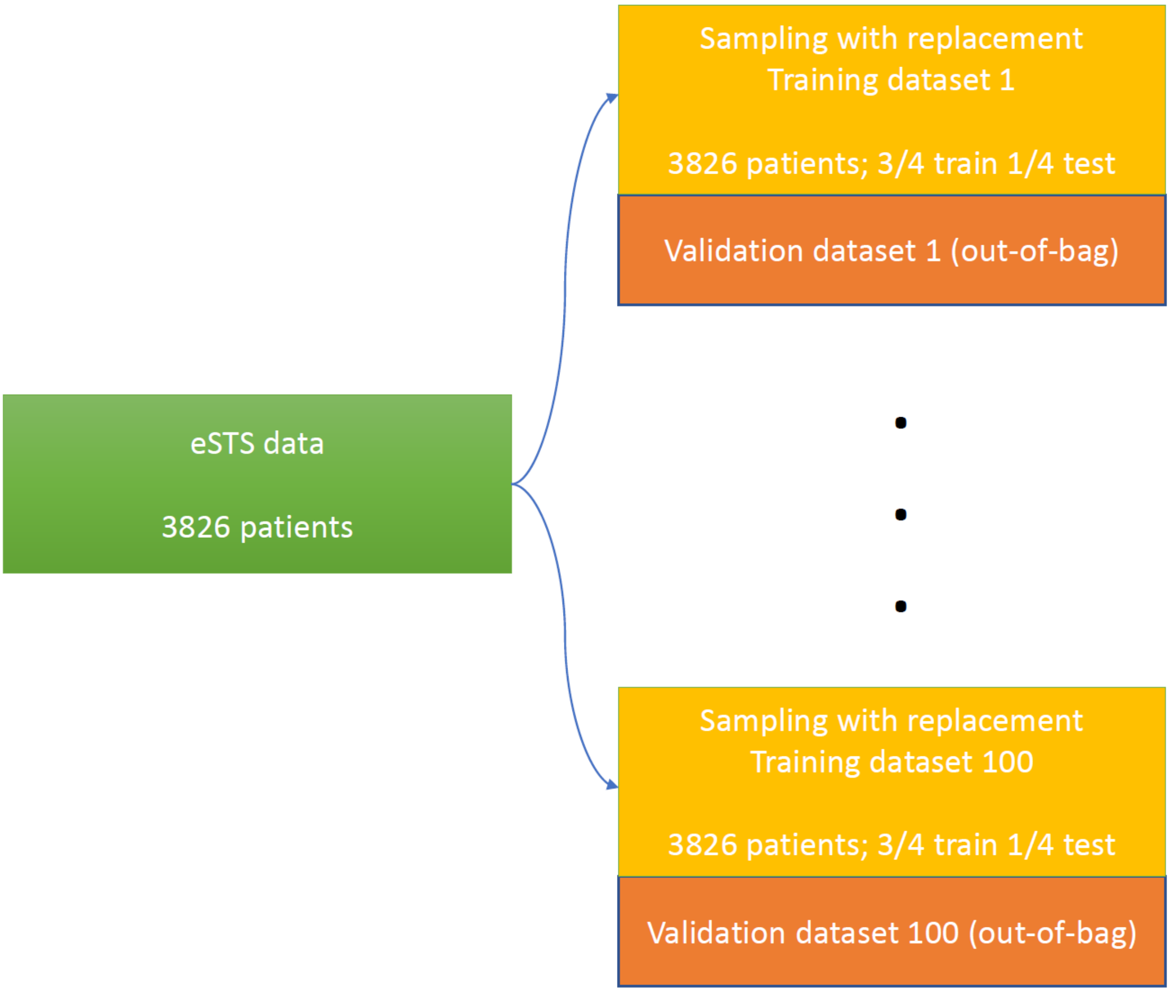
Sampling with replacement
Training dataset 1
3826 patients; 3/4 train 1/4 test
Validation dataset 1 (out-of-bag)
eSTS data
3826 patients
Sampling with replacement
Training dataset 100
3826 patients; 3/4 train 1/4 test
Validation dataset 100 (out-of-bag)

Figure 8.1: Illustration of the model training approach repeated 100 times. For the ML techniques, hyperparameters were tuned on the training datasets. Final performance for all models was assessed on the validation datasets (left out samples).

8.2.6 Predictive performance assessment

Predictive performance of the methods was assessed in terms of discrimination and calibration on each validation dataset. The Area Under the Curve (AUC) and Brier score with CRs were used. Miscalibration (absolute accuracy error) was also estimated.

Following Blanche *et al.* [38], we present the dynamic version of the measures with CRs. Let $\pi_i(\cdot, \cdot)$ be a subject-i specific prediction process ($i = 1, 2, \cdots, n$ independent and identically distributed subjects) for all landmark times s (times at which predictions are made) and prediction horizon t. Without loss of generality, we set $\pi_i(s,t) = 0$ for all subjects i who are no longer at risk at time s, and focus on prediction of event $D = 1$ (main event investigated). A dynamic AUC at landmark time s for a prediction horizon t can be defined as

$$AUC(s,t) = Prob\Big(\pi_i(s,t) > \pi_j(s,t)|\Delta_i(s,t) = 1, \Delta_j(s,t) = 0, T_i > s, T_j > s\Big), \tag{8.10}$$

where $\Delta_i(s,t) = \mathbb{1}_{s<T_i\leq s+t, D_i=1}$, $\Delta_i(s,t) = 1$ when subject i experiences the main event of interest within time interval $(s, s+t]$ (case), and $\Delta_i(s,t) = 0$ when subject i experiences a competing event within the time interval or is event-free at $s+t$ (control) [39].

Dynamic AUC with CRs is a measure of discrimination. It typically ranges from 0.5 to 1 (the higher the better). A good predictive accuracy is provided by a model that usually gives higher predicted risks of event for subjects who experience the event of interest compared to subjects who did not experience the event of interest.

A more complete predictive accuracy measure with CRs is the Brier score. The dynamic expected Brier score can be written as

$$BS(s,t) = \mathbb{E}[\big(\Delta(s,t) - \pi(s,t)\big)^2|T > s]. \tag{8.11}$$

This expression can be expanded based on Graaf *et al.* 1999 [40] taking the following form

$$BS(s,t) = \mathbb{E}[\big(\mathbb{E}[\Delta(s,t)|H(s)] - \pi(s,t)\big)^2|T > s] + \mathbb{E}[\big(\Delta(s,t) - \mathbb{E}[\Delta(s,t)|H(s)]\big)^2|T > s], \tag{8.12}$$

where $H(s) = \{\mathbf{X}, Y(s), T > s\}$ the information at time s used to compute the prediction of $\pi(s,t)$. The first term in (8.12) measures calibration - how close the predictions are to $\mathbb{E}[\Delta(s,t)|H(s)]$, the "true" underlying risk of event in $(s, s+t]$ given $H(s)$. In addition, the second term depends on the discrimination ability of $H(s)$. Thus, Brier score is a measure of both calibration and discrimination. Typically, it ranges from 0 to 0.25 (lower values mean smaller prediction error).

When censored data are present, the indicator $\Delta_i(s,t)$ is unknown (cannot be computed) for all subjects i censored within interval $(s, s+t]$. Therefore, the Inverse Probability of Censoring Weighting (IPCW) technique has to be applied for the estimation of both dynamic AUC and Brier score for CRs. For details see [38]. Here, the landmark time was set to $s = 0$ (baseline) for all analyses as all prognostic factors were time fixed.

Last, the predictive ability of the methods was evaluated based on their miscalibration on each validation dataset (see figure 8.1). Model calibration refers to the agreement between observed and predicted outcomes, in this case agreement between observed and predicted cumulative incidence event probabilities for a cause $D = k$ at time $t = t_0$ [41, 42]. For each SM and ML model, the predicted cumulative incidence event probabilities are estimated on a validation dataset, and the data is split into $m = 4$ equally sized groups based on the quantiles of the predicted event probabilities. Quantiles were selected instead of (for instance) deciles to avoid any computational issues. Then, the observed cumulative incidence probabilities are calculated for each group. Miscalibration is defined as the mean squared error (MSE) of the difference between the observed and the predicted cumulative probabilities of failure from a specific cause $D = k$ at time horizon $t = t_0$

$$MSE_k(t_0) = \frac{\sum_{m=1}^{4}\big[I_k^{(m)}(t_0) - \hat{I_k}^{(m)}(t_0)\big]^2}{4}, \tag{8.13}$$

with $I_k^{(m)}(t_0)$ and $\hat{I}_k^{(m)}(t_0)$ the observed and predicted cumulative event probability for group m, respectively.

8.3 Results

In this section, results for the eSTS data are presented. The following models are compared in terms of predictive performance: (1) Cause-specific Cox, (2) Fine-Gray, (3) PLANNCR original, (4) PLANNCR extended, (5) RSFCR. Each model is assessed on 100 validation datasets (see figure 8.1). More results about the comparison between the methods are provided in Supplementary file 3.

8.3.1 PLANNCR tuned with Brier score or AUC at 5 years

The hyperparameters selected for PLANNCR original and PLANNCR extended are provided in section 1 of Supplementary file 3. The most effective combinations are reported separately based on the Brier score / AUC at 5 years (5-year horizon was of major clinical interest).

For PLANN original, both performance measures selected the same values for the 2 hyperparameters (*size* and *decay*). On the other hand, separate hyperparameters were selected for PLANNCR extended on a 5-D space (*nodesize*, *dropout rate*, *learning rate*, *momentum*, *weak class weight*). The technical details can be found in Supplementary file 2. From the 3 activation functions tested for the hidden layer (”sigmoid”, ”relu”, ”tanh”), the ”sigmoid” provided the best performance on the training data for both Brier score and AUC. A *weak class weight* of 1 was selected (no adjustment for disease progression or death).

The performance of the tuned PLANNCR extended was compared for disease progression (event of interest). Results are presented in table 8.2. PLANNCR extended tuned with Brier score at 5 years had a better performance in terms of Brier score and miscalibration at 2, 5, or 10 years. However, PLANNCR extended tuned with AUC at 5 years had a better performance regarding AUC at 5 and 10 years. These results were expected as Brier score is a more complete measure taking into account both discrimination and calibration. For the rest of the results presented below, optimal combinations for Brier score at 5 years were selected for PLANNCR extended.

Performance	PLANNCR extended with Brier score	PLANNCR extended with AUC
Brier score at 2 years	0.208 (0.198 - 0.220)	0.214 (0.201 - 0.226)
Brier score at 5 years	0.228 (0.221 - 0.235)	0.231 (0.225 - 0.236)
Brier score at 10 years	0.238 (0.229 - 0.247)	0.240 (0.234 - 0.247)
AUC at 2 years	0.661 (0.637 - 0.688)	0.659 (0.640 - 0.683)
AUC at 5 years	0.652 (0.612 - 0.689)	0.660 (0.633 - 0.685)
AUC at 10 years	0.629 (0.576 - 0.681)	0.631 (0.582 - 0.678)
Miscalibration at 2 years	0.008 (0.003 - 0.017)	0.013 (0.006 - 0.022)
Miscalibration at 5 years	0.003 (0.001 - 0.008)	0.008 (0.004 - 0.014)
Miscalibration at 10 years	0.002 (0.000 - 0.008)	0.004 (0.001 - 0.009)

Table 8.2: Mean predictive performance of PLANNCR extended for disease progression (event of interest) tuned with Brier score or AUC at 5 years. The 95% confidence intervals are provided in parentheses based on 100 validation datasets.

8.3.2 Predictive performance comparison

In this section, the five methods are compared on the 100 validation datasets for different predictive performance measures: (1) Brier scores, (ii) AUC, (iii) miscalibration at 2, 5, and 10 years, respectively, for disease progression (local recurrence or distant metastasis). Optimal hyperparameters and additional plots for the event of interest (disease progression) and the competing event (death) are included in sections 1 and 2 of Supplementary file 3.

Brier score - AUC

Figure 8.2 shows the Brier score (lower values better) and AUC (higher values better) at 2, 5 and 10 years since surgery for all methods regarding disease progression.

For the time-dependent Brier score, the cause-specific Cox model had in general the best performance followed by the Fine-Gray model and RSFCR at 2 years, and the PLANNCR extended and Fine-Gray at 5 and 10 years. PLANNCR original had slightly the worst performance at these time points. 95% confidence intervals (CI) based on the percentile method for 100 validation datasets using the out-of-bag data overlapped. PLANNCR extended had marginally larger 95% CI at 2 years and RSFCR at 10 years. Regarding AUC at 2, 5, and 10 years, the cause-specific Cox model and the PLANNCR extended had the best performance (very close to each other) followed by Fine-Gray model, RSFCR and PLANNCR original in decreasing order of performance. The 95% confidence intervals were very similar for the methods, except for PLANNCR original which had much wider intervals at all times. This means that its discrimination ability (AUC) was not consistent (fluctuated) in the validation datasets.

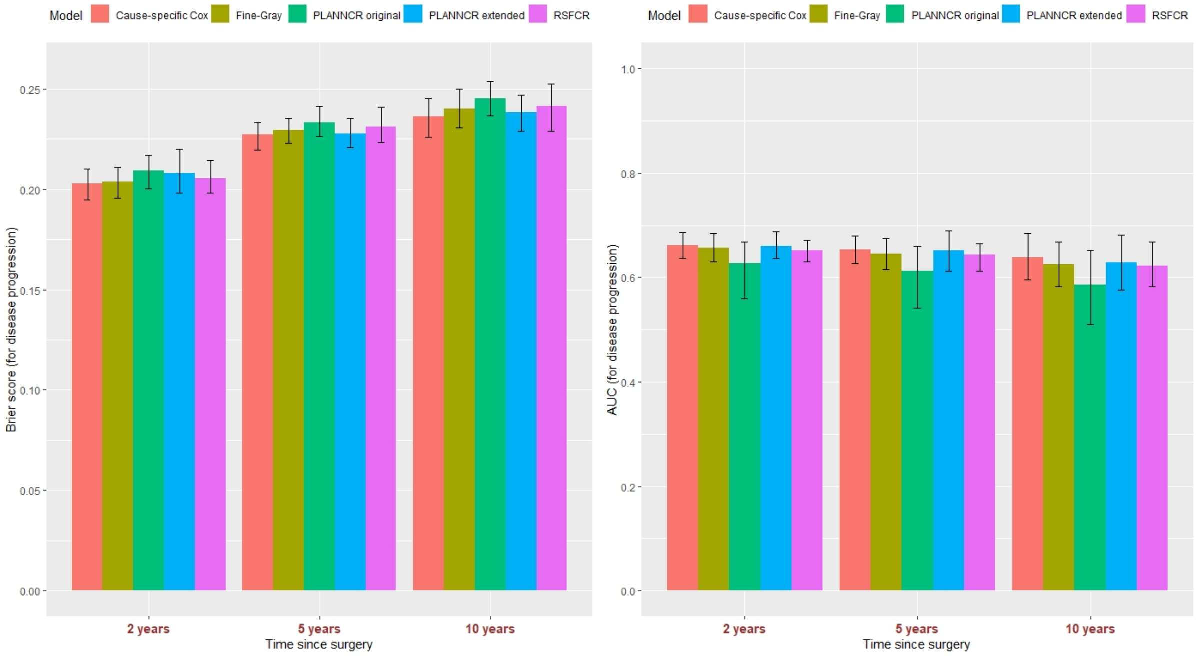

Figure 8.2: Predictive performance of cause-specific Cox model, Fine-Gray model, PLANNCR original, PLANNCR extended (tuned with Brier score at 5 years), and RSFCR for the event of interest: disease progression ± 95% percentile confidence intervals based on 100 validation datasets. Left panel: Brier score, right panel: AUC at 2, 5, and 10 years since surgery.

Figure S1 in Supplementary file 3 provides the same plot with PLANNCR extended tuned with AUC at 5 years. The predictive ability decreased in terms of Brier score but slightly increased regarding AUC at 5 and 10 years (see also table 8.2). Figures S3 and S5 in Supplementary file 3 illustrate the prognostic ability (Brier score, AUC) of all models for death (the competing event). The SM (cause-specific Cox and Fine-Gray) had the lowest Brier score followed by the RSFCR. PLANNCR models had worse performance and larger CI than the rest at 2 years. PLANNCR original continued to have larger CIs at 5 and 10 years, whereas PLANNCR extended had narrower

CIs at 5 and 10 years (more consistent performance). For AUC, the cause-specific Cox model and the PLANNCR extended had the highest values followed by the Fine-Gray model and the RSFCR. PLANNCR original the lowest performance and the largest 95% CI.

Miscalibration

The five models were investigated in terms of miscalibration (definition in section ”Predictive performance assessment”) at 2, 5, and 10 years. Results are depicted in figure 8.3 with boxplots. The SM (cause-specific Cox model, Fine-Gray) had by far the lowest miscalibration error at 2 years for disease progression (cause 1). The SM and then the PLANNCR original had the lowest miscalibration at 5 years (the SM and PLANNCR extended at 10 years). PLANNCR extended had the highest miscalibration error at 2 years, the second highest at 5 years and the lowest at 10 years (next to cause-specific Cox model for this time point). The RSFCR had the worst calibration at 5 and 10 years for the cumulative incidence of the event of interest.

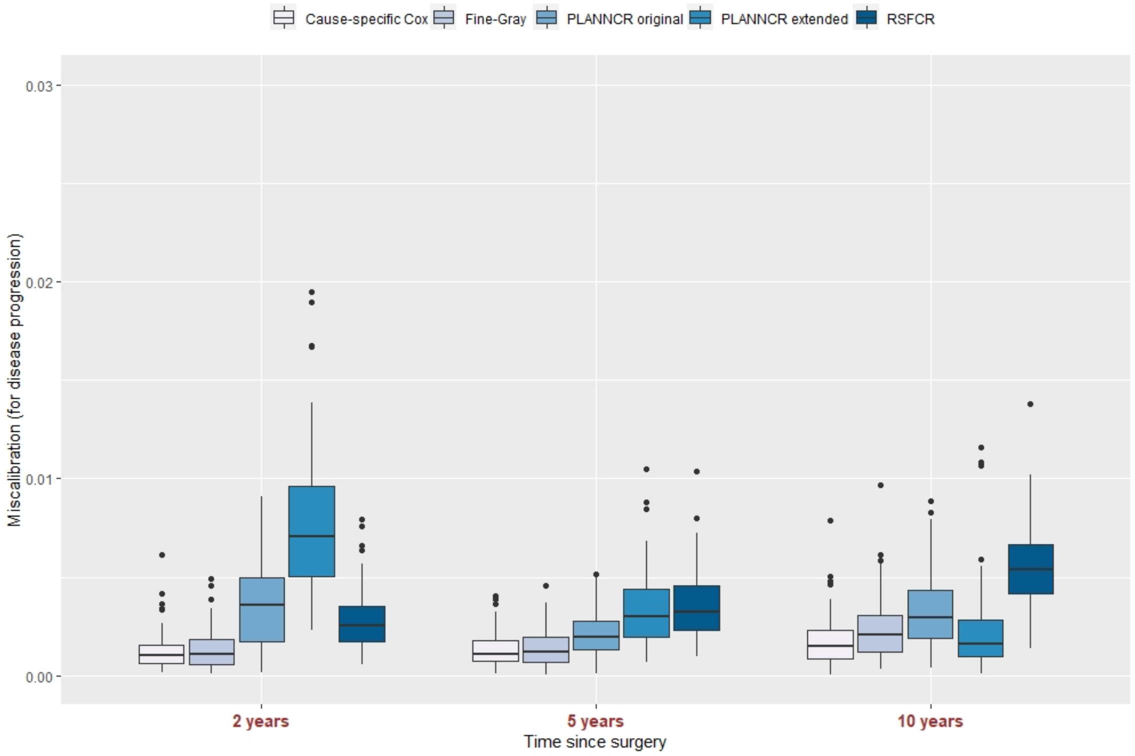

Figure 8.3: Miscalibration of cause-specific Cox model, Fine-Gray model, PLANNCR original, PLANNCR extended (tuned with Brier score at 5 years), and RSFCR at 2, 5, and 10 years for the event of interest: disease progression based on 100 validation datasets. Miscalibration was calculated as the mean squared error (MSE) between the observed and the predicted cumulative incidence event probabilities (for 4 groups).

The miscalibration plot for PLANNCR extended tuned with AUC at 5 years is available in Supplementary file 3 (figure S2). PLANNCR extended is less well calibrated compared to figure 8.3. This result was expected since in the supplementary figure the model was tuned for discrimination only (AUC at 5 years), whereas in figure 8.3 it was tuned taking into account both discrimination and calibration (Brier score at 5 years). Figures S4 and S6 show the miscalibration error for all five methods for the competing event (death). The cause-specific Cox model and the Fine-gray model had the lowest miscalibration error. RSFCR show a similar miscalibration error for death at 2 and 5 year and slightly worse error at 10 years. The two neural networks had the highest miscalibration error at any time point (distinct from the other three models). A tentative explanation of the higher PLANNCR miscalibration for the competing event is that it arises from heavier regularisation of the predicted death probabilities (for a given time point) resulting in a smaller spread of the predictions there. A solution to improve the calibration could be to tune the performance of PLANNCR (e.g. Brier score at 5 years) for the competing event. However, as disease progression was of major interest here, PLANNCR original and extended were both tuned for disease progression.

8.4 Discussion

To the best of our knowledge, this is the first study which compared SM with ML techniques for CRs in soft-tissue sarcoma. A total of 3826 retrospectively collected patients were analysed with high-grade eSTS based on nine prognostic factors (small/medium sample size, low-dimensional setting). The SM (cause-specific Cox, Fine-Gray) and the RSFCR used exact times to event whereas the neural networks (PLANNCR original, PLANNCR extended) required a data preparation into a long format where the exact time points were turned into L separate time intervals (years). The five methods predicted the cumulative incidence of disease progression (event of interest) and death (competing event) since the date of surgery.

The results showed that the ML models have similar performance to the SM in terms of Brier score and AUC at 2, 5, and 10 years for disease progression and death (95% confidence intervals overlapped). Predictive ability of PLANNCR extended was usually better than RSFCR and PLANNCR original especially for AUC. This means that PLANNCR extended had the ability to discriminate better between low and high risk groups of patients. Nevertheless, the SM were frequently better calibrated than the three ML techniques. Miscalibration of PLANNCR original and extended was more pronounced for the competing event. These findings are consistent with a simulation study of our group that compared the predictive performance of SNN (PLANN original and extensions) with Cox models for osteosarcoma data in a similar simple setting (250 or 1000 patients, five prognostic factors) [43]. Hence, more attention to model calibration (absolute predictive accuracy) is urgently needed for ML methods.

For this work, we sampled with replacement 100 times (bootstrapping) from the eSTS data to train the ML models. Then, the left out samples were used to internally validate all models' performance and obtain empirical 95% CIs (see figure 8.1). This can be an advantageous approach when the sample size is limited because it avoids decreasing the number of patients for model development / validation. However, it comes with a cost as this procedure is repeated multiple times and is therefore computationally expensive. The performance of all models was assessed with two time-dependent measures: Brier score (discrimination and calibration) and AUC (discrimination) at 2, 5, and 10 years, respectively. We chose the time-dependent AUC over the adaptation of Harrell's concordance index to the CRs setting [44, 45] - a global performance measure for discrimination - since the latter is not a proper measure for the evaluation of t-year predicted risks (see [46]).

Two regression models for CRs were applied for the comparison with ML techniques; the cause-specific hazard regression Cox and the Fine-Gray. The cause-specific Cox model might be better suited for addressing etiological questions, whereas the Fine-Gray for estimating the clinical prognosis of patients - which was the aim here [3, 5, 47]. Nonetheless, both SM were employed for a more comprehensive approach, providing similar results, and outperforming the ML models in calibration. Complex functional dependencies such as non-linear and non-additive effects were not investigated, which shows how effective the SM can be in simple settings (with small/medium sample size and limited number of predictors) despite they assume additivity of effects and proportionality of hazards over time. On the other hand, ML methods may be very flexible (no a priori modelling assumptions), but usually require (very) large datasets to ensure small overfitting of their developed clinical prediction models [48, 49].

Other ML-driven models have been recently proposed for survival analysis with CRs and their prognostic ability was compared with typical benchmarks such as the cause-specific Cox, Fine-Gray, and RSFCR. In 2017, Alaa and van der Schaar [50] proposed a non-parametric Bayesian model to jointly assess a patient's risk of multiple competing adverse events. The patient's cause-specific survival times are modelled as a function of the covariates using deep multi-task Gaussian processes. Bellot and van der Schaar [51] developed in 2018 a tree-based Bayesian mixture model for CRs. They constructed a hierarchical Bayesian mixture model through multivariate random survival forests and evaluated the importance of variables for each cause. Recently, a deep neural network (multiple hidden layers) was employed by Nagpal *et al.* called `deep survival machines` [52]. This is a parametric methodology to jointly learn a common deep non-linear representation of the input features. This network estimates separately the event distribution for each CR. Note that for this project, we only specified shallow neural networks (1 hidden layer) to avoid excessive danger of overfitting in this simple setting.

Focusing on the practical utility, the two SM have the advantage compared to three ML techniques examined. The latter require a substantial implementation time for data preprocessing, tuning of the parameters, and are computationally more intensive to run (in terms of hours here). At the same time model optimisation of PLANNCR is a delicate task which requires robust numerical methods and skillful use, else the network might converge in suboptimal minima in the error function [35]. From the three ML techniques, PLANNCR extended demanded more time and effort for training because of the larger number of tuning parameters (five versus two for PLANNCR original and RSFCR). On the contrary, the cause-specific Cox and Fine-Gray models do not require any hyperparameter tuning and offer a fast implementation.

Nowadays, the employment of ML is overhyped in some contexts of medicine due to the increased interest in applying modern techniques to create prediction models. Therefore, it is necessary to report prediction models powered by artificial intelligence completely and transparently to allow critical appraisal, reproducibility of the modelling steps and results by a wider audience, and to avoid research waste [14, 15, 53]. In general, a traditional regression approach may still provide more accurate predicted survival probabilities and prognostic performance compared to a state-of-the-art ML model, especially in non-complex medical settings (low-medium sample size, small number of predictors). In this instance, application of ML algorithms should only be motivated for exploration of the collected data.

In the future, it might be useful to compare the predictive ability of the cause-specific proportional hazard Cox model with the PLANNCR original / extended for time-dependent variables. The first method allows the inclusion of time-dependent covariates in standard software, and the second can naturally incorporate time-dependent covariates due to the essential data transformation into a long format for each patient. Moreover, the Fine-Gray and RSFCR can be extended to provide dynamic predictions with time-dependent covariates for CRs by creating a landmark dataset at a set of landmark time points t_{LM} [54]. Last but not least, it would be interesting to compare the SM and ML techniques regarding interpretation. Overall, SM offer a more straightforward interpretation via cause-specific hazard ratios, while PLANNCR can provide the shape of the cause-specific hazard function over time and covariates, and RSFCR the variable importance. Further research is needed on a common metric to directly compare all methods.

8.5 Conclusions

In this article, we discussed ML alternatives (PLANNCR original, PLANNCR extended, RSFCR) to SM (cause-specific Cox model, Fine-Gray) to build prognostic models for survival analysis with CRs in eSTS data with small/medium sample size and limited number of predictors (simple setting). Methods were compared in terms of discrimination and calibration. ML models reached an equivalent performance in terms of suitable predictive performance measures at 2, 5, or 10 years since surgery (95% confidence intervals overlapped), but the conventional regression models were generally better calibrated. Hence, more attention to calibration is needed. Modern ML-driven techniques are less practical as they require substantial implementation time (data preprocessing, hyperparameter tuning, computational intensity), whereas regression models are straightforward to use and can perform well without the additional workload of model training. Overall, complete and transparent reporting of all methods is required to allow critical appraisal, reproducibility, and avoid research waste. In our opinion, for non-complex real life data such as this, ML techniques should only be employed complementary to SM as exploratory tools of model's performance.

List of abbreviations

AUC, Area Under the Curve; CI, Confidence Interval; CIF, Cumulative Incidence Function; CRs, competing risks; eSTS, extremity soft-tissue sarcomas; IPCW, Inverse Probability of Censoring Weighting; KM, Kaplan-Meier; ML, machine learning; MSE, mean squared error; PLANN, partial logistic artificial neural network; PLANN-ARD, Partial logistic artificial neural network - automatic relevance determination; PLANNCR, Partial logistic artificial neural network for competing risks; PLANNCR-ARD, Partial logistic artificial neural network for competing risks - automatic relevance determination; ReLU, rectified linear unit; RSF, random survival forests; RSFCR, random survival forests for competing risks; SM, statistical models; SNNs, survival neural networks.

Declarations

Availability of data and materials

The clinical data used for this research project is private. The R-code developed to perform this analysis is provided in the following GitHub repository https://github.com/GKantidakis/SM-vs-ML-for-CRs. There, the reader will also find a zip file with R-codes, which is a comprehensive example of this analysis in publicly available R data for Follicular Cell Lymphoma (data ”follic”).

Funding statement

Georgios Kantidakis’s work as a Fellow at EORTC Headquarters was supported by a grant from the EORTC Soft Tissue and Bone Sarcoma Group and the Leiden University Medical Center (LUMC) Department of Medical Oncology. The funding sources had no role in the design of the study and collection, analysis, and interpretation of data or preparation of the manuscript.

Acknowledgements

This publication was supported by a donation from the Kom Op Tegen Kanker (Stand up to Cancer), the Flemish cancer society from Belgium. The authors would like to thank the members of the PERsonalised SARcoma Care (PERSARC) Study Group for the provision of the data used in this paper.

PERSARC Study Group

Ibtissam Acem, Will Aston, Han Bonenkamp, Ingrid M E Desar, Peter C Ferguson, Marta Fiocco, Hans Gelderblom, Anthony M Griffin, Dirk J Grünhagen, Rick L Haas, Andrew J Hayes, Lee M Jeys, Johnny Keller, Minna K Laitinen, Andreas Leithner, Katja Maretty-Kongstad, Rob Pollock, Anja Rueten-Budde, Myles Smith, Maria A Smolle, Emelie Styring, Joanna Szkandera, Per-Ulf Tunn, Jos A van der Hage, Robert J van Ginkel, Winan J van Houdt, Veroniek van Praag, Michiel van de Sande, Kees Verhoef, Madeleine Willegger, Reinard Windhager, Jay S Wunder, Olga Zaikova.

Online supplementary materials

The Supplementary files of this Chapter are available online at https://github.com/GKantidakis/Thesis_supplementary_materials/tree/main/Chapter8.

References

[1] H. Putter, M. Fiocco, and R. B. Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430, 2007. ISSN 1097-0258. doi: 10.1002/SIM.2712. URL https://onlinelibrary.wiley.com/doi/10.1002/sim.2712.

[2] R. Varadhan, C. O. Weiss, J. B. Segal, A. W. Wu, D. Scharfstein, and C. Boyd. Evaluating health outcomes in the presence of competing risks: A review of statistical methods and clinical applications. *Medical Care*, 48(6 SUPPL.):S96–105, 2010. ISSN 00257079. doi: 10.1097/MLR.0b013e3181d99107.

[3] R. B. Geskus. *Data Analysis with Competing Risks and Intermediate States*. Chapman and Hall/CRC, 1st edition, 2015. ISBN 9780367738051.

[4] Z. Zhang, G. Cortese, C. Combescure, R. Marshall, M. Lee, H. Lim, and B. Haller. Overview of model validation for survival regression model with competing risks using melanoma study data. *Annals of Translational Medicine*, 6(16):325, 2018. ISSN 23055839. doi: 10.21037/atm.2018.07.38.

[5] P. C. Austin, D. S. Lee, and J. P. Fine. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. *Circulation*, 133(6):601–609, 2016. ISSN 15244539. doi: 10.1161/CIRCULATIONAHA.115.017719.

[6] P. C. Austin and J. P. Fine. Accounting for competing risks in randomized controlled trials: a review and recommendations for improvement. *Statistics in Medicine*, 36(8):1203–1209, 2017. ISSN 10970258. doi: 10.1002/sim.7215.

[7] M. T. Koller, H. Raatz, W. Steyerberg, and M. Wolbers. Competing risks and the clinical community : irrelevance or ignorance ? *Statistics in Medicine*, 31(11-12):1089–1097, 2012. doi: 10.1002/sim.4384.

[8] E. L. Kaplan and P. Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958. doi: 10.2307/2281868.

[9] Z. Zhang. Survival analysis in the presence of competing risks. *Annals of Translational Medicine*, 5(3), 2016. ISSN 23055847. doi: 10.21037/atm.2016.08.62.

[10] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–220, 1972. URL http://www.jstor.org/stable/2985181.

[11] J. P. Fine and R. J. Gray. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999. ISSN 1537274X. doi: 10.1080/01621459.1999.10474144.

[12] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015. ISSN 20010370. doi: 10.1016/j.csbj.2014.11.005.

[13] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons. Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, 19(1):1–18, 2019. doi: 10.1186/s12874-019-0681-4.

[14] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *BMC Medicine*, 13(1), jan 2015. ISSN 17417015. doi: 10.1186/s12916-014-0241-z. URL http://www.biomedcentral.com/1741-7015/13/1.

[15] G. S. Collins and K. G. M. Moons. Reporting of artificial intelligence prediction models. *The Lancet*, 393(10181):1577–1579, apr 2019. ISSN 1474547X. doi: 10.1016/S0140-6736(19)30037-6. URL http://www.thelancet.com/article/S0140673619300376/fulltexthttp://www.thelancet.com/article/S0140673619300376/abstracthttps://www.thelancet.com/journals/lancet/article/PIIS0140-6736(19)30037-6/abstract.

[16] N. Shahid, T. Rappon, and W. Berta. Applications of artificial neural networks in health care organizational decision-making: A scoping review. *PLoS ONE*, 14(2):e0212356, 2019. ISSN 19326203. doi: 10.1371/journal.pone.0212356.

[17] P. Wang, Y. Li, and C. K. Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys*, 51(6):1–36, 2019. doi: https://doi.org/10.1145/3214306.

[18] E. Biganzoli, P. Boracchi, L. Mariani, and E. Marubini. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in medicine*, 17(10):1169–1186, 1998. doi: 10.1002/(sici)1097-0258(19980530)17:10<1169::aid-sim796>3.0.co;2-d.

[19] P. J. G. Lisboa, H. Wong, P. Harris, and R. Swindell. A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. *Artificial Intelligence in Medicine*, 28(1):1–25, 2003. ISSN 09333657. doi: 10.1016/S0933-3657(03)00033-2.

[20] G. Kantidakis, H. Putter, C. Lancia, J. de Boer, A. E. Braat, and M. Fiocco. Survival prediction models since liver transplantation - comparisons between Cox models and machine learning techniques. *BMC Medical Research Methodology*, 20(1):1–14, dec 2020. ISSN 14712288. doi: 10.1186/s12874-020-01153-1.

[21] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008. ISSN 1932-6157. doi: 10.1214/08-AOAS169. URL https://projecteuclid.org/journals/annals-of-applied-statistics/volume-2/issue-3/Random-survival-forests/10.1214/08-AOAS169.short.

[22] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://link.springer.com/article/10.1023/A:1010933404324.

[23] E. Biganzoli, P. Boracchi, F. Ambrogi, and E. Marubini. Artificial neural network for the joint modelling of discrete cause-specific hazards. *Artificial Intelligence in Medicine*, 37(2):119–130, 2006. doi: 10.1016/j.artmed.2006.01.004.

[24] P. J. G. Lisboa, T. A. Etchells, I. H. Jarman, C. T. C. Arsene, M. S. H. Aung, A. Eleuteri, A. F. G. Taktak, F. Ambrogi, P. Boracchi, and E. Biganzoli. Partial logistic artificial neural network for competing risks regularized with automatic relevance determination. *IEEE Transactions on Neural Networks*, 20(9):1403–1416, 2009. doi: 10.1109/TNN.2009.2023654.

[25] H. Ishwaran, T. A. Gerds, U. B. Kogalur, R. D. Moore, S. J. Gange, and B. M. Lau. Random survival forests for competing risks. *Biostatistics*, 15(4):757–773, 2014. ISSN 14684357. doi: 10.1093/biostatistics/kxu010.

[26] R Core Team. R: A Language and Environment for Statistical Computing, 2014. URL http://www.r-project.org/.

[27] A. J. Rueten-Budde, V. M. van Praag, M. A. J. van de Sande, M. Fiocco, W. Aston, H. Bonenkamp, D. Callegaro, P. D. S. Dijkstra, P. C. Ferguson, A. M. Griffin, A. Gronchi, D. J. Grünhagen, R. L. Haas, A. Hayes, L. M. Jeys, J. Keller, M. K. Laitinen, A. Leithner, K. Maretty-Kongstad, R. Pollock, F. Posch, M. Smith, M. A. Smolle, E. Styring, P. U. Tunn, J. A. van der Hage, Robert J. van G., W. J. van Houdt, K. Verhoef, M. Willegger, J. J. Willeumier, R. Windhager, J. S. Wunder, and O. Zaikova. External validation and adaptation of a dynamic prediction model for patients with high-grade extremity soft tissue sarcoma. *Journal of Surgical Oncology*, 123(4):1050–1056, 2021. ISSN 10969098. doi: 10.1002/jso.26337.

[28] L. Mariani, R. Miceli, M. W. Kattan, M. F. Brennan, M. Colecchia, M. Fiore, P. G. Casali, and A. Gronchi. Validation and adaptation of a nomogram for predicting the survival of patients with extremity soft tissue sarcoma using a three-grade system. *Cancer*, 103(2):402–408, jan 2005. ISSN 0008-543X. doi: 10.1002/CNCR.20778. URL https://pubmed.ncbi.nlm.nih.gov/15578681/.

[29] V. M. van Praag, A. J. Rueten-Budde, L. M. Jeys, M. Laitinen, R. Pollock, W. Aston, J. A. van de Hage, P. D. S. Dijkstra, P. C. Ferguson, A. M. Griffin, J. J. Willeumier, J. S. Wunder, M. A.J. van de Sande, and M. Fiocco. A prediction model for treatment decisions in high-grade extremity soft-tissue sarcomas: Personalised sarcoma care (PERSARC). *European Journal of Cancer*, 83:313–323, 2017. ISSN 18790852. doi: 10.1016/j.ejca.2017.06.032. URL http://dx.doi.org/10.1016/j.ejca.2017.06.032.

[30] D. Callegaro, R. Miceli, S. Bonvalot, P. Ferguson, D. C. Strauss, A. Levy, A. Griffin, A. J. Hayes, S. Stacchiotti, C. Le Pèchoux, M. J. Smith, M. Fiore, A. P. Dei Tos, H. G. Smith, C. Catton, P. G. Casali, J. S. Wunder, and A. Gronchi. Impact of perioperative chemotherapy and radiotherapy in patients with primary extremity soft tissue sarcoma: retrospective analysis across major histological subtypes and major reference centres. *European Journal of Cancer*, 105:19–27, 2018. ISSN 18790852. doi: 10.1016/j.ejca.2018.09.028.

[31] P. J. Woll, P. Reichardt, A. Le Cesne, S. Bonvalot, A. Azzarelli, H. J. Hoekstra, M. Leahy, F. Van Coevorden, J. Verweij, P. C. W. Hogendoorn, M. Ouali, S. Marreaud, V. H. C. Bramwell, and P. Hohenberger. Adjuvant chemotherapy with doxorubicin, ifosfamide, and lenograstim for resected soft-tissue sarcoma (EORTC 62931): A multicentre randomised controlled trial. *The Lancet Oncology*, 13(10):1045–1054, 2012. ISSN 14702045. doi: 10.1016/S1470-2045(12)70346-7. URL http://dx.doi.org/10.1016/S1470-2045(12)70346-7.

[32] D. J Stekhoven and P. Bühlmann. Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012. doi: 10.1093/bioinformatics/btr597.

[33] J. D. Holt. Competing risk analyses with special reference to matched pair experiments. *Biometrika*, 65(1): 159–165, apr 1978. ISSN 0006-3444. doi: 10.1093/BIOMET/65.1.159.

[34] O. O.; Aalen and S. Johansen. An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations. *Scandinavian Journal of Statistics*, 5(3):141–150, 1978. URL https://www.jstor.org/stable/4615704.

[35] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006. ISBN 978-0-387-31073-2.

[36] F. Chollet. keras, 2015. URL https://github.com/keras-team/keras.

[37] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2nd edition, 2009. ISBN 978-0-387-84857-0. doi: 10.1007/978-0-387-84858-7. URL http://link.springer.com/10.1007/978-0-387-84858-7.

[38] P. Blanche, C. Proust-Lima, L. Loubère, C. Berr, J. F. Dartigues, and H. Jacqmin-Gadda. Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics*, 71(1):102–113, 2015. ISSN 15410420. doi: 10.1111/biom.12232.

[39] P. Blanche, J. F. Dartigues, and H. Jacqmin-Gadda. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine*, 32(30):5381–5397, 2013. doi: 10.1002/sim.5958.

[40] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999. doi: 10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5. URL http://www.ncbi.nlm.nih.gov/pubmed/10474158.

[41] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan. Assessing the performance of prediction models: A framework for some traditional and novel measures. *Epidemiology*, 21(1):128–138, jan 2010. ISSN 10443983. doi: 10.1097/EDE.0b013e3181c30fb2. URL https://pubmed.ncbi.nlm.nih.gov/20010215/.

[42] T. A. Gerds, P. K. Andersen, and M. W. Kattan. Calibration plots for risk prediction models in the presence of competing risks. *Statistics in Medicine*, 33(18):3191–3203, 2014. ISSN 10970258. doi: 10.1002/sim.6152.

[43] G. Kantidakis, E. Biganzoli, H. Putter, and M. Fiocco. A Simulation Study to Compare the Predictive Performance of Survival Neural Networks with Cox Models for Clinical Trial Data. *Computational and Mathematical Methods in Medicine*, 2021:1–15, 2021. ISSN 1748-670X. doi: 10.1155/2021/2160322.

[44] M. Wolbers, M. T. Koller, J. C. M. Witteman, and E. W. Steyerberg. Prognostic models with competing risks methods and application to coronary risk prediction. *Epidemiology*, 20(4):555–561, 2009. ISSN 10443983. doi: 10.1097/EDE.0b013e3181a39056.

[45] M. Wolbers, P. Blanche, M. T. Koller, J. C. M. Witteman, and T. A. Gerds. Concordance for prognostic models with competing risks. *Biostatistics*, 15(3):526–539, 2014. ISSN 14684357. doi: 10.1093/biostatistics/kxt059.

[46] P. Blanche, M. W. Kattan, and T. A. Gerds. The c-index is not proper for the evaluation of t-year predicted risks. *Biostatistics*, 20(2):347–357, 2019. ISSN 14684357. doi: 10.1093/biostatistics/kxy006.

[47] A. Tullio, A. Magli, E. Moretti, and F. Valent. Why we should take care of the competing risk bias in survival analysis: A phase II trial on the toxicity profile of radiotherapy for prostate cancer. *Reports of Practical Oncology and Radiotherapy*, 24(6):511–519, 2019. ISSN 15071367. doi: 10.1016/j.rpor.2019.08.001. URL https://doi.org/10.1016/j.rpor.2019.08.001.

[48] T. Van Der Ploeg, P. C. Austin, and E. W. Steyerberg. Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology*, 14(1):1–13, 2014. ISSN 14712288. doi: 10.1186/1471-2288-14-137.

[49] R. D. Riley, J. Ensor, K. I. E. Snell, F. E. Harrell, G. P. Martin, J. B. Reitsma, K. G. M. Moons, G. Collins, and M. Van Smeden. Calculating the sample size required for developing a clinical prediction model. *The BMJ*, 368(March):1–12, 2020. ISSN 17561833. doi: 10.1136/bmj.m441. URL http://dx.doi.org/doi:10.1136/bmj.m441.

[50] A. M. Alaa and M. Van Der Schaar. Deep multi-task Gaussian processes for survival analysis with competing risks. *Advances in Neural Information Processing Systems*, pages 2326–2334, 2017. ISSN 10495258. URL http://medianetlab.ee.ucla.edu/papers/Alaa-Deep-Competing-Risk.pdf.

[51] A. Bellot and M. van der Schaar. Tree-based Bayesian mixture model for competing risks. *International Conference on Artificial Intelligence and Statistics, PMLR 2018*, pages 910–918, 2018. URL http://proceedings.mlr.press/v84/bellot18a/bellot18a.pdf.

[52] C. Nagpal, X. Li, and A. Dubrawski. Deep Survival Machines: Fully Parametric Survival Regression and Representation Learning for Censored Data with Competing Risks. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3163–3175, 2021. ISSN 21682208. doi: 10.1109/JBHI.2021.3052441.

[53] P. Dhiman, J. Ma, C. A. Navarro, B. Speich, G. Bullock, J. A. A. Damen, S. Kirtley, L. Hooft, R. D. Riley, B. Van Calster, K. G. M. Moons, and G. S. Collins. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. *Journal of Clinical Epidemiology*, 138:60–72, 2021. ISSN 18785921. doi: 10.1016/j.jclinepi.2021.06.024. URL https://doi.org/10.1016/j.jclinepi.2021.06.024.

[54] M. A. Nicolaie, J. C. van Houwelingen, T. M. de Witte, and H. Putter. Dynamic prediction by landmarking in competing risks. *Statistics in Medicine*, 32(12):2031–2047, 2013. ISSN 02776715. doi: 10.1002/sim.5665.