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

Kantidakis, G.

1 General introduction

1.1 Research in context

This thesis springs from an interdisciplinary collaboration between the European Organisation for Research and Treatment of Cancer (EORTC), the Mathematical Institute of Leiden University, and the Leiden University Medical Center (LUMC) Department of Medical Oncology to perform statistical analyses for the European Organisation for Research and Treatment of Cancer - Soft Tissue and Bone Sarcoma Group (EORTC - STBSG) and, in addition, to investigate the potential of survival prediction models with machine learning compared with traditional statistical benchmarks for sarcoma and non-sarcoma clinical data. Methodology developed during this PhD project is applied to data of the STBSG database, the PERsonalised SARcoma Care (PERSARC) Study Group, and to liver transplantation data of the Scientific Registry of Transplant Recipients (SRTR).

1.1.1 Part I: Clinical trials in soft-tissue sarcomas

This part provides modern efficacy thresholds to design new phase II clinical trials for common histotypes of locally advanced or metastatic soft-tissue sarcoma (STS) patients. Research is timely as well-established values were reported in 2002 by the EORTC - STBSG [1], and thus an update is necessary. A further goal here is to identify high-risk patient populations by investigating the prognostic significance of bone metastasis in STS.

Soft-tissue sarcomas background

STS are very heterogeneous rare malignancies developed in cells of the connective or soft supporting tissues of the body such as muscles, nerves, blood vessels, fatty and fibrous tissues [2]. They account for about 1% of all adult tumours and commonly affect arms, legs or trunk. Over the years, more than 100 histologic subtypes have been recognised with widely varying presentation, sensitivity to treatment, and long-term outcomes [3]. Prognosis of advanced STS is poor with median progression-free survival (PFS) of about 6 months, i.e., for first-line systemic

therapy with doxorubicin plus ifosfamide, and a median overall survival (OS) just above one year [4].

The most common site of metastasis is the lungs [3–5]. Depending on the histology, STS metastasise sometimes to lymph nodes, bones, liver, and brain [6]. Other organs may also be affected depending on the sarcoma entity. Skeletal metastasis is part of the natural history affecting prognosis and quality of life of patients with advanced/metastatic disease as a pathological fracture may occur in 20–30% of them together with other skeletal-related events (hypercalcaemia, spinal cord compression, and need for surgery or palliative radiotherapy for refractory pain) [7].

The mainstay of disease management for inoperable locally advanced or metastatic population is systemic treatment, which is often palliative to delay progression and severe morbidity [3, 5]. Gastrointestinal stromal tumours (GIST) are generally considered separately because of their responsiveness to receptor tyrosine kinase inhibitors, predominantly imatinib [8]. The selection of treatment is based on clinical performance, age, histology, disease biology, patient preferences, and availability of novel treatments and studies [5, 9]. For first-line treatment of (non-GIST) STS, doxorubicin alone or in combination with ifosfamide is considered the most active drug (combination) for many years [4]. Beyond first-line, commonly used drugs depending on the histologic subtype are gemcitabine with/without docetaxel, trabectedin, pazopanib, and dacarbazine with/without gemcitabine which have been associated with a progression-free survival (PFS) benefit in doxorubicin-treated patients [10]. Eribulin is the only drug to have shown a survival benefit although curiously no benefit in PFS.

The most common histotypes are leiomyosarcoma ($\sim$ 20%), liposarcoma ($\sim$ 20%), undifferentiated pleomorphic sarcoma ($\sim$ 15%), and synovial sarcoma ($\sim$ 6%), with the remaining histotypes being individually rarer [11]. In the previous decades, STS studies were designed based on a one-size-fits-all principle mixing several histologic subtypes. Since 2002, well-established criteria have been widely used to design new phase II studies for all STS (> 420 citations) [1]. We refer to them with detail in the following section. Nevertheless, more recently, there is a clear trend towards histology-specific tailored research as there is a need to better diversify the eligibility criteria of clinical trials [3, 5]. Two meta-analyses projects are conducted to provide an update to reflect current treatment practices, and to evaluate these thresholds separately for the most prevalent STS subtypes to aid the design of histology-specific trials.

Historical benchmarking for STS clinical trials

In 2002, Van Glabbeke and colleagues published in the European Journal of Cancer (EJC) a pooled analysis on behalf of the EORTC - STBSG, estimating progression-free rates for various groups of STS patients who participated in phase II trials [1]. The clinical trials database of EORTC was used to estimate reference values for conducting new phase II trials with progression-free rate as the primary endpoint. Drugs were separated as active or inactive (pooling in distinct groups drugs which had demonstrated activity/inactivity) and activity thresholds were calculated separately for first-line and pre-treated patients.

Three groups of patients were defined:

(a) Patients treated with a first-line active drug or combination (used for the activity threshold in first-line patients - namely P_1).
(b) Patients with drugs that did not demonstrate substantial anti-tumor activity at the tested dose and schedule (used for non-activity threshold in pre-treated patients - namely P_0).
(c) Patients treated with an active drug after failure of an anthracycline-containing regimen (used to provide P_1 in pre-treated patients).

Table 1.1 shows the historical benchmarks for first-line and pre-treated advanced or metastatic STS patients. The authors suggested for the first-line treatment a rate of 38-56% at 6 months as a reference value (depending on the histology), and for second-line or further a 3-month rate $\geq$ 40% for drug activity and $\leq$ 20% for inactivity (for any STS subgroup). It is important to note that most of the phase II trials were conducted before the classification of GIST as a separate entity.

Line of treatment	Treatment regimen	Number of patients	Histology specific	3-month rate	6-month rate
First-line	Active	1154	Yes	57-77 %	38-56 %
Pre-treated	Inactive	234	No	21 %	8 %
Pre-treated	Active	146	No	39 %	14 %

Table 1.1: Progression-free rates for first line or pre-treated advanced / metastatic STS patients based on the paper by Van Glabbeke *et al.* in 2002 [1].

Meta-analysis of proportions

Meta-analysis is a statistical method that combines results of multiple scientific studies all addressing the same research question. The basic situation in meta-analysis is dealing with k studies in which a parameter of interest is estimated. In this thesis, two meta-analyses concerning drugs / drug combinations are performed from clinical trials with advanced or metastatic STS patients identified via a literature review in PubMed. Our goal is to update the historical benchmarks by Van Glabbeke *et al.* (2002) [1] for leiomyosarcoma, liposarcoma, and synovial sarcoma patients. A random effects model is introduced for each drug / drug combination to estimate the overall effect estimate (i.e., PFS proportion) per line of treatment (first line vs pre-treated) for the aforementioned STS subtypes. This approach is commonly referred to as the "DerSimonian and Laird method" in medical research [12, 13].

A random effects model allows the included studies (drug / drug combinations per treatment line) to have true effect sizes normally distributed. Such a model takes both within- and between-studies variance (defined as τ^2) into account. Parameter τ^2 is calculated using the method of moments (DerSimonian and Laird method) [14]. The overall effect size (i.e., the summary proportion) is estimated as a weighted average of the observed effect sizes for each drug (individual study). The weighting for each study is the inverse of its total variance (sum of the within- and the between-study variance). A larger study is given more weight so its effect size has greater impact on the overall effect. Using the logit (log-odds) transformation [15], the effect size for a

study i is written as $y_i = \ln\left(\frac{p_i}{1-p_i}\right)$, with sampling variance $v_i = \frac{1}{n_i p_i} + \frac{1}{n_i(1-p_i)}$, and the inverse variance weight $w_i = \frac{1}{v_i+\tau^2}$, where p and n are the proportion and sample size, respectively. Then, the weighted mean proportion and sampling variance can be computed as:

$$y_{mean} = \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i}, \tag{1.1}$$

and

$$v_{mean} = \frac{1}{\sum_{i=1}^{k} w_i}, \tag{1.2}$$

where k is the total number of studies (here drugs or drug combinations per treatment line).

The overall heterogeneity between studies is provided by the I^2 statistic (variability between the study-specific effect sizes which cannot be explained by random variation) [14]. Note that, equivalently, a PFS rate is the PFS proportion*100.

Outline: research objectives of part I

The historical benchmarking work by Van Glabbeke *et al.* [1] on behalf on the EORTC - STBSG estimated progression-free rates for various groups of STS patients who participated in phase II trials of the EORTC. These thresholds have been used extensively (currently > 420 citations) to design new studies for all STS or for specific histology subgroups but remain unchanged since 2002. Therefore, it is not difficult to see the first gap that this thesis aims to cover. To elaborate on this, the 2002 thresholds should not only be updated but also be evaluated separately for the most prevalent STS subtypes to reflect modern clinical practice, as future agents should perform better than currently available standards of care, and there is a need to better diversify the eligibility criteria of clinical trials. We performed an extensive in-house literature search to identify all phase II or subsequent clinical trials of advanced or metastatic STS (2003 to 2018), thus documenting the current landscape. Because of the substantial heterogeneity among clinical trials, it was decided to focus first on leiomyosarcoma (LMS) - the most commonly occurring STS subtype in our literature review. The primary endpoint of interest is progression-free survival rate (PFSR), which is nowadays a preferred and more frequently reported endpoint than progression-free rates (censoring non-disease-related death). Hereto, in **Chapter** 2 [16], a meta-analysis is performed to provide a new benchmark for designing phase II studies of advanced or metastatic LMS patients using PFSR.

Historically, the majority of the STS trials have been designed with a one-size-fits-all principle mixing several histologic subtypes [3, 5]. However, our recent study [16] is in accordance with a trend towards histology-specific tailored research. In **Chapter** 3, a second meta-analysis is performed for advanced or metastatic liposarcoma or synovial sarcoma - the second and third most common subtypes in our literature review - to propose new efficacy thresholds for the design of future histology-specific phase II trials with these entities.

As a further step in our work, we felt it would be of interest to identify high-risk patient populations in STS clinical trials examining patient characteristics from EORTC - STBSG clinical

trials. Our group has recently conducted some research in this direction [17, 18]. Lindner *et al.* confirmed and identified prognostic factors for first-line chemotherapy patients with lung metastases from 15 EORTC - STBSG trials [17]. Younger *et al.* analysed outcomes of elderly patients (age $\geq$ 65) with a pooled analysis of 12 EORTC - STBSG trials. They found that elderly patients with metastatic STS treated with first-line chemotherapy were largely under-represented in these trials [18]. Skeletal metastasis is part of the natural history affecting the prognosis and quality of life of patients with advanced/metastatic STS as a pathological fracture may occur in 20–30% of them together with other skeletal-related events [7]. Hence, as an extension in **Chapter** 4 [19], we investigate whether, and if so, to what extent, bone metastases at presentation affect the outcome (regarding OS and PFS) of patients with advanced or metastatic disease, performing a pooled analysis of 5 trials from the EORTC - STBSG database. If presence of bone metastases at study entry is an important risk factor per line of treatment (first vs second or further), then stratification should be considered in randomised trials. An additional goal is to identify which metastatic organ site (among liver, lymph node, lung, soft-tissue, or other) has the largest impact for patient's OS/PFS combined with skeletal metastasis at diagnosis in this database (per treatment line).

1.1.2 Part II: Statistical models versus machine learning to predict survival for sarcoma and non-sarcoma clinical data

This part focuses on the comparison of statistical models (SM) with machine learning (ML) techniques. Nowadays, there is a growing interest by the medical community in applying ML to predict clinical outcomes [20]. In this part of the thesis, new developments regarding ML for time-to-event data proposed by the author are discussed. A comparison with traditional benchmarks for real-life clinical data (small/medium or large sample sizes, low- or high-dimensional settings) is performed.

Survival analysis

Survival analysis (also referred to as time-to-event analysis) is used to estimate the lifespan of a particular population under study [21, 22]. Survival analysis is one of the primary statistical methods for analysing medical data concerning time until the occurrence of an event of interest (such as death, disease-progression, heart attack, organ failure etc.) occurs. These kind of data are often right-censored; they can be seen as a specific type of missing data in which time to the event of interest may be unobserved, either due to subjects being lost to follow-up, or due to time limitations such as study termination (called administrative censoring). The presence of censored observations makes the analysis of these data challenging requiring modifications to the traditional approaches aiming at using all available information (including also censored observation). In this thesis, survival analysis is performed for different endpoints such as overall survival (time to death from any cause since the date of surgery) and overall graft-survival (time between liver transplantation and graft failure or death).

The most popular statistical benchmark for right-censored data is the Cox proportional hazards model (traditional benchmark) [23], which is typically employed to estimate the effect of risk factors on the outcome of interest. This model assumes that each covariate has a multiplicative constant over time effect on the hazard function. Data with sample size n consist of independent observations from the triplet (T, D, X) i.e., $(t_1, d_1, x_1), \cdots, (t_n, d_n, x_n)$. For the i^{th} individual, t_i is the survival time, d_i the censoring indicator ($d_i = 1$ if the event occurred and $d_i = 0$ if the observation is right-censored) and x_i is the vector of predictors $(x_1, \cdots, x_p)$. The hazard function of the Cox model with time-fixed covariates is specified as:

$$h(t|X) = h_0(t) \exp(X^T \boldsymbol{\beta}), \tag{1.3}$$

where $h(t|X)$ is the hazard at time t given predictor values X, $h_0(t)$ is an arbitrary baseline hazard, and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ is a parameter vector. The baseline hazard h_0 is the hazard if all predictors are equal to zero. Formula (1.3) shows that the hazard function depends on chosen predictors $(x_1, \cdots, x_p)$ expressed through the size of the coefficients $(\beta_1, \cdots, \beta_p)$. For example, suppose that we are interested in modelling the effect size of an experimental treatment A versus the standard of care for OS in STS patients adjusted for the impact of other variables $(x_2, \cdots, x_p)$. Then, a hazard ratio of 0.70 versus the standard of care means that treatment A has a protective effect (30% reduction in the risk of death), given that the other covariates remain constant.

Survival analysis with competing risks

A competing risk (CR) is an event whose occurrence precludes the occurrence of an event of interest (for instance death may preclude the occurrence of disease relapse) [24, 25]. Typically in clinical applications for survival data, if several types of events occur, a model describing progression for each of the CRs is needed [26, 27]. CRs are unlikely to be independent as the biology suggests at least some dependence between them (a competing event can alter the probability of occurrence of the event of interest). In several chronic diseases attributable to aging and frailty such as cancer, chronic heart failure, or dementia, study populations are susceptible to CRs [28].

The most popular non-parametric approach (does not make any assumptions about the sample characteristics) to estimate survival is the Kaplan-Meier (KM) [29]. However, in the presence of CRs, the KM methodology overestimates the probability of failure which might lead to over-treatment of patients [24, 30]. Different SM have been developed to model the effect of covariates on cumulative incidence (absolute risk) of an event in the presence of CRs such as the cause-specific hazard regression model a competing risks analogue of a Cox proportional hazards model [23], and the Fine-Gray sub-distribution hazards regression model [31]. 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 competing event. The latter models the covariate effects directly on the cumulative incidence function (a proper summary statistic for CRs) over time reporting on the sub-distribution hazard ratio [30].

Survival prediction and performance measures

Prediction is the cornerstone of clinical practice. It is inherent in every diagnosis about the course of an illness. At the same time, every therapeutic medication stimulates prediction regarding a response to treatment (prognosis; a forecast of the medical outcome) [32]. The number of studies that focus on prediction models is rapidly expanding in the medical field. Survival prediction aims to predict the future survival risk of patients based on a set of predictive features (also called prognostic factors, risk factors, or covariates; e.g., gender, age, treatment group etc.). Hence, survival prediction is conducted using prognostic models.

The performance of prediction models can be assessed in different ways. The most popular measure of model performance in a survival context is the concordance index (also called the C-index) [33], which computes the proportion of pairs of observations for which the survival times and model predictions' (risk of the outcome e.g., death) order are concordant taking into account censoring. Patients who experience the event of interest should have higher risk than patients who do not experience the event. It takes values typically in the range 0.5 - 1 with higher values denoting higher ability of the model to discriminate between patients with different risk; 0.5 indicating no discrimination (similar to flipping a coin). In real-world data, a C-index above 0.60 indicates a clinically acceptable model, above 0.70 shows a good model, and above 0.80 an excellent model in terms of discrimination.

The C-index provides a rank statistic between the observations that is not time-dependent. Following van Houwelingen and le Cessie [34] a time-dependent prediction error is defined as

$$\textit{Brier}(y, \hat{S}(t_0|x)) = (y - \hat{S}(t_0|x))^2, \tag{1.4}$$

where $\hat{S}(t_0|x)$ is the model-based probabilistic prediction for the survival of an individual beyond t_0 given the predictor x, and $y = 1\{t > t_0\}$ is the actual observation ignoring censoring. To assess the performance of a prediction rule in actual data, censored observations before time t_0 must be considered. To calculate Brier Score when censored observations are present, Graf proposed the use of inverse probability of censoring weighting [35]. An estimate of the average prediction error of the prediction model $\hat{S}(t|x)$ at time $t = t_0$ is as follows

$$Err_{Score}(\hat{S}, t_0) = \frac{1}{n}\sum_i 1\{d_i = 1 \vee t_i > t_0\}\frac{Score(1\{t_i > t_0\}, \hat{S}(t_0|x_i))}{\hat{C}(\min(t_i-, t_0)|x_i)}, \tag{1.5}$$

where $\frac{1}{\hat{C}(\min(t_i-,t_0)|x_i)}$ is the probability of censoring weighting (IPCW) and $Score$ is the Brier Score for the prediction model. The Brier Score is a more complete performance measure than C-index as it takes into account both model discrimination and calibration (how well the predicted survival probabilities match the expected). It ranges typically from 0 to 0.25 with a lower value meaning a smaller prediction error.

The Brier score is calculated at different time-points. An overall measure of prediction error is the Integrated Brier Score (IBS) which can be used to summarise the prediction error over the whole range up to the time horizon $\int_0^{t_{hor}} Err_{Score}(\hat{S}, t_0)dt_0$ [36]. IBS provides the cumulative

prediction error up to t_{hor} at all available times ($t^* = 1, 2, \cdots, t_{hor}$ years) and takes values in the same range as the Brier score.

An essential measure of model performance is calibration which refers to the agreement between observed outcomes and predictions (absolute predictive accuracy) [37]. In particular, one way that is used in this thesis to assess model calibration is as follows: the predicted survival probabilities are estimated, and the clinical data are split into $m = 4$ equally sized groups based on the quantiles of the predicted probabilities. Quantiles were chosen over for instance deciles to avoid any computational issues. Then, the observed survival probabilities are calculated using KM methodology [29]. Miscalibration for each group is defined in terms of mean squared error (MSE) of the difference between the observed and the predicted survival probabilities:

$$MSE(t_0) = \frac{\sum_{m=1}^{4} \left[S_{KM}^{(m)}(t_0) - \hat{S}^{(m)}(t_0)\right]^2}{4}, \tag{1.6}$$

at t_0 years.

Reporting discrimination and calibration is always important for a prediction model [37]. In the presence of CRs, these measures have to be modified to address competing events [38–40].

Machine learning in medicine

ML - a subfield of artificial intelligence (a wide-ranging branch of computer science concerned with building smart machines capable of performing tasks that typically require human intelligence) - is the study of computer algorithms that can automatically learn from data in order to make predictions rather than being explicitly programmed with rules to do so [20, 41, 42]. It arises at the intersection of computer science with mathematics, statistics, and other disciplines including medicine. The long-term promise of ML in medicine is that the care of each patient will be informed by the wisdom contained in the decisions made by nearly all clinicians, and the outcomes of potentially millions of patients [42]. Every diagnosis, management decision, or therapy could be personalised not only based on all known information about a patient, but also incorporating lessons learnt from the medical community. Still, these days, little in healthcare is driven by ML despite the early claims (even half a century ago) that computers will augment and, in some cases, largely replace the intellectual functions of physicians [43].

From last decade, there is a growing interest in applying ML for prediction by medical researchers and clinicians sparked by the improvements in computer processing power, the storage of large amount of patient data in electronic databases, and the development of new algorithms [20]. Successful applications of ML in medicine include for instance automated interpretation of the electrocardiogram for a cardiologist (diagnosis selection from a limited list of options), or automated detection of a lung nodule from a chest x-ray for a radiologist [41]. These are labelled tasks where the computer is approximating what a trained physician can do with high accuracy. ML has also been used to find naturally occurring unlabeled patterns or groupings within the data. One successful example from the field of genomics is the identifi-

cation of an eosinophilic subtype of asthma [44], which led to the discovery of a novel target therapy called lebrikizumab to tackle it [45]. An excellent overview of current and future potential applications of ML in the field of oncology can be found in Ref. [46]. The authors focus on the role of ML for risk assessment, lesion detection and grading, imaging, staging, prognosis and therapy response of cancer patients.

Machine learning for survival analysis

Amongst ML techniques, artificial neural networks have been a common choice of methodology in healthcare. Over the years, neural networks and other ML techniques have been developed and adapted to survival data. Wang *et al.* in 2019 provide a comprehensive overview of conventional and modern approaches for right-censored time-to-event data [47]. 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) [48]. 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) to select the most relevant prognostic factors [49]. In this work, we propose ”PLANN extended” which provides new specifications to the original PLANN by Biganzoli in terms of architecture (i.e., time interval inputs, hyperparameters, activation functions). 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) [50]. RSF adapt Breiman’s random forest method [51] by using a collection of survival trees. Note that survival trees and forests are popular non-parametric (no assumptions are made about the characteristics of the sample) alternatives to the Cox model for time-to-event analysis.

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 [52]. The time (in discrete time intervals) is used 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) [53]. Here, we develop ”PLANNCR extended” which provides new specifications to the original PLANNCR by Biganzoli in terms of architecture (i.e., hyperparameters, activation functions). Ishwaran et al. extended RSF for CRs (RSFCR) in 2014 to estimate the cumulative incidence function of competing events [54].

The split sample approach

To develop and evaluate a ML prediction model, the ”split sample” approach is typically employed [55]. The original dataset is split randomly into two complementary parts: training data (2/3), and testing data (1/3). The training data are used for model development and the testing data for the evaluation of the final model's performance. This division of the data is required to limit overfitting (an analysis excessively tailored to a particular set of data) which can affect the model's ability to predict on new data reliably (model generalizability). An illustration of the procedure is provided in figure 1.1.

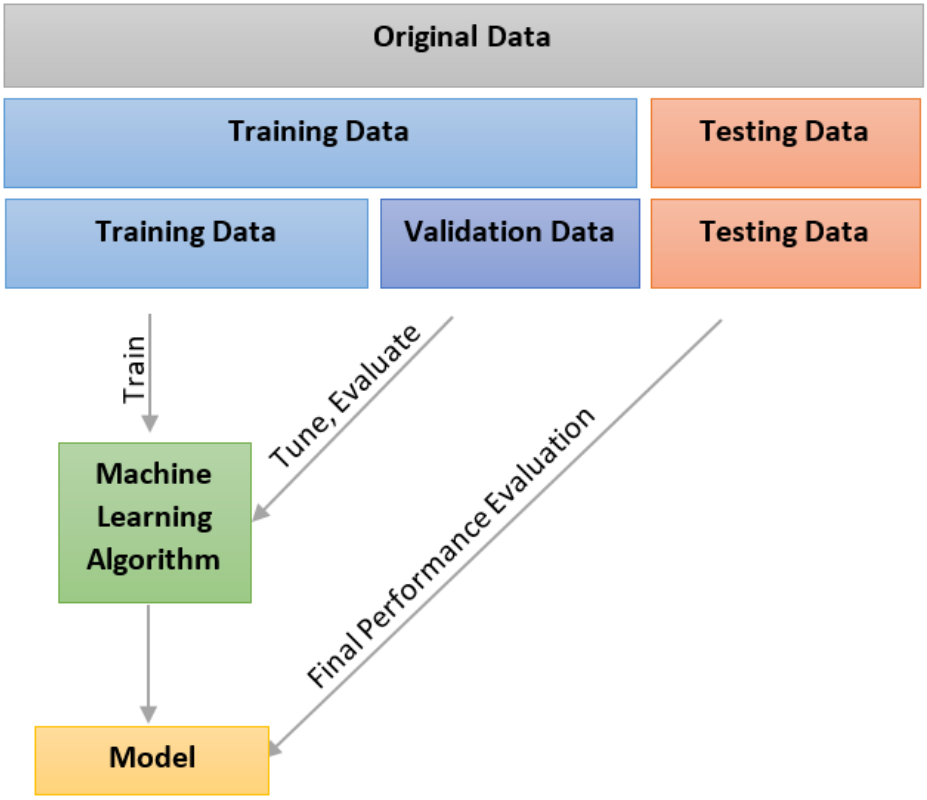

Figure 1.1: General strategy for model development and evaluation of a ML prediction model.

ML algorithms have a set of parameters (called the ”hyper-parameters”) whose values control the learning process. The prefix ”hyper” suggests that these are top level parameters which control the resulting model parameters derived via training with data. Different model training algorithms require different hyper-parameters. To tune the ML model (to find the best model or hyper-parameters for a given task), one common practice is to perform a 5-fold cross validation on the training data. Training data are divided into 5 folds. Each time 4 folds are used to train a model, and the remaining fold is used to validate (evaluate) its performance on the training data. This procedure is repeated with all combinations of folds used as training or validation data. In this thesis, tuning of hyper-parameters is done using grid search (a tuning technique that attempts to compute the optimum values of hyper-parameters using an exhaustive search on specific parameter values).

The partial logistic artificial neural network

Artificial neural networks were inspired from the human brain activity and more specifically from the neurons that transmit information between different areas of the brain. They have a layered structure based on a collection of units called nodes (or neurons) for each layer. The input layer fetches the signals and passes them to the next layer which is called "hidden" after the application of a non-linear transformation (activation) function. There might be a stack of hidden layers next to each other that connect with the previous layer and transmit signals towards the output layer. Connections between the artificial neurons of different layers are called edges. Artificial neurons and edges have a weight which adjusts through training increasing or decreasing the strength of each connection's signal. To train the network, a target is defined in the output layer which is the observed outcome for each individual. The simplest form of a feed forward ANN has the input layer, a single hidden layer and the output layer.

PLANN [48] is a SNN with a single output node (unit) which estimates discrete hazards as conditional probabilities of failure. It can be used for flexible modelling of survival data, as it relaxes the proportional hazards assumption in intervals. To implement this approach, survival times are discretized into a set of $l = 1, \cdots, L$ non-overlapping intervals $A_l = (\tau_{l-1}, \tau_l]$, with mid-points α_l (time variable), $0 = \tau_0 < \tau_1 < \cdots < \tau_l$ a set of pre-defined time points (usually years) and l_i the last observation interval for subject i. Data have to be transformed into a longitudinal format where the time variable is added as part of the input features next to the prognostic factors. On the training data each subject is repeated for the number of intervals being observed, whereas on the test data for all time intervals. By adding hidden layers, PLANN naturally models time-dependent interactions and non-linearities between the prognostic features. Activation function of both hidden and output layers is the logistic (sigmoid) function:

$$f(\theta) = \frac{1}{1 + e^{-\theta}}. \tag{1.7}$$

The output node is one large target vector with 0 if the event did not occur and 1 if the event of interest occurred in a specific time interval (due to the necessary data transformation). PLANN provides the discrete conditional probability of failure $\mathcal{P}(T \in A_l \mid T > \tau_{l-1})$ for each patient at each time interval. Hence, the hazard $h_l = P(\tau_{l-1} < T \leq \tau_l | T > \tau_{l-1})$ is estimated first in each interval, and then, the survival probabilities are given as $S(t) = \prod_{l:t_l \leq t}(1 - h_l)$.

Random survival forests

RSFs are an ensemble tree method for survival analysis of right-censored data [50] adapted from random forests [51]. The main idea of random forests is to get a series of decision trees - which can capture complex interactions but are notorious for their high variance - and obtain a collection averaging their characteristics. In this way weak learners (the individual trees) are turned into strong learners (the ensemble) [55]. Randomness is introduced in two ways for RSFs: bootstrapping (sampling with replacement) a number of patients at each tree $\mathcal{B}$ times

and selecting a subset of variables for growing each node. During growing each survival tree, a recursive application of binary splitting is performed per region (called node) on a specific predictor in such a way that survival difference between daughter nodes is maximised and difference within them is minimised. Splitting is terminated when a certain criterion is reached (these nodes are called terminal).

The fundamental principle behind each survival tree is the conservation of events. This principle asserts that the sum of estimated cumulative hazard estimate over time is equal to the total number of deaths, therefore the total number of deaths is conserved within each terminal node $\mathcal{H}$. RSFs can handle both data with large sample size and vast number of predictors. Moreover, they can reach remarkable stability combining the results of many trees.

Machine learning versus statistical models for survival prediction: An open discussion

Nowadays, there is an open debate regarding the added value of ML versus SM within clinical and healthcare practice [56, 57]. For survival data, the most commonly applied SM for prediction is the Cox proportional hazards regression model [23]. This model allows a straightforward interpretation (via hazard ratios), but it is at the same time restricted to the proportional hazards assumption (the ratio of the hazards for any two individuals is constant over time), and requires a manual pre-specification of interaction terms. On the contrary, ML techniques are assumption-free and data adaptive which means that they can naturally incorporate interactions between the predictive features. However, ML models are prone to overfitting of the training data and they lack extensive assessment of predictive accuracy (i.e., absence of calibration) [58, 59]. Needless to say, overfitting might also occur with a traditional regression model if it is too complex (estimation of too many parameters) thus limiting generalisability (the ability of the model to make good predictions) outside training data.

The choice of the appropriate methodology should be motivated by the available real-life data and their complexity. SM usually perform well if the sample size is low/moderate, if there is a small number of variables (low-dimensional setting) with a low signal to noise ratio, or when predictors are associated with the outcome in a linear or additive way. On the other hand, ML techniques may be a better choice if the sample size is large/huge, if there is a large number of variables (high-dimensional setting) with a high signal to noise ratio, or when nonlinearity and nonadditivity are expected to be strong [60]. In this thesis, ML techniques are compared with SM for right-censored data in terms of prediction to address different real-life situations (small/medium or large clinical datasets, low- or high-dimensional settings).

Outline: research objectives of part II

There is a growing interest by the medical community in applying ML to predict clinical outcomes [20]. In a recent comprehensive survey, Wang et al. (2019) [47] discuss conventional and modern methods for survival analysis with right-censored data. The authors conclude that SNNs are well-suited to predicting survival and estimating disease risk, and are able to provide

personalised treatment recommendations. Even so, despite their non-negligible use in medicine, a comprehensive review on SNNs using prognostic factors is missing. In **Chapter** 5, we fill this gap with a structured overview of SNNs with prognostic factors for clinical prediction, which can be used as a guideline for future research. Our objective is to provide a broad understanding of the literature (1st January 1990 - 31st August 2021). We discuss how SNNs are employed in the medical field for prediction and detail how researchers have tried to adapt a classification method to right-censored survival data. We also provide a critical appraisal of model aspects to be designed and reported more carefully in future studies. We identify key characteristics of prediction models (i.e., number of patients/predictors, evaluation measures, calibration), and compare neural networks' predictive performance to the Cox proportional hazards model [23]. We conclude with recommendations about the correct application of SNNs in context of clinical prediction models, and discuss limitations and potential directions of future research.

There is an open discussion about the value of ML versus SM within clinical and healthcare practice. ML techniques might be an attractive choice for modelling complex data. In **Chapter** 6 [61], ML techniques (RSF [50], and two methodological extensions of PLANN [48]) developed in this thesis are tested to large retrospective data of 62294 patients from the United States provided by the Scientific Registry for Transplant Recipients. A total of 97 predictors are selected to predict survival since transplantation on clinical/statistical grounds. A comparison is performed between three different Cox models [23] and the ML techniques. Clinical endpoint is overall graft-survival defined as the time between transplantation and the date of graft-failure or death. Well-established predictive measures are employed from the survival field. The main aims of this project can be outlined as: (i) investigate the potential role of ML as competitor to traditional methods when complexity of the data is high (large sample size, high dimensional setting), (ii) identify the strongest potential risk factors for each method, (iii) evaluate the predictions and goodness of fit (calibration) of each method, and (iv) discuss the clinical relevance of the findings (potential for medical applications).

In the previous study, our group provided new methodological extensions of PLANN [61]. "PLANN extended" was developed and validated for complex liver transplantation data with a large sample size and within a high-dimensional setting. However, it is common to have a small number of patients recruited in clinical trials and a limited set of predictive features, for instance in sarcoma trials. Even so, there is an expectation by clinicians that ML models may perform better than SM. Hence, in **Chapter** 7 [62], the focus is on ML techniques versus SM for non-complex clinical data (small/medium sample size, low-dimensional) to investigate a different real-life setting. A Monte-Carlo simulation study is performed to compare PLANN original or extended [48, 61] with Cox models [23] for right-censored survival data in terms of prediction (discrimination and calibration). More specifically, real-life clinical data is mimicked to simulate synthetic data (5 predictors, 250 or 1000 observations) and to address different scenarios (20, 40, 61, or 80% censoring) which are representative of the real disease (bone sarcoma). The dataset originates from a randomised phase III European Osteosarcoma Intergroup study (MRC BO06/EORTC 80931) that investigated the effect of dose-intense chemotherapy in patients with localised extremity osteosarcoma [63]. The endpoint of interest is OS defined as the time to death from any cause since the date of surgery. As part of our objectives, we investi-

gate the robustness of PLANN original [48] and PLANN extended [61] in scenarios with less observations or less information available (due to the presence of censoring), and we explain the practical relevance of the findings (SM versus ML).

In health research, several chronic diseases are susceptible to CRs. Initially, SM were developed to estimate the cumulative incidence of an event of interest in the presence of CRs. As recently there is a growing interest in applying ML for clinical prediction, these techniques have also been extended to CRs but the literature is limited. In **Chapter** 8, two SM (cause-specific Cox [23], Fine-Gray [31]) and three ML techniques (PLANNCR original [52], a new model called "PLANNCR extended", and RSFCR [54]) are employed for CRs. Our goal is to develop and validate prognostic clinical prediction models. A dataset with 3826 retrospectively collected patients from the PERSARC Study Group with extremity soft-tissue sarcoma (eSTS) and nine predictors is used to systematically assess the predictive performance (discrimination and calibration) of all methods in a simple clinical setting. The endpoint of interest is the time in years between surgery and disease progression (event of interest) or death (competing event). This work examines 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), and compares the methods with regards to practical utility for prediction.

In **Chapter** 9, our findings are summarized and put into a broader perspective. We end the thesis with some suggestions for future research.

References

[1] M. Van Glabbeke, J. Verweij, I. Judson, and O. S. Nielsen. Progression-free rate as the principal end-point for phase II trials in soft-tissue sarcomas. *European Journal of Cancer*, 38(4):543–549, 2002. doi: 10.1016/S0959-8049(01)00398-7.

[2] M. E. Kallen and J. L. Hornick. The 2020 WHO classification: What's new in soft tissue tumor pathology? *American Journal of Surgical Pathology*, 45(1):1–23, 1 2021. ISSN 15320979. doi: 10.1097/PAS.0000000000001552. URL https://pubmed.ncbi.nlm.nih.gov/32796172/.

[3] A. C. Gamboa, A. Gronchi, and K. Cardona. Soft-tissue sarcoma in adults: An update on the current state of histiotype-specific management in an era of personalized medicine. *CA: A Cancer Journal for Clinicians*, 70(3):200–229, 2020. ISSN 0007-9235. doi: 10.3322/caac.21605.

[4] A. M. Frezza, S. Stacchiotti, and A. Gronchi. Systemic treatment in advanced soft tissue sarcoma: What is standard, what is new. *BMC Medicine*, 15(1):1–12, 2017. ISSN 17417015. doi: 10.1186/s12916-017-0872-y.

[5] N. T. Hoang, L. A. Acevedo, M. J. Mann, and B. Tolani. A review of soft-tissue sarcomas: Translation of biological advances into treatment measures. *Cancer Management and Research*, 10:1089–1114, 2018. ISSN 11791322. doi: 10.2147/CMAR.S159641.

[6] M. P. Vezeridis, R. Moore, and C. P. Karakousis. Metastatic Patterns in Soft-Tissue Sarcomas. *Archives of Surgery*, 118(8):915–918, 1983. ISSN 15383644. doi: 10.1001/archsurg.1983.01390080023007. URL https://pubmed.ncbi.nlm.nih.gov/6307217/.

[7] B. Vincenzi, A. M. Frezza, G. Schiavon, D. Santini, P. Dileo, M. Silletta, F. Bertoldo, G. Badalamenti, G. G. Baldi, S. Zovato, R. Berardi, M. Tucci, J. Whelan, R. Tirabosco, A. P. Dei Tos, and G. Tonini. Bone metastases in soft tissue sarcoma patients: A survey of natural, prognostic value, and treatment. *Clinical sarcoma research*, 3(1):1–5, 2013. ISSN 0732-183X. doi: 10.1200/jco.2012.30.15{_}suppl.10063.

[8] P. Reichardt. The Story of Imatinib in GIST - A Journey through the Development of a Targeted Therapy. *Oncology Research and Treatment*, 41(7-8):472–477, 7 2018. ISSN 22965262. doi: 10.1159/000487511. URL https://www.karger.com/Article/Abstract/487511.

[9] L. M. Nystrom, N. B. Reimer, J. D. Reith, L. Dang, R. A. Zlotecki, M. T. Scarborough, and C. P. Gibbs. Multidisciplinary management of soft tissue sarcoma. *The Scientific World Journal*, 2013, 2013. ISSN 1537744X. doi: 10.1155/2013/852462. URL /pmc/articles/PMC3745982/.

[10] A. Smrke, Y. Wang, and C. Simmons. Update on systemic therapy for advanced soft-tissue sarcoma. *Current Oncology*, 27(s1):25–33, 2020. ISSN 17187729. doi: 10.3747/CO.27.5475.

[11] O. Mir, T. Brodowicz, A. Italiano, J. Wallet, J. Y. Blay, F. Bertucci, C. Chevreau, S. Piperno-Neumann, E. Bompas, S. Salas, C. Perrin, C. Delcambre, B. Liegl-Atzwanger, M. Toulmonde, S. Dumont, I. Ray-Coquard, S. Clisant, S. Taieb, C. Guillemet, M. Rios, O. Collard, L. Bozec, D. Cupissol, E. Saada-Bouzid, C. Lemaignan, W. Eisterer, N. Isambert, L. Chaigneau, A. L. Cesne, and N. Penel. Safety and efficacy of regorafenib in patients with advanced soft tissue sarcoma (REGOSARC): a randomised, double-blind, placebo-controlled, phase 2 trial. *The Lancet Oncology*, 17(12):1732–1742, 12 2016. ISSN 14745488. doi: 10.1016/S1470-2045(16)30507-1.

[12] R. Dersimonian and N. Laird. Meta-Analysis in Clinical Trials. *Controlled clinical trials*, 7(3):177–188, 1986.

[13] R. Dersimonian and N. Laird. Meta-Analysis in Clinical Trials Revisited. *Contemporary clinical trials*, 45:139–145, 2015. doi: 10.1016/j.cct.2015.09.002.Meta-Analysis.

[14] M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein. *Introduction to Meta-Analysis*. John Wiley & Sons, 2011. ISBN 1119964377. URL https://books.google.be/books/about/Introduction_to_Meta_Analysis.html?id=JQg9jdrq26wC&source=kp_cover&redir_esc=y.

[15] H. Sahai and M. I. Ageel. *The analysis of variance : fixed, random, and mixed models*. Springer Science and Business Media, 2012. doi: https://doi.org/10.1007/978-1-4612-1344-4.

[16] G. Kantidakis, S. Litière, A. Neven, M. Vinches, I. Judson, P. Schöffski, E. Wardelmann, S. Stacchiotti, L. D'Ambrosio, S. Marréaud, W. T. A. van der Graaf, B. Kasper, M. Fiocco, and H. Gelderblom. Efficacy thresholds for clinical trials with advanced or metastatic leiomyosarcoma patients: A European Organisation for Research and Treatment of Cancer Soft Tissue and Bone Sarcoma Group meta-analysis based on a literature review for soft-tissue sarcoma. *European Journal of Cancer*, 154:253–268, 2021. ISSN 18790852. doi: 10.1016/j.ejca.2021.06.025.

[17] L. H. Lindner, S. Litière, S. Sleijfer, C. Benson, A. Italiano, B. Kasper, C. Messiou, H. Gelderblom, E. Wardelmann, A. Le Cesne, J. Y. Blay, S. Marreaud, N. Hindi, I. M. E. Desar, A. Gronchi, and W. T. A. van der Graaf. Prognostic factors for soft tissue sarcoma patients with lung metastases only who are receiving first-line chemotherapy: An exploratory, retrospective analysis of the European Organization for Research and Treatment of Cancer-Soft Tissue and Bone Sarcoma. *International Journal of Cancer*, 142(12): 2610–2620, 2018. ISSN 10970215. doi: 10.1002/ijc.31286.

[18] E. Younger, S. Litière, A. Le Cesne, O. Mir, H. Gelderblom, A. Italiano, S. Marreaud, R. L. Jones, A. Gronchi, and W. T. A. van der Graaf. Outcomes of Elderly Patients with Advanced Soft Tissue Sarcoma Treated with First-Line Chemotherapy: A Pooled Analysis of 12 EORTC Soft Tissue and Bone Sarcoma Group Trials. *The Oncologist*, 23(10): 1250–1259, 2018. ISSN 1083-7159. doi: 10.1634/theoncologist.2017-0598.

[19] G. Kantidakis, S. Litière, H. Gelderblom, M. Fiocco, I. Judson, W. T. A. van der Graaf, A. Italiano, S. Marréaud, S. Sleijfer, G. Mechtersheimer, C. Messiou, and B. Kasper. Prognostic Significance of Bone Metastasis in Soft Tissue Sarcoma Patients Receiving Palliative Systemic Therapy: An Explorative, Retrospective Pooled Analysis of the EORTC-Soft Tissue and Bone Sarcoma Group (STBSG) Database. *Sarcoma*, 2022:1–13, 4 2022. ISSN 1369-1643. doi: 10.1155/2022/5815875. URL https://www.hindawi.com/journals/sarcoma/2022/5815875/.

[20] 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.

[21] J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, 2nd edition, 2003. ISBN 038795399X,9780387953991,9780387216454. doi: 10.1007/b97377. URL https://www.springer.com/gp/book/9780387953991.

[22] J. O'Quigley. *Survival Analysis*. Springer, 1st edition, 2021. ISBN 978-3-03-033438-3, 978-3-03-033439-0. doi: 10.1007/978-3-030-33439-0.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] D. E. Leisman, M. O. Harhay, D. J. Lederer, M. Abramson, A. A. Adjei, J. Bakker, Z. K. Ballas, E. Barreiro, S. C. Bell, R. Bellomo, J. A. Bernstein, R. D. Branson, V. Brusasco, J. D. Chalmers, S. Chokroverty, G. Citerio, N. A. Collop, C. R. Cooke, J. D. Crapo, G. Donaldson, D. A. Fitzgerald, E. Grainger, L. Hale, F. J. Herth, P. M. Kochanek, G. Marks, J. R. Moorman, D. E. Ost, M. Schatz, A. Sheikh, A. R. Smyth, I. Stewart, P. W. Stewart, E. R. Swenson, R. Szymusiak, J. L. Teboul, J. L. Vincent, J. A. Wedzicha, and D. M. Maslove. Development and Reporting of Prediction Models: Guidance for Authors from Editors of Respiratory, Sleep, and Critical Care Journals. *Critical Care Medicine*, 48(5):623–633, 2020. ISSN 15300293. doi: 10.1097/CCM.0000000000004246.

[33] F. E. Harrell, K. L. Lee, and D. B. Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387, 1996. doi: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

[34] J. C. van Houwelingen and S. Le Cessie. Predictive value of statistical models. *Statistics in Medicine*, 9(11):1303–1325, 1990. doi: https://doi.org/10.1002/sim.4780091109.

[35] 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.

[36] J. C. van Houwelingen and H. Putter. *Dynamic prediction in clinical survival analysis*. CRC Press, 1st edition, 2012. ISBN 9781439835333. URL https://www.crcpress.com/Dynamic-Prediction-in-Clinical-Survival-Analysis/van-Houwelingen-Putter/p/book/9781439835333.

[37] 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, 1 2010. ISSN 10443983. doi: 10.1097/EDE.0b013e3181c30fb2. URL https://pubmed.ncbi.nlm.nih.gov/20010215/.

[38] 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.

[39] 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.

[40] 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.

[41] R. C. Deo. Machine learning in medicine. *Circulation*, 132(20):1920–1930, 2015. ISSN 15244539. doi: 10.1161/CIRCULATIONAHA.115.001593.

[42] A. Rajkomar, J. Dean, and I. Kohane. Machine Learning in Medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019. ISSN 0028-4793. doi: 10.1056/nejmra1814259.

[43] William B. Schwartz. Medicine and the computer. The promise and problems of change. *The New England journal of medicine*, 283(23):1257–1264, 12 1970. ISSN 0028-4793. doi: 10.1056/NEJM197012032832305. URL https://pubmed.ncbi.nlm.nih.gov/4920342/.

[44] P. G. Woodruff, B. Modrek, D. F. Choy, G. Jia, A. R. Abbas, A. Ellwanger, J. R. Arron, L. L. Koth, and J. V. Fahy. T-helper type 2-driven inflammation defines major subphenotypes of asthma. *American Journal of Respiratory and Critical Care Medicine*, 180(5): 388–395, 2009. ISSN 1073449X. doi: 10.1164/rccm.200903-0392OC.

[45] J. Corren, R. F. Lemanske, N. A. Hanania, P. E. Korenblat, M. V. Parsey, J. R. Arron, J. M. Harris, H. Scheerens, L. C. Wu, Z. Su, S. Mosesova, M. D. Eisner, S. P. Bohen, and J. G. Matthews. Lebrikizumab Treatment in Adults with Asthma. *New England Journal of Medicine*, 365(12):1088–1098, 9 2011. ISSN 0028-4793. doi: 10.1056/NEJMOA1106469/SUPPL{_}FILE/NEJMOA1106469{_}DISCLOSURES.PDF. URL https://www.nejm.org/doi/full/10.1056/nejmoa1106469.

[46] R. Cuocolo, M. Caruso, T. Perillo, L. Ugga, and M. Petretta. Machine Learning in oncology: A clinical appraisal. *Cancer Letters*, 481(February):55–62, 2020. ISSN 18727980. doi: 10.1016/j.canlet.2020.03.032.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] Ian A. Scott, David Cook, Enrico W. Coiera, and Brent Richards. Machine learning in clinical practice: prospects and pitfalls. *Medical Journal of Australia*, 211(5):203–205, 9 2019. ISSN 1326-5377. doi: 10.5694/MJA2.50294. URL https://onlinelibrary.wiley.com/doi/10.5694/mja2.50294.

[57] Rishi J. Desai, Shirley V. Wang, Muthiah Vaduganathan, Thomas Evers, and Sebastian Schneeweiss. Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes. *JAMA Network Open*, 3(1):e1918962–e1918962, 1 2020. ISSN 25743805. doi: 10.1001/JAMANETWORKOPEN.2019.18962. URL https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2758475.

[58] 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), 1 2015. ISSN 17417015. doi: 10.1186/s12916-014-0241-z. URL http://www.biomedcentral.com/1741-7015/13/1.

[59] G. S. Collins and K. G. M. Moons. Reporting of artificial intelligence prediction models. *The Lancet*, 393(10181):1577–1579, 4 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.

[60] F. E. Harrell Jr. Road Map for Choosing Between Statistical Modeling and Machine Learning | Statistical Thinking. URL https://www.fharrell.com/post/stat-ml/.

[61] 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, 12 2020. ISSN 14712288. doi: 10.1186/s12874-020-01153-1.

[62] 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.

[63] I. J. Lewis, M. A. Nooij, J. Whelan, M. R. Sydes, R. Grimer, P. C. W. Hogendoorn, M. A. Memon, S. Weeden, B. M. Uscinska, M. Ven Glabbeke, A. Kirkpatrick, E. I. Hauben, A. W. Craft, and A. H. M. Taminiau. Improvement in histologic response but not survival in osteosarcoma patients treated with intensified chemotherapy: A randomized phase III trial of the european osteosarcoma intergroup. *Journal of the National Cancer Institute*, 99(2):112–128, 2007. ISSN 14602105. doi: 10.1093/jnci/djk015.