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

Kantidakis, G.

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

Georgios Kantidakis

Analysis of sarcoma and non-sarcoma clinical data with statistical methods and machine learning techniques.
PhD thesis. Department of Medical Oncology, Leiden University Medical Center, Leiden, the Netherlands.

ISBN: 978 90 361 0562 0
Printing & cover design: HAVEKA BV | de grafische partner

Financial support for the publication of this thesis was kindly provided by the European Organisation for Research and Treatment of Cancer - Soft Tissue and Bone Sarcoma Group (EORTC - STBSG), the EORTC Cancer Research Fund (ECRF), and the Leiden University Medical Center (LUMC) Department of Medical Oncology.

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

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof. dr. ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op woensdag 23 november 2022
klokke 10.00 uur

door

Georgios Kantidakis

geboren te Athene, Griekenland
in 1993

Promotores:

prof. dr. A.J. Gelderblom	
prof. dr. M. Fiocco	

Copromotor:

dr. S. Litière	European Organisation for Research and Treatment of Cancer

Promotiecommissie:

prof. dr. M.A.J. van de Sande	
prof. dr. C. Legrand	Université catholique de Louvain
prof. dr. M.R. Spruit	
dr. J.H.M. Merks	Prinses Máxima Center, Universiteit van Amsterdam
dr. F. Ieva	Politecnico di Milano

To my father:

Ioannis Kantidakis

and to my grandmother:

Angeliki Varvaropoulou

Contents

Chapter 1 General introduction 1

I Clinical trials in soft-tissue sarcomas 23

Chapter 2 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 sarcomas 25

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, H. Gelderblom.

European Journal of Cancer, 154:253–268, 2021

Chapter 3 New benchmarks to design clinical trials with advanced or metastatic liposarcoma or synovial sarcoma patients: A EORTC - Soft Tissue and Bone Sarcoma Group (STBSG) meta-analysis based on a literature review for soft-tissue sarcomas 43

G. Kantidakis, S. Litière, A. Neven, M. Vinches, I. Judson, J. Y. Blay, E. Wardelmann, S. Stacchiotti, L. D'Ambrosio, S. Marréaud, W. T. A. van der Graaf, B. Kasper, M. Fiocco, H. Gelderblom.

European Journal of Cancer, 174:261-276, 2022

Chapter 4 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 63

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, B. Kasper.

Sarcoma, 2022:5815875, 2022

II Statistical models versus machine learning to predict survival for sarcoma and non-sarcoma clinical data 81

Chapter 5 Neural networks for survival prediction in medicine using prognostic factors: a review and critical appraisal 83

G. Kantidakis, A. D. Hazewinkel, M. Fiocco.
Computational and Mathematical Methods in Medicine, 2022:1176060, 2022

Chapter 6 Survival prediction models since liver transplantation - comparisons between Cox models and machine learning techniques **109**
G. Kantidakis, H. Putter, C. Lancia, J. Boer, A. E. Braat, M. Fiocco.
BMC Medical Research Methodology, 20(1):1-14, 2020

Chapter 7 A simulation study to compare the predictive performance of survival neural networks with Cox models for clinical trial data **129**
G. Kantidakis, E. Biganzoli, H. Putter, M. Fiocco.
Computational and Mathematical Methods in Medicine, 2021:2160322, 2021

Chapter 8 Statistical models versus machine learning for competing risks: development and validation of prognostic models **149**
G. Kantidakis, H. Putter, S. Litière, M. Fiocco.
Submitted

Chapter 9 General discussion **169**

Summary **191**

Nederlandse samenvatting **195**

Acknowledgements **199**

Curriculum vitae **201**

List of publications **203**

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 ≥ 40% for drug activity and ≤ 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 ≥ 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

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

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

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

where $\frac{1}{\hat{C}(\text{min}(t_i-,t_0)|x_i)}$ 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 identification 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.

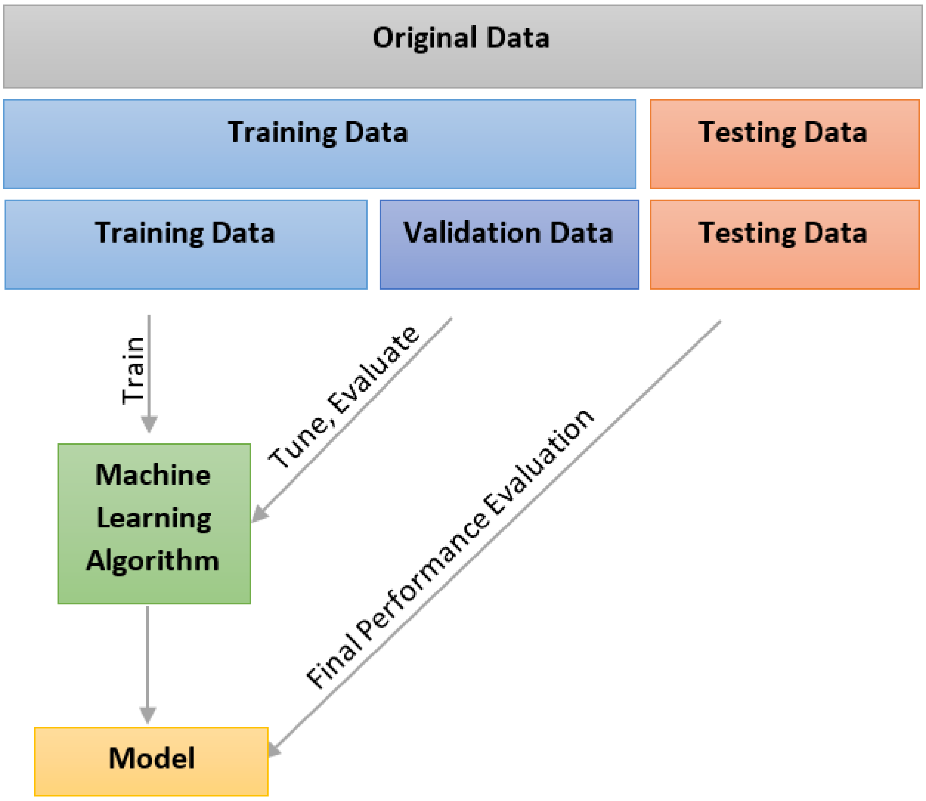


Figure 1.1: General strategy for model development and evaluation of a ML prediction model.

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 investigate 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.

Part I

Clinical trials in soft-tissue sarcomas

2

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 sarcomas

This chapter is based on: 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.

Abstract

Background: In 2002, the European Organisation for Research and Treatment of Cancer Soft Tissue and Bone Sarcoma Group reported well-established values for conducting phase II trials for soft-tissue sarcomas. An update is provided for leiomyosarcoma (LMS).

Materials and methods: Clinical trials with advanced or metastatic LMS were identified via literature review in PubMed (published 2003–2018, ≥10 adult LMS patients). End-points were 3- and 6-month progression-free survival rates (PFSR-3m and PFSR-6m). When estimates could not be derived from publications, data requests were sent out. Treatments were classified as recommended (R-T) or non-recommended (NR-T) according to the ESMO 2018 guidelines. A random effects meta-analysis was used to pool trial-specific estimates for first-line (1L) or pre-treated (2L+) patients separately. The ESMO Magnitude of Clinical Benefit Scale was used to guide the treatment effect to target in future trials.

Results: From 47 studies identified, we obtained information on 7 1L and 16 2L+ trials for 1500 LMS patients. Overall, in 1L, PFSR-3m and PFSR-6m were 74% (95% confidence interval [CI] 64–82%) and 58% (95% CI 50–66%), respectively. For 2L+, PFSR-3m was 48% (95% CI 41–54%), and PFSR-6m was 28% (95% CI 22–34%). No difference was observed between R-T and NR-T for first or later lines. Under the alternative that the true benefit amounts to a hazard ratio of 0.65, a PFSR-6m ≥70% can be considered to suggest drug activity in 1L. For 2L+, a PFSR-3m ≥62% or PFSR-6m ≥44% would suggest drug activity. Specific results are also provided for uterine LMS.

Conclusions: This work provides a new benchmark for designing phase II studies for advanced or metastatic

LMS.

2.1 Introduction

Non-gastrointestinal stromal tumour soft-tissue sarcomas (STS) constitute a very heterogeneous group of mesenchymal rare malignancies, accounting for 1% of all adult malignancies, with widely varying genetics, prognostic factors, and sensitivity to treatments [1]. The tumours metastasise predominantly to the lungs [1, 2]. Gastrointestinal stromal tumour (GIST) is generally considered separately because it is responsive to receptor tyrosine kinase inhibitors, most notably imatinib. The prognosis of patients with advanced or metastatic STS is poor, with a median overall survival (OS) of 12–17 months after first-line treatment and an estimated 2-year OS of 20–30% after treatment with standard cytotoxic chemotherapy drugs [3, 4]. In these patients, treatment is often palliative to delay progression and severe morbidity. Doxorubicin and ifosfamide are considered the most active drugs used either singly or in combination for first line with a response rate (RR) of 10–25% [5]. Dacarbazine and the combination of docetaxel and gemcitabine are also treatments with some recognised activity [6, 7]. Frequently used drugs, particularly for the second and further lines of treatment of LMS, are trabectedin, dacarbazine, pazopanib, and gemcitabine [8].

In total, more than 100 histologic subtypes have been recognised occurring in the trunk, extremity, and retroperitoneum [1]. The commonest histotypes are leiomyosarcoma (LMS; ∼20%), liposarcoma (∼20%), undifferentiated pleiomorphic sarcoma (∼15%), and synovial sarcoma (∼6%), with the remaining histotypes being individually rarer [9].

LMS — one of the most common STS — has a wide anatomical distribution exhibiting complex genetic alterations. LMS occurs most frequently in the uterus and is the most prevalent form of gynaecologic sarcoma. It comprises ∼20% of STS being rare but aggressive [10, 11]. First-line patients with locally advanced or metastatic LMS have poor prognosis (median OS ∼17 months) and are usually treated with doxorubicin alone, or in combination with ifosfamide, or dacarbazine [7, 12]. Non-uterine and uterine LMS (uLMS) should be considered separately since different gene patterns are expressed and different clinical behaviour has been reported that might make uLMS more chemosensitive [13, 14]. Systemic treatment for advanced uterine LMS with doxorubicin or gemcitabine-based regimens results in median progression-free survival (PFS) of 6–8 months and median OS of <2 years [15].

As historical benchmarking, Van Glabbeke et al. published in 2002 a pooled analysis on behalf of the European Organisation for Research and Treatment of Cancer Soft Tissue and Bone Sarcoma Group (EORTC - STBSG) estimating progression-free rate for various groups of STS patients who participated in EORTC phase II trials [16]. In this work, thresholds for activity were provided separately for first-line and pre-treated patients dividing drugs into active and inactive: a rate at 6 months of 30–56% was suggested as a reference for first line (depending on histology), and for second line, a 3-month rate was ≥40% for drug activity and ≤20% for inactivity (for any STS subgroup).

The aforementioned thresholds have been widely used (more than 400 citations) to design new studies for all STS or for specific histology subgroups. As they were calculated almost two decades ago, it is of great importance to provide updates to reflect current treatment practices. Moreover, in the previous decade, STS studies were designed based on the one-size-fits-all principle mixing several histologic subtypes. However, more recently, there is a clear trend towards histology-specific tailored research [1, 13]. To elaborate on this, the 2002 thresholds should not only be updated but also be evaluated separately for the most prevalent STS subtypes to aid the design of histology-specific trials. This is more relevant with the increased survival trend from the standard of care (i.e. doxorubicin) and multiple other agents such as eribulin, pazopanib, and trabectedin; all associated with improvements in supportive and multidisciplinary care [17, 18].

An extensive literature search was performed to identify all phase II or subsequent clinical trials of advanced or metastatic STS from 2003 to 2018, thus documenting the current landscape. Because of the heterogeneity among clinical trials (e.g. different treatments, subtypes, and phases), it was decided to focus first on LMS – the most commonly occurring STS subtype. Moreover, given the fact that PFS rates (PFSRs; counting death as an event) are nowadays a preferred and more frequently reported end-point than progression-free rates (censoring non-disease–related death), the primary end-point of interest in this work is PFSR at 3 and 6 months. The aim is to provide a new benchmark for designing phase II studies for advanced or metastatic LMS patients using PFS rates as the primary end-point.

2.2 Methods

2.2.1 Search strategy and selection criteria

This literature review and meta-analysis was performed according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines [19]. The details are provided in the Appendix pp 3–5. In summary, MEDLINE was searched through PubMed for phases II, III, or IV clinical trials for advanced or metastatic STS published from 1 January 2003 to 31 December 2018. Three investigators (Georgios Kantidakis, Anouk Neven, and Marie Vinches) independently examined the database. Two search algorithms were combined using the terms ‘sarcoma’, ‘clinical trial’, ‘advanced’, ‘metastatic’, and ‘human’.

Only articles published in English were included. Eligible study designs included randomised controlled or non-randomised clinical trials as well as prospective real-life studies. The study domain included any systemic therapy in non-resectable advanced or metastatic STS for first or later lines of treatment. Case–control studies, case series, review papers, early phase trials (phase I, I-II), reports, pooled analyses, and substudies were excluded. Articles with paediatric population or with retrospective clinical data were considered ineligible, as well as those dedicated exclusively to GIST or bone sarcomas. A two-step procedure was performed by the three investigators. The first step included screening of titles and abstracts, the second step of full text. During the first step, the name of study, first author and year of publication were extracted. At the second step, study design, study phase, number of patients registered, line of treatment, subtypes included/excluded, primary end-points, drugs used in the trial, and more summary estimates filling in total 41 variables in our database. In case of discordance, discussion followed to find a compromise. It was decided to first focus on LMS, the most frequent STS subtype in the screened trials.

2.2.2 Data extraction

To perform the meta-analysis, a line per treatment arm database was designed. For each line, Georgios Kantidakis extracted the year of study activation, LMS subgroup (all or uterine only), number of evaluable LMS patients (those who meet the statistical plan criteria for inclusion in efficacy data sets) for PFSR at 3/6 months with 95% confidence intervals (95% CIs). Placebo arms, treatment arms with less than 10 LMS patients, studies activated before 2000, or those with mixed treatment lines were excluded. When information on the end-points could not be extracted from a publication, first authors and/or study sponsors were contacted.

2.2.3 Statistical methods

The main analysis focused on the activity of drugs or drug combination, distinguishing between recommended (R-T) / non-recommended treatment (NR-T) regimens for LMS patients, measured in terms of the overall PFSR at 3/6 months. The ESMO 2018 guidelines were used as a criterion to perform drug classification [7]. A random effects model was used for each drug (or drug combination) to estimate an overall PFSR. A necessary component for the calculation of study heterogeneity was the variance of PFS (not available in publications). Therefore, for each treatment arm, the number of cases (patients alive and progression-free) at 3 and 6 months was approximated

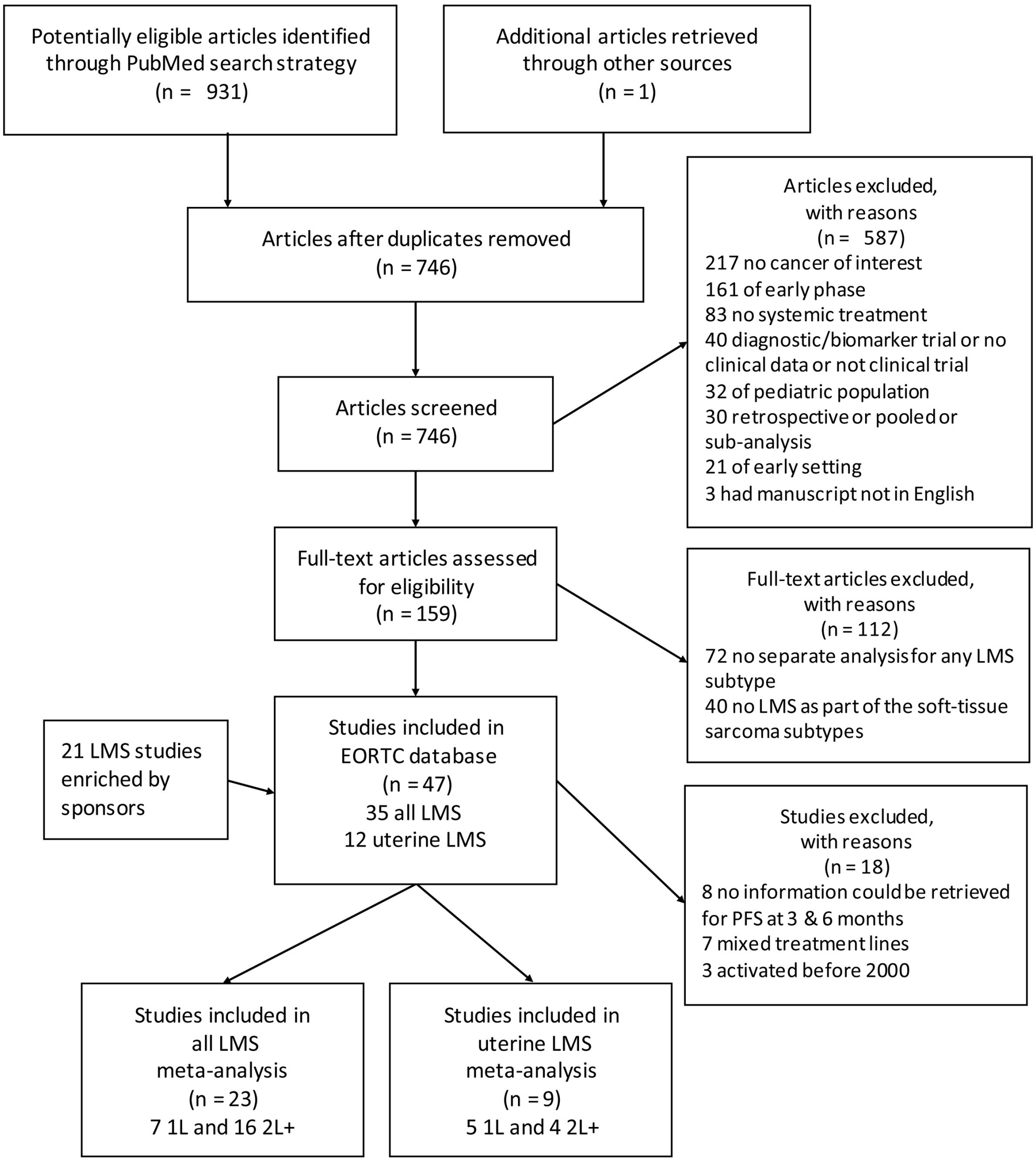

Potentially eligible articles identified through PubMed search strategy (n = 931)
Additional articles retrieved through other sources (n = 1)
Articles after duplicates removed (n = 746)
Articles excluded, with reasons (n = 587)
217 no cancer of interest
161 of early phase
83 no systemic treatment
40 diagnostic/biomarker trial or no clinical data or not clinical trial
32 of pediatric population
30 retrospective or pooled or sub-analysis
21 of early setting
3 had manuscript not in English
Articles screened (n = 746)
Full-text articles assessed for eligibility (n = 159)
Full-text articles excluded, with reasons (n = 112)
72 no separate analysis for any LMS subtype
40 no LMS as part of the soft-tissue sarcoma subtypes
21 LMS studies enriched by sponsors
Studies included in EORTC database (n = 47)
35 all LMS
12 uterine LMS
Studies excluded, with reasons (n = 18)
8 no information could be retrieved for PFS at 3 & 6 months
7 mixed treatment lines
3 activated before 2000
Studies included in all LMS meta-analysis (n = 23)
7 1L and 16 2L+
Studies included in uterine LMS meta-analysis (n = 9)
5 1L and 4 2L+

Figure 2.1: **Study selection**. For the uterine LMS meta-analysis, nine studies were included: six studies designed for uterine LMS and three designed for (all) LMS for which estimates for the uterine LMS subgroup were provided.

according to the number of evaluable LMS patients and a given PFS proportion (defined as cases/evaluable patients). Followingly, the estimated number of cases was used under a binomial distribution to calculate the variance and the 95% CIs for each drug/combination (see more details in Appendix pp 11–12) [20].

The inverse variance method, giving more weight to larger trials, was used to pool treatment-specific PFS estimates. These are reported on forest plots alongside the 95% CIs. To estimate the between-study variance, the DerSimonian-Laird's method was employed [21, 22]. An overall test on heterogeneity between studies was performed for each meta-analysis (value I^2 in figures) [23]. The association of drug groups (R-T/NR-T) with PFS was tested with a Z-statistic. The risk of publication bias was assessed with funnel plots and formal regression tests [24–26]. The Baujat plot was applied to detect sources of heterogeneity and potentially influential studies [27]. Meta-regressions were performed to test the effect of phase, study design, year of activation, and sample size on efficacy for all LMS, but not for uLMS (because of the small number of specific studies). First, the predictors were tested separately in univariate models and then any prognostic factors were added in multivariate models, including the drug groups, to investigate whether some part of the residual heterogeneity can be explained.

The ESMO Magnitude of Clinical Benefit Scale (MCBS) was used to guide the choice of treatment effect to target in future trials [28]. All reported P values are two sided. Analyses were performed using the packages metafor and meta in R (version 4.0.2) [29, 30].

2.3 Results

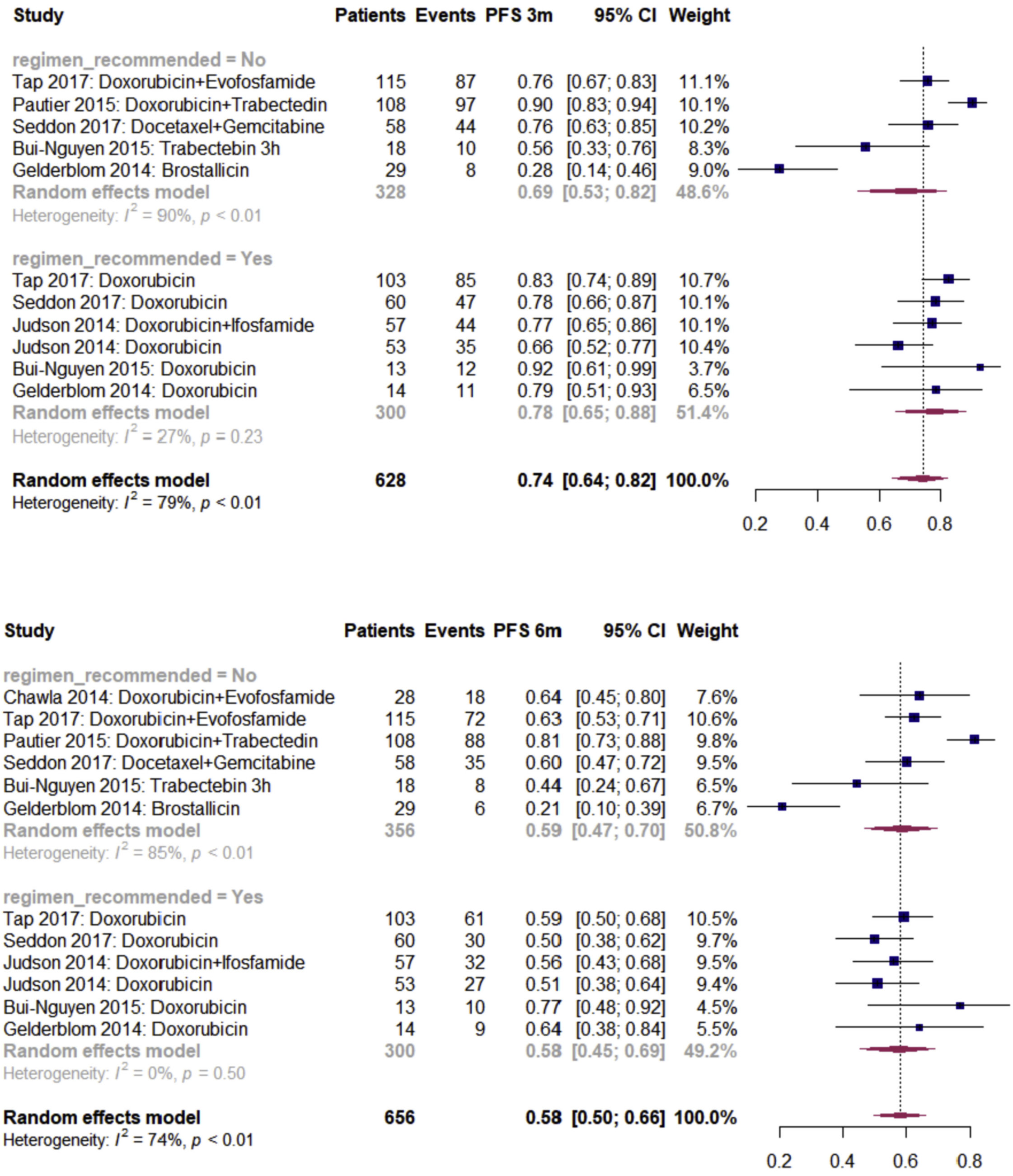


Figure 2.2: Forest plots of PFS at 3 (upper panel) and 6 (low panel) months for first line (all) LMS patients. PFS proportion at 3 or 6 months was defined as the (approximate) proportion of patients alive and without progression at 3 or 6 months after the start of treatment. Treatments were classified as recommended or non-recommended according to ESMO 2018 guidelines [7]. Heterogeneity refers to the variability between the study-specific effect sizes that cannot be explained by a random variation.

2.3.1 Included clinical trials

The search strategy identified 745 publications; 159 potentially relevant articles for STS were selected after abstract and full-text screening. A noticeable amount of variation was observed (e.g. different treatments, subtypes, and end-points). For this work, the focus is on LMS, which appeared more than 100 times (as LMS, uLMS, soft-tissue LMS etc). Forty-seven studies were identified for the meta-analyses. Overall, twenty-three trials were included in the all LMS meta-analysis (excluding trials designed only for uLMS patients) [3, 5, 9, 18, 31–49], and nine trials were included in the uLMS-specific meta-analysis [37, 45, 49–55] (see study selection in Figure 2.1).

First author (year of publication)	Study period	Study type	Phase	Treatment line	Total patients registered	Drug or drug combination	Recommended	Evaluable LMS patients for PFS (%)	Analysed group
Long et al. (2005)	2002–2003	Non-randomised trial	2	1	18	D+M+D+C+S	No	18 (100.00%)	Uterine LMS
Hartmann et al. (2007)	2002–2006	Non-randomised trial	2	2+	36	Bendamustine	No	15 (41.67%)	All LMS
Reichardt et al. (2007)	2002–2004	Non-randomised trial	2	2+	39	Exatecan	No	16 (41.03%)	All LMS
Hensley et al. (2008)	2003–2006	Non-randomised trial	2	1	42	Docetaxel + Gemcitabine	No	42 (100.00%)	Uterine LMS
Hensley et al. (2008)	2003–2006	Non-randomised trial	2	2+	51	Docetaxel + Gemcitabine	Yes	48 (94.12%)	Uterine LMS
Hensley et al. (2009)	2006–2007	Non-randomised trial	2	2+	25	Sunitinib	No	23 (92.00%)	Uterine LMS
Sleijfer et al. (2009)	2005–2007	Non-randomised trial	2	2+	142	Pazopanib	Yes	41 (28.87%)	All LMS
Schöffski et al. (2011)	2007–2009	Non-randomised trial	2	2+	128	Eribulin	No	38 (29.69%)	All LMS
Chawla et al. (2011)	2004–2005	Non-randomised trial	2	2+	216	Ridaforolimus	No	57 (26.39%)	All LMS
van der Graaf et al. (2012)	2008–2010	Randomised trial	3	2+	372	Pazopanib	Yes	92 (24.73%)	All LMS
Pautier et al. (2012)	2006–2008	Randomised trial	2	2+	90	Docetaxel + gemcitabine	Yes	21 (23.33%)	Uterine LMS
						Docetaxel + gemcitabine	Yes	40 (44.44%)	All LMS
						Gemcitabine	Yes	21 (23.33%)	Uterine LMS
						Gemcitabine	Yes	43 (47.78%)	All LMS
Schuetze et al. (2012)	2008–2009	Non-randomised trial	2	2+	49	Cyclophosphamide + sirolimus	No	16 (32.66%)	All LMS
Cassier et al. (2013)	2010	Non-randomised trial	2	2+	47	Panobinostat	No	10 (21.28%)	All LMS
Santoro et al. (2013)	2006–2010	Non-randomised trial	2	2+	100	Sorafenib	No	30 (30.00%)	All LMS
Schöffski et al. (2013)	2008–2012	Non-randomised trial	2	2+	113	Cixutumumab	No	22 (19.47%)	All LMS
Chawla et al. (2014)	2009–2011	Non-randomised trial	2	1	91	Doxorubicin + evofosfamide	No	28 (30.77%)	All LMS
Duska et al. (2014)	2010–2014	Non-randomised trial	2	2+	26	Ixabepilone	No	23 (88.46%)	Uterine LMS
Gelderblom et al. (2014)	2006–2008	Randomised trial	2	1	118	Brostallicin	No	29 (24.58%)	All LMS
						Doxorubicin	Yes	14 (11.86%)	All LMS
Judson et al. (2014)	2003–2010	Randomised trial	3	1	455	Doxorubicin + ifosfamide	Yes	57 (12.53%)	All LMS
						Doxorubicin	Yes	53 (11.65%)	All LMS
Bui-Nguyen et al. (2015)	2011–2012	Randomised trial	2|3	1	133	Trabectedin 3h	No	18 (13.53%)	All LMS
						Doxorubicin	Yes	13 (9.77%)	All LMS
Eroglu et al. (2015)	2010–2013	Randomised trial	2	2+	71	Selumetinib	No	10 (14.08%)	All LMS
						Selumetinib + temsirolimus	No	11 (15.49%)	All LMS
Hensley et al. (2015)	2009–2013	Randomised trial	3	1	107	Bevacizumab + docetaxel + gemcitabine	No	53 (49.53%)	Uterine LMS
						Docetaxel + gemcitabine	No	54 (50.47%)	Uterine LMS
Pautier et al. (2015)	2010–2013	Non-randomised trial	2	1	109	Doxorubicin + trabectedin	No	47 (43.12%)	Uterine LMS
						Doxorubicin + trabectedin	No	108 (99.08%)	All LMS
Mir et al. (2016)	2013–2014	Randomised trial	2	2+	182	Regorafenib	No	28 (15.38%)	All LMS
Schöffski et al. (2016)	2011–2013	Randomised trial	3	2+	452	Eribulin	No	157 (34.73%)	All LMS
						Dacarbazine	Yes	152 (33.63%)	All LMS
Schuetze et al. (2016)	2007–2009	Non-randomised trial	2	2+	200	Dasatinib	No	47 (23.50%)	All LMS
Kawai et al. (2017)	2011–2012	Non-randomised trial	2	2+	52	Eribulin	No	19 (36.54%)	All LMS
Tap et al. (2017)	2011–2014	Randomised trial	3	1	640	Doxorubicin + evofosfamide	No	115 (17.97%)	All LMS
						Doxorubicin	Yes	103 (16.09%)	All LMS
Seddon et al. (2017)	2010–2014	Randomised trial	3	1	257	Docetaxel + gemcitabine	No	35 (13.62%)	Uterine LMS
						Docetaxel + gemcitabine	No	58 (22.57%)	All LMS
						Doxorubicin	Yes	36 (14.01%)	Uterine LMS
						Doxorubicin	Yes	60 (23.35%)	All LMS

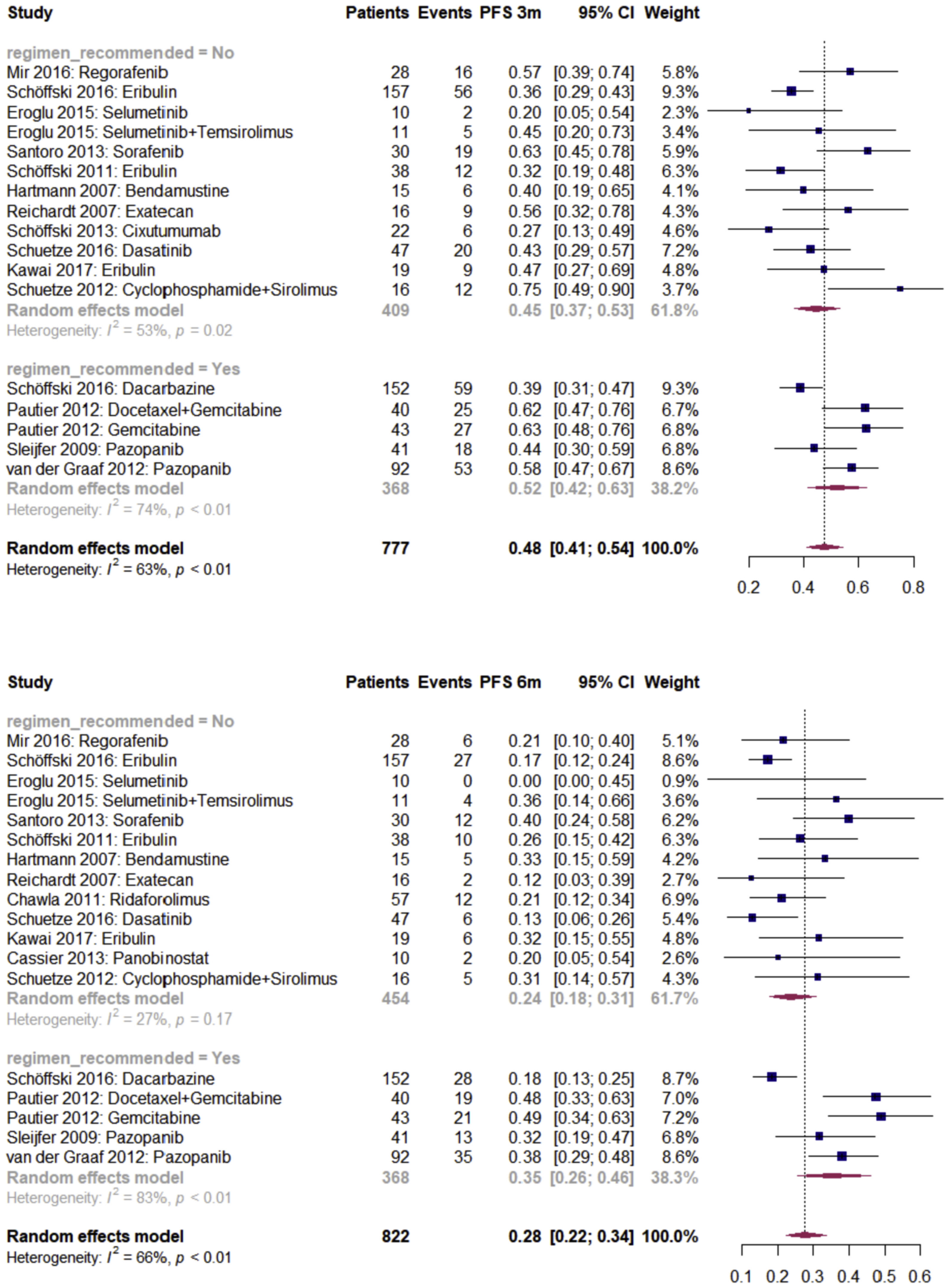
Table 2.1: **Main characteristics and results of studies included in the LMS meta-analyses**. Treatments were classified as recommended (yes or no) according to ESMO 2018 guidelines [7]. Study period = period of first to last patient accrual. NA = not available. Evaluable are those patients who meet the study's statistical plan criteria for inclusion in efficacy data sets. D + M + D + C + S = dacarbazine, mitomycin, doxorubicin, and cisplatin with sargramostim. Trabectedin 3h = trabectedin 3-h infusion treatment schedule. The 24-h infusion treatment arm was excluded from the meta-analysis because of the limited number of LMS patients (n = 6). You can find the full online version of this table here.

2.3.2 Characteristics of the included trials

A total of 1500 patients were evaluable for the LMS analysis (range 10–157; Table 1) and 421 for the uLMS analysis (range 18–54; Table 2.1). The most common drug regimen in first line for LMS was doxorubicin, either monotherapy (five times) or in combination with evofosfamide, ifosfamide, or trabectedin. Eribulin was the most common drug in pre-treated population (three times). For uLMS patients, the most frequent therapeutic option for any line was docetaxel + gemcitabine (five times).

2.3.3 Risk of bias

Contour-enhanced funnel plots did not portray any systematic asymmetry between studies for all LMS. Formal tests for publication bias for all LMS patients were non-significant ($P > 0.05$), indicating low risk of bias. On the contrary, a number of formal tests were significant for uLMS subanalysis ($P < 0.05$), indicating high risk of



Study	Patients	Events	PFS 3m	95% CI	Weight
regimen_recommended = No					
Mir 2016: Regorafenib	28	16	0.57	[0.39; 0.74]	5.8%
Schöffski 2016: Eribulin	157	56	0.36	[0.29; 0.43]	9.3%
Eroglu 2015: Selumetinib	10	2	0.20	[0.05; 0.54]	2.3%
Eroglu 2015: Selumetinib+Temsirolimus	11	5	0.45	[0.20; 0.73]	3.4%
Santoro 2013: Sorafenib	30	19	0.63	[0.45; 0.78]	5.9%
Schöffski 2011: Eribulin	38	12	0.32	[0.19; 0.48]	6.3%
Hartmann 2007: Bendamustine	15	6	0.40	[0.19; 0.65]	4.1%
Reichardt 2007: Exatecan	16	9	0.56	[0.32; 0.78]	4.3%
Schöffski 2013: Cixutumumab	22	6	0.27	[0.13; 0.49]	4.6%
Schuetze 2016: Dasatinib	47	20	0.43	[0.29; 0.57]	7.2%
Kawai 2017: Eribulin	19	9	0.47	[0.27; 0.69]	4.8%
Schuetze 2012: Cyclophosphamide+Sirolimus	16	12	0.75	[0.49; 0.90]	3.7%
Random effects model	409		0.45	[0.37; 0.53]	61.8%
Heterogeneity: $I^2 = 53\%$, $p = 0.02$					
regimen_recommended = Yes					
Schöffski 2016: Dacarbazine	152	59	0.39	[0.31; 0.47]	9.3%
Pautier 2012: Docetaxel+Gemcitabine	40	25	0.62	[0.47; 0.76]	6.7%
Pautier 2012: Gemcitabine	43	27	0.63	[0.48; 0.76]	6.8%
Sleijfer 2009: Pazopanib	41	18	0.44	[0.30; 0.59]	6.8%
van der Graaf 2012: Pazopanib	92	53	0.58	[0.47; 0.67]	8.6%
Random effects model	368		0.52	[0.42; 0.63]	38.2%
Heterogeneity: $I^2 = 74\%$, $p < 0.01$					
Random effects model	**777**		**0.48**	**[0.41; 0.54]**	**100.0%**
Heterogeneity: $I^2 = 63\%$, $p < 0.01$					

Study	Patients	Events	PFS 6m	95% CI	Weight
regimen_recommended = No					
Mir 2016: Regorafenib	28	6	0.21	[0.10; 0.40]	5.1%
Schöffski 2016: Eribulin	157	27	0.17	[0.12; 0.24]	8.6%
Eroglu 2015: Selumetinib	10	0	0.00	[0.00; 0.45]	0.9%
Eroglu 2015: Selumetinib+Temsirolimus	11	4	0.36	[0.14; 0.66]	3.6%
Santoro 2013: Sorafenib	30	12	0.40	[0.24; 0.58]	6.2%
Schöffski 2011: Eribulin	38	10	0.26	[0.15; 0.42]	6.3%
Hartmann 2007: Bendamustine	15	5	0.33	[0.15; 0.59]	4.2%
Reichardt 2007: Exatecan	16	2	0.12	[0.03; 0.39]	2.7%
Chawla 2011: Ridaforolimus	57	12	0.21	[0.12; 0.34]	6.9%
Schuetze 2016: Dasatinib	47	6	0.13	[0.06; 0.26]	5.4%
Kawai 2017: Eribulin	19	6	0.32	[0.15; 0.55]	4.8%
Cassier 2013: Panobinostat	10	2	0.20	[0.05; 0.54]	2.6%
Schuetze 2012: Cyclophosphamide+Sirolimus	16	5	0.31	[0.14; 0.57]	4.3%
Random effects model	454		0.24	[0.18; 0.31]	61.7%
Heterogeneity: $I^2 = 27\%$, $p = 0.17$					
regimen_recommended = Yes					
Schöffski 2016: Dacarbazine	152	28	0.18	[0.13; 0.25]	8.7%
Pautier 2012: Docetaxel+Gemcitabine	40	19	0.48	[0.33; 0.63]	7.0%
Pautier 2012: Gemcitabine	43	21	0.49	[0.34; 0.63]	7.2%
Sleijfer 2009: Pazopanib	41	13	0.32	[0.19; 0.47]	6.8%
van der Graaf 2012: Pazopanib	92	35	0.38	[0.29; 0.48]	8.6%
Random effects model	368		0.35	[0.26; 0.46]	38.3%
Heterogeneity: $I^2 = 83\%$, $p < 0.01$					
Random effects model	**822**		**0.28**	**[0.22; 0.34]**	**100.0%**
Heterogeneity: $I^2 = 66\%$, $p < 0.01$					

Figure 2.3: Forest plots of PFS at 3 (upper panel) and 6 (low panel) months for pre-treated (all) LMS patients.

publication bias there (see Appendix section 2.4 for further details).

2.3.4 All LMS meta-analyses

Starting with the all LMS meta-analyses, the pooled PFSR-3m for the first-line setting (Figure 2.2) were 78% (95% CI 65–88%) and 69% (95% CI 53–82%) for drugs classified as recommended/non-recommended, respectively. At 6 months, PFSR were 58% (95% CI 45–69%) and 59% (95% CI 47–70%), respectively. Differences between R-T and NR-T were not significant at 3 or 6 months (P value 0.32 and 0.90). Variability between the effect sizes that could not be explained was very high as indicated by overall heterogeneity ($I^2 > 70\%$, P < 0.01). Univariate meta-regressions showed that sample size >38 (median value) is a prognostic factor for PFS at 3 months. Nevertheless, multivariate meta-regression adding this variable did not explain much of the residual 3-month heterogeneity ($I^2 =$ 73%, P < 0.01). No significant factor was identified for PFSR-6m (see Appendix). For the pre-treated population (Figure 2.3), the pooled PFSR-3m were 52% (95% CI 42–63%) for R-T and 45% (95% CI 37–53%) for NR-T. PFSR-6m for R-T and NR-T were 35% (95% CI 26–46%) and 24% (95% CI 18–31%), respectively. Similarly, differences were not significant between the R-T/NR-T (P values 0.27 and 0.06). Remaining variability was high ($I^2 > 60\%$, P < 0.01). None of the tested variables was prognostic at 3 months. Year of activation was a prognostic factor for PFSR-6m. Multivariate adjustment with it explained a part of the residual heterogeneity at 6 months (I^2 = 39%, P = 0.06).

2.3.5 Uterine LMS meta-analyses

For first-line treatment of uLMS patients (Figure 2.4), the pooled PFSR-3m were 75% (95% CI 51–90%) and 70% (95% CI 60–78%) for R-T and NR-T, respectively. The PFSR-6m for R-T and NR-T were 39% (95% CI 18–65%) and 51% (95% CI 40–62%), respectively. Differences were not significant at 3 and 6 months (P values 0.66 and 0.41). Overall heterogeneity was moderate to high at 3 months (I^2 = 48%; P = 0.07) and high at 6 months (I^2 = 62%; P = 0.01). For pre-treated patients (Figure 2.5), the PFSR-3m for R-T and NR-T were 68% (95% CI 52–81%) and 23% (95% CI 10–44%), respectively. The PFSR-6m for R-T and NR-T were 50% (95% CI 40–60%) and 13% (95% CI 5–28%), respectively. Notably, there was a statistically significant difference between the classified drugs (P values < 0.01 at both 3 and 6 months). Overall variation between studies was high ($I^2 > 70\%$, P < 0.01).

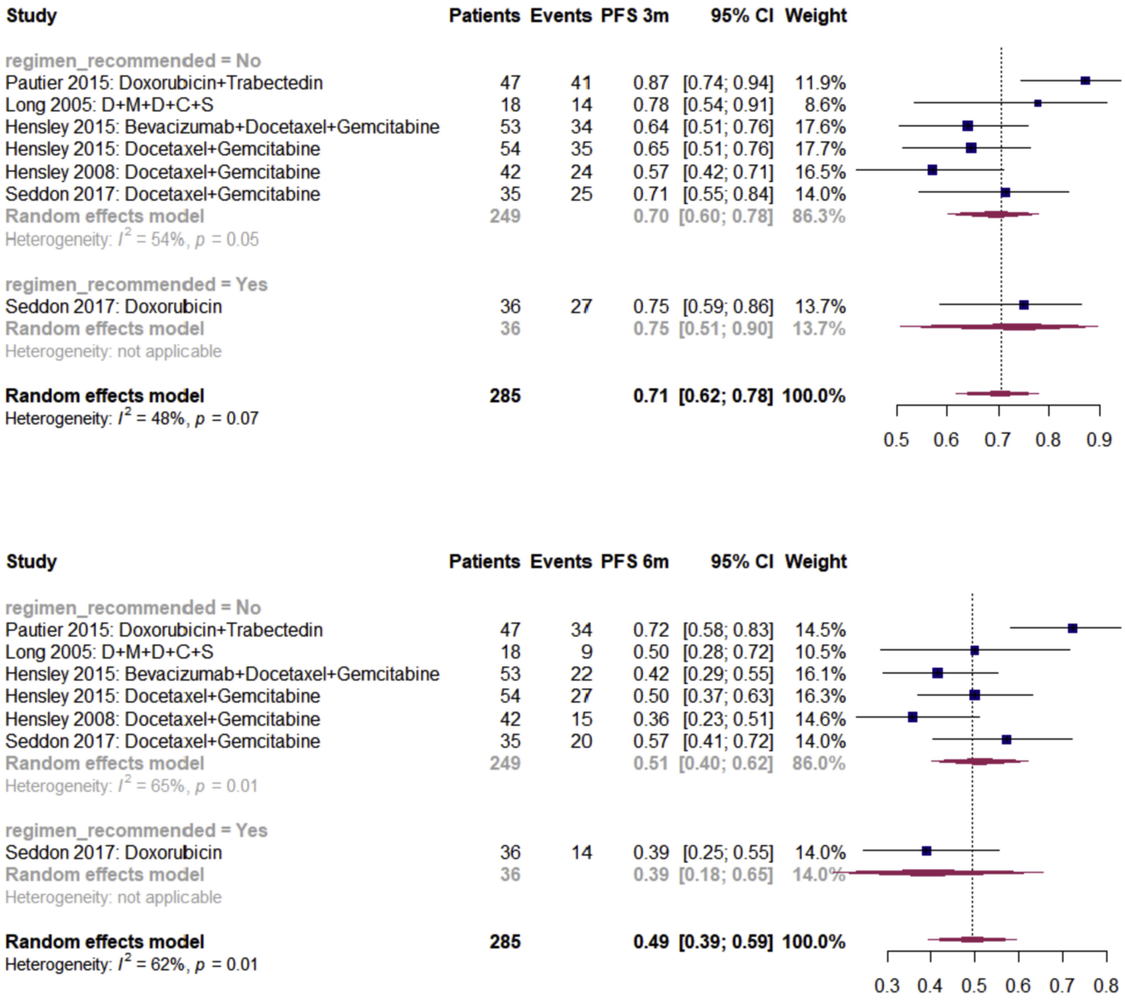


Figure 2.4: Forest plots of PFS at 3 (upper panel) and 6 (low panel) months for first-line uterine LMS patients.

2.3.6 Sensitivity analyses

Baujat plots for all LMS identified ‘Gelderblom 2014: Brostallicin’ as potentially influential for first-line analyses (pooled PFSR at 3 and 6 months increased 4% and 3% if this treatment arm is excluded), and in the pre-treated population ‘Schuetze 2012: Cyclophosphamide+Sirolimus’ (rate decreases 1% if excluded) and ‘Schöffski 2016: Dacarbazine’ (rate increases 1% if excluded) at 3 and 6 months, respectively [38, 42, 46]. Removing these treatment arms reduced overall heterogeneity insignificantly. The results in the first-line setting were less robust to the potential outlier than those in the pre-treated setting. Sensitivity analyses specific to uLMS showed low robustness because of the small sample size (seven treatment arms in first line and five in pre-treated). Baujat plots and forest plots removing potential outliers are provided in the Appendix sections 2.3 for all LMS and 2.4 for uLMS.

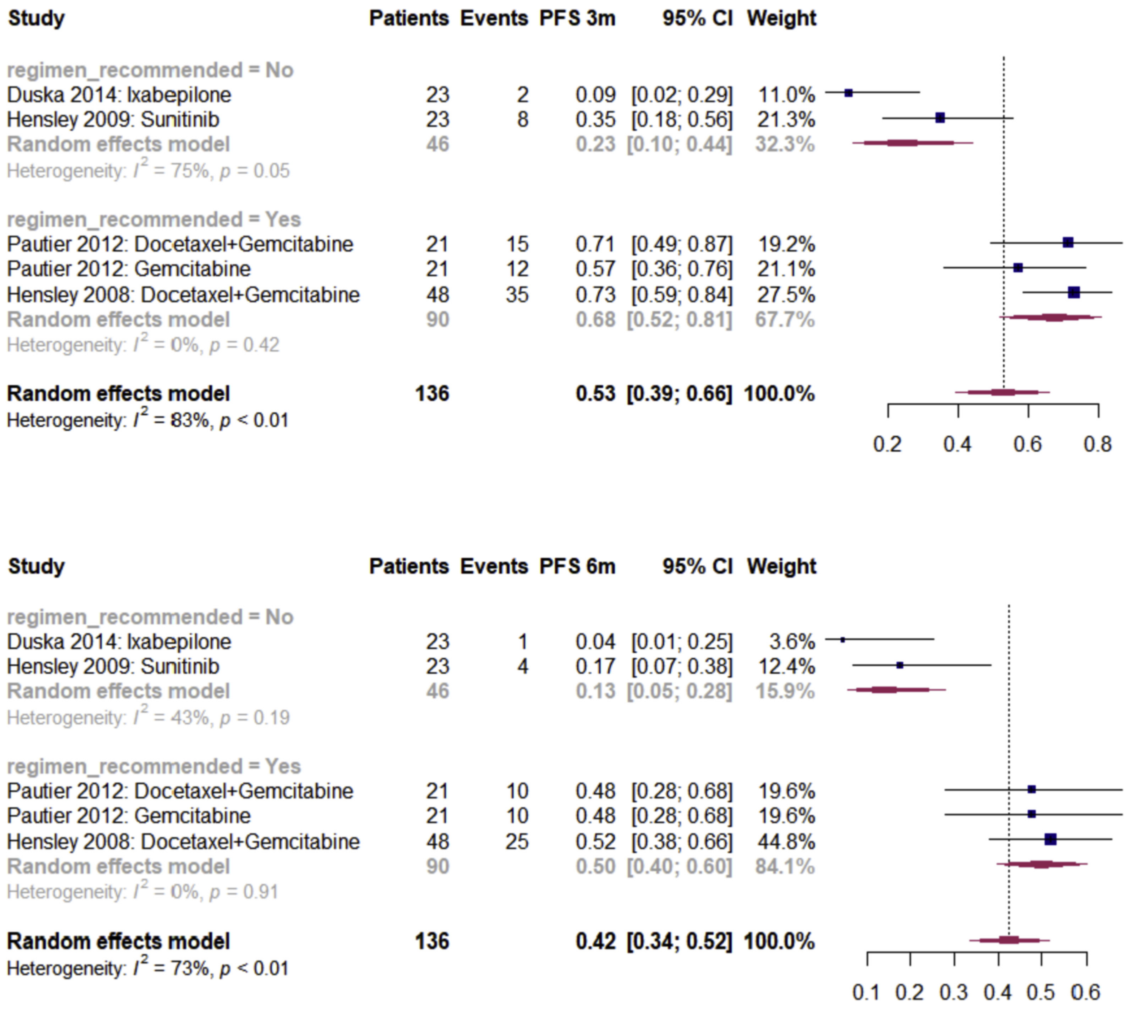


Figure 2.5: Forest plots of PFS at 3 (upper panel) and 6 (low panel) months for pre-treated uterine LMS patients.

2.3.7 Benchmarking

To derive the new benchmark for the LMS cohorts, our proposal is to use the overall pooled PFSR estimated from our analysis as reference value for the null hypothesis (H_0) parameter P_0. This choice is guided by the fact that there was no significant difference between R-T and NR-T for all LMS patients but can also be justified that future agents should do better than those currently available. As the ESMO-MCBS recommends a hazard ratio (HR) of at least 0.65 for PFS in advanced or metastatic setting (scale evaluation form 2b) [28], the reference value for the alternative hypothesis (H_1) parameter P_1 is estimated to detect an effect size of HR = 0.65. Table 2.2 summarises the P0 and P1 parameters. A PFSR-3m ≥82% or a PFSR-6m ≥70% (80% and 63% for uLMS) can be considered to suggest drug activity in first-line studies. For two or further lines, the recommended thresholds are 62% and 44% (66% and 57% for uLMS) at 3 and 6 months, respectively.

It should be underlined that if the minimum required level of efficacy is P_1, the design of the phase II trial focuses on demonstrating that this level is plausible, given the trial results and the efficacy is greater than P_0. In other words, the new agent deserves further testing at the end of the phase II trial if the estimated CI does not contain P_0. Following the ESMO-MCBS guidelines, the estimated CI should also encompass P_1. An example is provided in Figure 2.6.

	3 months		6 months	
Treatment line and analysed group	**Ref (P_0)**	**Min target (P_1)**	**Ref (P_0)**	**Min target (P_1)**
First-line uterine LMS	71%	80%	49%	63%
First line all LMS	74%	82%	58%	70%
Pre-treated uterine LMS	53%	66%	42%	57%
Pre-treated all LMS	48%	62%	28%	44%

Table 2.2: **Treatment effect (PFSR) for the null hypothesis (H_0) parameter P_0 and the alternative hypothesis (H_1) parameter P_1 of a study for LMS**. LMS, leiomyosarcoma. Reference values for P_0 are the overall pooled PFSR at 3 and 6 months. Minimum values to target for P_1 are calculated using the recommended treatment effect for PFS by the ESMO Magnitude of Clinical Benefit Scale (MCBS) [28].

2.4 Discussion

In the present study, we provided updated thresholds for PFS rates to be used for the design of clinical trials in advanced/metastatic and inoperable LMS by a meta-analysis of available data from clinical trials published between 2003 and 2018. Reference values for H_0 and H_1 have been estimated using the ESMO-MCBS recommendations [28].

The historical benchmarking analysis by Van Glabbeke et al. (2002) provided pooled progression-free rates for various STS patients who participated in phase II trials [16]. Notably, these have been used to design a large number of new studies. The results and thresholds cannot be directly compared for several reasons: Our meta-analysis focused on defining thresholds for LMS patients using phase II and phase III trials. In addition, most of the phase II trials included in the 2002 publication were conducted before the classification of GIST as a separate entity, and GIST patients were consequently classified as LMS patients. The primary end-point shifted from progression-free rates to PFSR, counting any death as an event. Van Glabbeke et al. exploited individual patient data (IPD, N = 1534 overall) from the STBSG database, whereas we used summary estimates, which are less reliable than IPD. On the other hand, we were able to conduct a meta-analysis including over 1500 LMS patients.

We chose not to meta-analyse other common end-points in clinical trials, such as RR and OS. Here, rather low objective RRs were obtained for the majority of the drugs/drug combinations in our LMS database (several times 0%, frequently less than 15%), which is expected in this population as a decrease of tumour volume greater than 30% (needed to qualify a partial response according to RECIST 1.1 [56] is unlikely with the studied agents. Hence, RR is not the best end-point for simple screening phase II studies in LMS as a basis for further drug development. Furthermore, OS is usually not the primary end-point in phase II studies. On the contrary, PFS (and/or time to progression) is a valuable alternative end-point for the estimation of the biological antitumor activity of a new treatment and thus to justify further investigation in phase III trials. An extensive discussion is provided in the Van Glabbeke paper [16].

Thresholds were defined for all LMS and were shown to be robust by sensitivity analysis. A uLMS-specific subgroup meta-analysis was performed. The results should be interpreted with caution because of the potential publication bias indicated in this subanalysis and the small sample size (seven rows from five trials for first line and five rows from four trials for pre-treated population).

This analysis showed that R-T based on standard clinical practice guidelines do not necessarily exhibit a significant

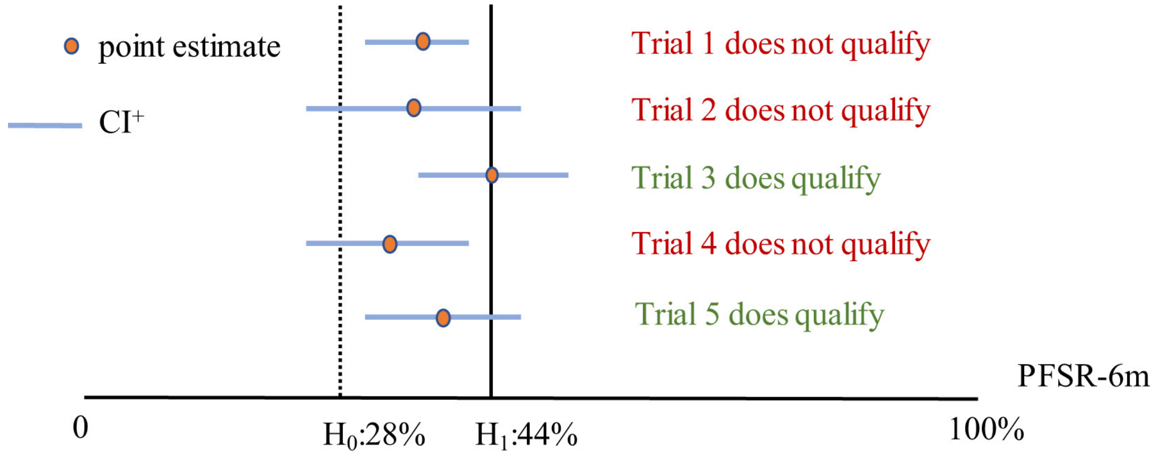


Figure 2.6: Example regarding the thresholds estimated for the PFS rate at 6 months of pre-treated all LMS patients. The parameter of null hypothesis (P_0) was calculated at 28% and the parameter of the alternative hypothesis at 44%. Trial 1 does not qualify because the point estimate or the upper limit of the CI do not reach 44% (P_1). Trial 2 does not qualify because the lower limit of the CI does not surpass 28% (P_0). Trial 3 does qualify because the point estimate reaches P1 and the lower limit of the CI surpasses P_0. Trial 4 does not qualify because the lower limit of the CI does not surpass P_0 and the point estimate or the upper limit of the CI do not reach P_1. Trial 5 does qualify because the lower limit of the CI surpasses P_0 and the upper limit of the CI surpasses P_1.
+ The confidence level of the confidence interval (CI) is to be defined based on the statistical parameters of the study design.

difference in PFSR at 3/6 months versus NR-T for advanced or metastatic LMS, apart from the pre-treated setting for uLMS [7]. This could be explained by the fact that the majority of the trials used as a basis for the clinical practice guidelines were designed for multiple STS subtypes and as a result are underpowered for specific subgroup analyses. They did therefore not lead to specific recommendations.

To the best of our knowledge, this is the first attempt at a meta-analysis of the outcome of patients with advanced or metastatic LMS for both first and further lines. Overall, 1500 patients were included in the analysis for all LMS and 421 patients for uLMS, which is a key strength of this work. A meta-regression was performed to investigate whether the phase of the trial, study design, year of activation, and sample size are prognostic for PFSR separately and if they can mitigate heterogeneity. Sample size was prognostic and could explain a small part of residual heterogeneity (variability between study outcomes not accounted for by the variables) for first line at 3 months and year of activation a larger part for pre-treated population at 6 months. For uLMS patients, meta-regression was not performed because of the limited number of therapeutic combinations. Future research should shed light to whether other factors could explain heterogeneity across studies.

A condition of any meta-analysis is the implied independence of effect sizes between drugs of the same trial [23, 57]. In our meta-analysis, a random effects model was used for each treatment regimen in the database and not for each trial. However, for randomised studies (10/23 trials for all LMS), there might be some dependence, as treatment arms were designed for the same patient population/centres. And finally, a source of bias is the use of progression-free rate instead of PFSR for 4/31 treatment regimens, as the required data could not be retrieved. This could lead to a small overestimation of the overall PFSR, as deaths are not taken into account at 3 and 6 months in these four regimens.

Last but not least, the ultimate aim of a clinical trial is to provide evidence of improved OS or improved quality of life. Nonetheless, two recent meta-analyses do not support strong surrogacy properties between PFS and OS in advanced STS randomised clinical trials [58, 59]. Consequently, PFS carries the risk of misleading conclusions because of erroneous extrapolation of the results. On the other hand, PFS remains an attractive end-point to identify benefit earlier than OS, and phase II trials are not intended to provide definite proof of the new treatment but rather a justification to further investigation. PFS (or PFSR-3m, PFSR-6m) can thus be used as primary end-points in phase II trials or as futility end-points in phase III trials, but OS should remain the primary end-point in phase III trials (whenever possible).

In conclusion, last decade research in STS shifted to a histology-specific approach. Because of the unmet medical need in standard of care alternatives, new studies tailoring therapy to specific histological subtypes should be based on modern thresholds for drug activity. Hereto, we suggest a new benchmark for designing phase II studies for all LMS or uLMS using the overall PFSR-3m and PFSR-6m as primary end-point. Future research is warranted using similar methodology to update thresholds of other common STS subgroups (e.g. liposarcomas).

Declarations

Role of the funding source

This work was supported by the European Organisation for Research and Treatment of Cancer – Soft Tissue and Bone Sarcoma Group (EORTC – STBSG). The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Acknowledgements

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 as well as from the EORTC Cancer Research Fund (ECRF). This publication was supported by a donation from the Kom Op Tegen Kanker from Belgium. The authors would like to express our gratitude to all primary investigators who shared 3- and 6-month PFS estimates of their clinical trials with leiomyosarcoma patients for undertaking this research project, namely, Dr Beatrice Seddon (one study), Prof Armando Santoro (one study), Dr William D. Tap (one study), Prof Scott M. Schuetze (two studies), Dr Akira Kawai (one study), and Dr Amit M. Oza (one study). The authors thank NRG Oncology Statistics and Data Management Center (NRG SDMC) for the provision of data from the NRG/GOG 087K, 087L, 131H, 231C clinical trials related to uterine Leiomyosarcoma. This article was prepared using data from Dataset GOG-0250 from the NCTN Data Archive of the National Cancer Institute's (NCI's) National Clinical Trials Network (NCTN). Data were originally collected from clinical trial NCT01012297 'A Randomised Phase III Evaluation of Docetaxel (NSC #628503) and Gemcitabine (NSC #613327) Plus G-CSF With Bevacizumab (NSC #704865) Versus Docetaxel (NSC #628503) and Gemcitabine (NSC #613327) Plus G-CSF With Placebo in the Treatment of Recurrent or Advanced Leiomyosarcoma of the Uterus'. All analyses and conclusions in this article are the sole responsibility of the authors and do not necessarily reflect the opinions or views of the clinical trial investigators, the NCTN, or the NCI.

Online supplementary materials

The Appendix of this Chapter is available online at https://github.com/GKantidakis/Thesis_supplementary_materials/blob/main/Chapter2/Appendix.docx.

References

[1] 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.

[2] K. G. Billingsley, M. E. Burt, E. Jara, R. J. Ginsberg, J. M. Woodruff, D. H. Y. Leung, and M. F. Brennan. Pulmonary metastases from soft tissue sarcoma: Analysis of patterns of disease and postmetastasis survival.

Annals of Surgery, 229(5):602–612, 1999. ISSN 00034932. doi: 10.1097/00000658-199905000-00002. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1420804/.

[3] I. Judson, J. Verweij, H. Gelderblom, J. T. Hartmann, P. Schöffski, J. Y. Blay, J. M. Kerst, J. Sufliarsky, J. Whelan, P. Hohenberger, A. Krarup-Hansen, T. Alcindor, S. Marréaud, S. Litière, C. Hermans, C. Fisher, P. C. W. Hogendoorn, A. P. Dei Tos, and W. T. A. Van der Graaf. Doxorubicin alone versus intensified doxorubicin plus ifosfamide for first-line treatment of advanced or metastatic soft-tissue sarcoma: A randomised controlled phase 3 trial. *The Lancet Oncology*, 15(4):415–423, 2014. ISSN 14745488. doi: 10.1016/S1470-2045(14)70063-4.

[4] C. W. Ryan, O. Merimsky, M. Agulnik, J. Y. Blay, S. M. Schuetze, B. A. Van Tine, R. L. Jones, A. D. Elias, E. Choy, T. Alcindor, V. L. Keedy, D. R. Reed, R. N. Taub, A. Italiano, X. G. Del Muro, I. R. Judson, J. Y. Buck, F. Lebel, J. J. Lewis, R. G. Maki, and P. Schöffski. PICASSO III: A phase III, placebo-controlled study of doxorubicin with or without palifosfamide in patients with metastatic soft tissue sarcoma. *Journal of Clinical Oncology*, 34(32):3898–3905, 2016. ISSN 15277755. doi: 10.1200/JCO.2016.67.6684.

[5] S. P. Chawla, L. D. Cranmer, B. A. Van Tine, D. R. Reed, S. H. Okuno, J. E. Butrynski, D. R. Adkins, A. E. Hendifar, S. Kroll, and K. N. Ganjoo. Phase II study of the safety and antitumor activity of the hypoxia-activated prodrug TH-302 in combination with doxorubicin in patients with advanced soft tissue sarcoma. *Journal of Clinical Oncology*, 32(29):3299–3306, oct 2014. ISSN 15277755. doi: 10.1200/JCO.2013.54.3660.

[6] J. M. Buesa, H. T. Mouridsen, A. T. Van Oosterom, J. Verweij, T. Wagener, W. Steward, A. Poveda, P. M. Vestlev, D. Thomas, and R. Sylvester. High-dose DTIC in advanced soft-tissue sarcomas in the adult: A phase II study of the E.O.R.T.C. soft tissue and bone Sarcoma group. *Annals of Oncology*, 2(4):307–309, 1991. ISSN 09237534. doi: 10.1093/oxfordjournals.annonc.a057942. URL https://doi.org/10.1093/oxfordjournals.annonc.a057942.

[7] P. G. Casali, N. Abecassis, H. T. Aro, S. Bauer, R. Biagini, S. Bielack, S. Bonvalot, I. Boukovinas, J. V. M. G. Bovee, T. Brodowicz, J. M. Broto, A. Buonadonna, E. De Álava, A. P. Dei Tos, X. G. Del Muro, P. Dileo, M. Eriksson, A. Fedenko, V. Ferraresi, A. Ferrari, S. Ferrari, A. M. Frezza, S. Gasperoni, H. Gelderblom, T. Gil, G. Grignani, A. Gronchi, R. L. Haas, B. Hassan, P. Hohenberger, R. Issels, H. Joensuu, R. L. Jones, I. Judson, P. Jutte, S. Kaal, B. Kasper, K. Kopeckova, D. A. Krákorová, A. Le Cesne, I. Lugowska, O. Merimsky, M. Montemurro, M. A. Pantaleo, R. Piana, P. Picci, S. Piperno-Neumann, A. L. Pousa, P. Reichardt, M. H. Robinson, P. Rutkowski, A. A. Safwat, P. Schöffski, S. Sleijfer, S. Stacchiotti, K. Sundby Hall, M. Unk, F. Van Coevorden, W. T. A. Van Der Graaf, J. Whelan, E. Wardelmann, O. Zaikova, and J. Y. Blay. Soft tissue and visceral sarcomas: ESMO-EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 29(Supplement_4):iv51—-iv67, 2018. ISSN 15698041. doi: 10.1093/annonc/mdy321.

[8] X. García-del Muro, A. López-Pousa, J. Maurel, J. Martín, J. Martínez-Trufero, A. Casado, A. Gómez-España, J. Fra, J. Cruz, A. Poveda, A. Meana, C. Pericay, R. Cubedo, J. Rubió, A. De Juan, N. Laínez, J. A. Carrasco, R. De Andrés, and J. M. Buesa. Randomized phase II study comparing gemcitabine plus dacarbazine versus dacarbazine alone in patients with previously treated soft tissue sarcoma: A Spanish group for research on sarcomas study. *Journal of Clinical Oncology*, 29(18):2528–2533, jun 2011. ISSN 15277755. doi: 10.1200/JCO.2010.33.6107.

[9] 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, dec 2016. ISSN 14745488. doi: 10.1016/S1470-2045(16)30507-1.

[10] X. Guo, V. Y. Jo, A. M. Mills, S. X. Zhu, C. H. Lee, I. Espinosa, M. R. Nucci, S. Varma, E. Forgó, T. Hastie, S. Anderson, K. Ganjoo, A. H. Beck, R. B. West, C. D. Fletcher, and M. Van De Rijn. Clinically relevant molecular subtypes in leiomyosarcoma. *Clinical Cancer Research*, 21(15):3501–3511, 2015. ISSN 15573265. doi: 10.1158/1078-0432.CCR-14-3141.

[11] D. Y. S. Kuo, P. Timmins, S. V. Blank, A. L. Fields, G. L. Goldberg, A. Murgo, P. Christos, S. Wadler, and C. D. Runowicz. Phase II trial of thalidomide for advanced and recurrent gynecologic sarcoma: A brief communication from the New York Phase II consortium. *Gynecologic Oncology*, 100(1):160–165, jan 2006. ISSN 00908258. doi: 10.1016/j.ygyno.2005.08.033.

[12] N. Penel, A. Italiano, N. Isambert, E. Bompas, G. Bousquet, and F. Duffaud. Factors affecting the outcome of patients with metastatic leiomyosarcoma treated with doxorubicin-containing chemotherapy. *Annals of Oncology*, 21(6):1361–1365, 2009. ISSN 15698041. doi: 10.1093/annonc/mdp485.

[13] 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.

[14] A. W. Oosten, C. Seynaeve, P. I.M. Schmitz, M. A. Den Bakker, J. Verweij, and S. Sleijfer. Outcomes of first-line chemotherapy in patients with advanced or metastatic leiomyosarcoma of uterine and non-uterine origin. *Sarcoma*, 2009(348910), 2009. ISSN 1357714X. doi: 10.1155/2009/348910.

[15] E. Ben-Ami, C. M. Barysauskas, S. Solomon, K. Tahlil, R. Malley, M. Hohos, K. Polson, M. Loucks, M. Severgnini, T. Patel, A. Cunningham, S. J. Rodig, F. S. Hodi, J. A. M., P. Merriam, A. J. Wagner, G. I. Shapiro, and S. George. Immunotherapy with single agent nivolumab for advanced leiomyosarcoma of the uterus: Results of a phase 2 study. *Cancer*, 123(17):3285–3290, sep 2017. ISSN 10970142. doi: 10.1002/cncr.30738.

[16] 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.

[17] V. Y. Jo and C. D. M. Fletcher. WHO classification of soft tissue tumours: An update based on the 2013 (4th) edition. *Pathology*, 46(2):95–104, 2014. ISSN 14653931. doi: 10.1097/PAT.0000000000000050.

[18] W. D. Tap, Z. Papai, B. A. Van Tine, S. Attia, K. N. Ganjoo, R. L. Jones, S. Schuetze, D. Reed, S. P. Chawla, R. F. Riedel, A. Krarup-Hansen, M. Toulmonde, I. Ray-Coquard, P. Hohenberger, G. Grignani, L. D. Cranmer, S. Okuno, M. Agulnik, W. Read, C. W. Ryan, T. Alcindor, X. F. G. del Muro, G. T. Budd, H. Tawbi, T. Pearce, S. Kroll, D. K. Reinke, and P. Schöffski. Doxorubicin plus evofosfamide versus doxorubicin alone in locally advanced, unresectable or metastatic soft-tissue sarcoma (TH CR-406/SARC021): an international, multicentre, open-label, randomised phase 3 trial. *The Lancet Oncology*, 18(8):1089–1103, aug 2017. ISSN 14745488. doi: 10.1016/S1470-2045(17)30381-9.

[19] A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. A. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen, and D. Moher. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of clinical epidemiology*, 62(10):e1—-e34, 2009. ISSN 18785921. doi: 10.1016/j.jclinepi.2009.06.006.

[20] W. Feller. On the Normal Approximation to the Binomial Distribution. *The Annals of Mathematical Statistics*, 16(4):319–329, 1945. ISSN 0003-4851. doi: 10.1214/aoms/1177731058. URL https://projecteuclid.org/euclid.aoms/1177731058.

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

[22] 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.

[23] 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.

[24] J. L. Peters, A. J. Sutton, D. R. Jones, K. R. Abrams, and L. Rushton. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology*, 61(10):991–996, oct 2008. ISSN 08954356. doi: 10.1016/j.jclinepi.2007.11.010.

[25] M. Egger, G. D. Smith, M. Schneider, and C. Minder. Bias in meta-analysis detected by a simple, graphical test measures of funnel plot asymmetry. *BMJ*, 315(7109):629–634, 1997. doi: 10.1136/bmj.315.7109.629.

[26] C. B. Begg and M. Mazumdar. Operating Characteristics of a Rank Correlation Test for Publication Bias. *Biometrics*, 50(4):1088–1101, 1994. doi: 10.2307/2533446. URL https://www.jstor.org/stable/pdf/2533446.pdf.

[27] B. Baujat, C. Mahé, J. P. Pignon, and C. Hill. A graphical method for exploring heterogeneity in meta-analyses: Application to a meta-analysis of 65 trials. *Statistics in Medicine*, 21(18):2641–2652, 2002. doi: 10.1002/sim.1221.

[28] N. I. Cherny, U. Dafni, J. Bogaerts, N. J. Latino, G. Pentheroudakis, J. Y. Douillard, J. Tabernero, C. Zielinski, M. J. Piccart, and E. G. E. de Vries. ESMO-Magnitude of Clinical Benefit Scale version 1.1. *Annals of Oncology*, 28(10):2340–2366, 2017. doi: 10.1093/annonc/mdx310.

[29] W. Viechtbauer. Conducting meta-analyses in R with the metafor. *Journal of Statistical Software*, 36(3): 1–48, 2010. ISSN 15487660.

[30] G. Swarzer. meta: an R package for meta-analysis. *R News*, 7(3):40–45, 2007. ISSN 1609-3631. URL https://www.researchgate.net/publication/285729385{_}meta{_}An{_}R{_}Package{_}for{_}Meta-Analysis.

[31] J. T. Hartmann, F. Mayer, J. Schleicher, M. Horger, J. Huober, I. Meisinger, J. Pintoffl, G. Käfer, L. Kanz, and V. Grünwald. Bendamustine hydrochloride in patients with refractory soft tissue sarcoma: A noncomparative multicenter phase 2 study of the German sarcoma group (AIO-001). *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 110(4):861–866, 2007. ISSN 0008543X. doi: 10.1002/cncr.22846.

[32] P. Reichardt, O. S. Nielsen, S. Bauer, J. T. Hartmann, P. Schöffski, T. B. Christensen, D. Pink, S. Daugaard, S. Marréaud, M. Van Glabbeke, and J. Y. Blay. Exatecan in pretreated adult patients with advanced soft tissue sarcoma: Results of a phase II - Study of the EORTC Soft Tissue and Bone Sarcoma Group. *European Journal of Cancer*, 43(6):1017–1022, 2007. ISSN 09598049. doi: 10.1016/j.ejca.2007.01.014.

[33] S. Sleijfer, I. Ray-Coquard, Z. Papai, A. Le Cesne, M. Scurr, P. Schöffski, F. Collin, L. Pandite, S. Marréaud, A. De Brauwer, M. Van Glabbeke, J. Verweij, and J. Y. Blay. Pazopanib, a multikinase angiogenesis inhibitor, in patients with relapsed or refractory advanced soft tissue sarcoma: A phase II study from the European organisation for research and treatment of cancer-soft tissue and bone sarcoma group (EORTC Study 620. *Journal of Clinical Oncology*, 27(19):3126–3132, 2009. ISSN 0732183X. doi: 10.1200/JCO.2008.21.3223.

[34] P. Schöffski, I. L. Ray-Coquard, A. Cioffi, N. B. Bui, S. Bauer, J. T. Hartmann, A. Krarup-Hansen, V. Grünwald, R. Sciot, H. Dumez, J. Y. Blay, A. Le Cesne, J. Wanders, C. Hayward, S. Marréaud, M. Ouali, and P. Hohenberger. Activity of eribulin mesylate in patients with soft-tissue sarcoma: A phase 2 study in four independent histological subtypes. *The Lancet Oncology*, 12(11):1045–1052, 2011. ISSN 14702045. doi: 10.1016/S1470-2045(11)70230-3.

[35] S. P. Chawla, A. P. Staddon, L. H. Baker, S. M. Schuetze, A. W. Tolcher, G. Z. D'Amato, J. Y. Blay, M. M. Mita, K. K. Sankhala, L. Berk, V. M. Rivera, T. Clackson, J. W. Loewy, F. G. Haluska, and G. D. Demetri. Phase II study of the mammalian target of rapamycin inhibitor ridaforolimus in patients with advanced bone

and soft tissue sarcomas. *Journal of Clinical Oncology*, 30(1):78–84, 2012. ISSN 15277755. doi: 10.1200/JCO.2011.35.6329.

[36] W. T. A. Van Der Graaf, J. Y. Blay, S. P. Chawla, D. W. Kim, B. Bui-Nguyen, P. G. Casali, P. Schöffski, M. Aglietta, A. P. Staddon, Y. Beppu, A. Le Cesne, H. Gelderblom, I. R. Judson, N. Araki, M. Ouali, S. Marréaud, R. Hodge, M. R. Dewji, C. Coens, G. D. Demetri, C. D. Fletcher, A. P. Dei Tos, and P. Hohenberger. Pazopanib for metastatic soft-tissue sarcoma (PALETTE): A randomised, double-blind, placebo-controlled phase 3 trial. *The Lancet*, 379(9829):1879–1886, 2012. ISSN 1474547X. doi: 10.1016/S0140-6736(12)60651-5.

[37] P. Pautier, A. Floquet, N. Penel, S. Piperno-Neumann, N. Isambert, A. Rey, E. Bompas, A. Cioffi, C. Delcambre, D. Cupissol, F. Collin, J. Y. Blay, M. Jimenez, and F. Duffaud. Randomized Multicenter and Stratified Phase II Study of Gemcitabine Alone Versus Gemcitabine and Docetaxel in Patients with Metastatic or Relapsed Leiomyosarcomas: A Federation Nationale des Centres de Lutte Contre le Cancer (FN-CLCC) French Sarcoma Group . *The Oncologist*, 17(9):1213–1220, sep 2012. ISSN 1083-7159. doi: 10.1634/theoncologist.2011-0467.

[38] S. M. Schuetze, L. Zhao, R. Chugh, D. G. Thomas, D. R. Lucas, G. Metko, M. M. Zalupski, and L. H. Baker. Results of a phase II study of sirolimus and cyclophosphamide in patients with advanced sarcoma. *European Journal of Cancer*, 48(9):1347–1353, 2012. ISSN 09598049. doi: 10.1016/j.ejca.2012.03.022. URL http://dx.doi.org/10.1016/j.ejca.2012.03.022.

[39] P. A. Cassier, A. Lefranc, E. Y Amela, C. Chevreau, B. N. Bui, A. Lecesne, I. Ray-Coquard, S. Chabaud, N. Penel, Y. Berge, J. Dômont, A. Italiano, F. Duffaud, A. C. Cadore, V. Polivka, and J. Y. Blay. A phase II trial of panobinostat in patients with advanced pretreated soft tissue sarcoma. A study from the French Sarcoma Group. *British Journal of Cancer*, 109(4):909–914, 2013. ISSN 00070920. doi: 10.1038/bjc.2013.442.

[40] A. Santoro, A. Comandone, U. Basso, H. Soto Parra, R. De Sanctis, E. Stroppa, I. Marcon, L. Giordano, F. R. Lutman, A. Boglione, and A. Bertuzzi. Phase II prospective study with sorafenib in advanced soft tissue sarcomas after anthracycline-based therapy. *Annals of Oncology*, 24(4):1093–1098, 2013. ISSN 09237534. doi: 10.1093/annonc/mds607.

[41] P. Schöffski, D. Adkins, J. Y. Blay, T. Gil, A. D. Elias, P. Rutkowski, G. K. Pennock, H. Youssoufian, H. Gelderblom, R. Willey, and D. O. Grebennik. An open-label, phase 2 study evaluating the efficacy and safety of the anti-IGF-1R antibody cixutumumab in patients with previously treated advanced or metastatic soft-tissue sarcoma or Ewing family of tumours. *European Journal of Cancer*, 49(15):3219–3228, 2013. ISSN 09598049. doi: 10.1016/j.ejca.2013.06.010.

[42] H. Gelderblom, J. Y. Blay, B. M. Seddon, M. Leahy, I. Ray-Coquard, S. Sleijfer, J. M. Kerst, P. Rutkowski, S. Bauer, M. Ouali, S. Marréaud, R. J. H. M. Van Der Straaten, H. J. Guchelaar, S. D. Weitman, P. C. W. Hogendoorn, and P. Hohenberger. Brostallicin versus doxorubicin as first-line chemotherapy in patients with advanced or metastatic soft tissue sarcoma: An European Organisation for Research and Treatment of Cancer Soft Tissue and Bone Sarcoma Group randomised phase II and pharmacogeneti. *European Journal of Cancer*, 50(2):388–396, 2014. ISSN 09598049. doi: 10.1016/j.ejca.2013.10.002.

[43] B. Bui-Nguyen, J. E. Butrynski, N. Penel, J. Y. Blay, N. Isambert, M. Milhem, J. M. Kerst, A. K. L. Reyners, S. Litière, S. Marréaud, F. Collin, and W. T. A. Van Der Graaf. A phase IIb multicentre study comparing the efficacy of trabectedin to doxorubicin in patients with advanced or metastatic untreated soft tissue sarcoma: The TRUSTS trial. *European Journal of Cancer*, 51(10):1312–1320, 2015. ISSN 18790852. doi: 10.1016/j.ejca.2015.03.023.

[44] Z. Eroglu, H. A. Tawbi, J. Hu, M. Guan, P. H. Frankel, N. H. Ruel, S. Wilczynski, S. Christensen, D. R. Gandara, and W. A. Chow. A randomised phase II trial of selumetinib vs selumetinib plus temsirolimus for soft-tissue sarcomas. *British Journal of Cancer*, 112(10):1644–1651, 2015. ISSN 15321827. doi: 10.1038/bjc.2015.126.

[45] P. Pautier, A. Floquet, C. Chevreau, N. Penel, C. Guillemet, C. Delcambre, D. Cupissol, F. Selle, N. Isambert, S. Piperno-Neumann, A. Thyss, F. Bertucci, E. Bompas, J. Alexandre, O. Collard, S. Lavau-Denes, P. Soulié, M. Toulmonde, A. Le Cesne, B. Lacas, and F. Duffaud. Trabectedin in combination with doxorubicin for first-line treatment of advanced uterine or soft-tissue leiomyosarcoma (LMS-02): A non-randomised, multicentre, phase 2 trial. *The Lancet Oncology*, 16(4):457–464, 2015. ISSN 14745488. doi: 10.1016/S1470-2045(15) 70070-7.

[46] P. Schöffski, S. Chawla, R. G. Maki, A. Italiano, H. Gelderblom, E. Choy, G. Grignani, V. Camargo, S. Bauer, S. Y. Rha, J. Y. Blay, P. Hohenberger, D. D'Adamo, M. Guo, B. Chmielowski, A. Le Cesne, G. D. Demetri, and S. R. Patel. Eribulin versus dacarbazine in previously treated patients with advanced liposarcoma or leiomyosarcoma: A randomised, open-label, multicentre, phase 3 trial. *The Lancet*, 387(10028):1629–1637, 2016. ISSN 1474547X. doi: 10.1016/S0140-6736(15)01283-0.

[47] S. M. Schuetze, J. K. Wathen, D. R. Lucas, E. Choy, B. L. Samuels, A. P. Staddon, K. N. Ganjoo, M. Von Mehren, W. A. Chow, D. M. Loeb, H. A. Tawbi, D. A Rushing, S. R. Patel, D. G. Thomas, R. Chugh, D. K. Reinke, and L. H. Baker. SARC009: Phase 2 study of dasatinib in patients with previously treated, high-grade, advanced sarcoma. *Cancer*, 122(6):868–874, 2016. ISSN 10970142. doi: 10.1002/cncr.29858.

[48] A. Kawai, N. Araki, Y. Naito, T. Ozaki, H. Sugiura, Y. Yazawa, H. Morioka, A. Matsumine, K. Saito, S. Asami, and K. Isu. Phase 2 study of eribulin in patients with previously treated advanced or metastatic soft tissue sarcoma. *Japanese Journal of Clinical Oncology*, 47(2):137–144, 2017. ISSN 14653621. doi: 10.1093/jjco/hyw175.

[49] B. Seddon, S. J. Strauss, J. Whelan, M. Leahy, P. J. Woll, F. Cowie, C. Rothermundt, Z. Wood, C. Benson, N. Ali, M. Marples, G. J. Veal, D. Jamieson, K. Küver, R. Tirabosco, S. Forsyth, S. Nash, H. M. Dehbi, and S. Beare. Gemcitabine and docetaxel versus doxorubicin as first-line treatment in previously untreated advanced unresectable or metastatic soft-tissue sarcomas (GeDDiS): a randomised controlled phase 3 trial. *The Lancet Oncology*, 18(10):1397–1410, 2017. ISSN 14745488. doi: 10.1016/S1470-2045(17)30622-8.

[50] H. J. Long III, J. A. Blessing, and J. Sorosky. Phase II trial of dacarbazine, mitomycin, doxorubicin, and cisplatin with sargramostim in uterine leiomyosarcoma: A gynecologic oncology group study. *Gynecologic Oncology*, 99(2):339–342, 2005. ISSN 00908258. doi: 10.1016/j.ygyno.2005.06.002.

[51] M. L. Hensley, J. A. Blessing, R. Mannel, and P. G. Rose. Fixed-dose rate gemcitabine plus docetaxel as first-line therapy for metastatic uterine leiomyosarcoma: A Gynecologic Oncology Group phase II trial. *Gynecologic Oncology*, 109(3):329–334, 2008. ISSN 00908258. doi: 10.1016/j.ygyno.2008.03.010.

[52] M. L. Hensley, M. W. Sill, D. R. Scribner, J. Brown, R. L. DeBernardo, E. M. Hartenbach, C. K. McCourt, J. R. Bosscher, and P. A. Gehrig. Sunitinib malate in the treatment of recurrent or persistent uterine leiomyosarcoma: A Gynecologic Oncology Group phase II study. *Gynecologic Oncology*, 115(3):460–465, 2009. ISSN 00908258. doi: 10.1016/j.ygyno.2009.09.011.

[53] M. L. Hensley, J. A. Blessing, K. DeGeest, O. Abulafia, P. G. Rose, and H. D. Homesley. Fixed-dose rate gemcitabine plus docetaxel as second-line therapy for metastatic uterine leiomyosarcoma: A Gynecologic Oncology Group phase II study. *Gynecologic Oncology*, 109(3):323–328, jun 2008. ISSN 0090-8258. doi: 10.1016/J.YGYNO.2008.02.024. URL `http://www.gynecologiconcology-online.net/article/S0090825808001765/fulltext`.

[54] L. R. Duska, J. A. Blessing, J. Rotmensch, R. S. Mannel, P. Hanjani, P. G. Rose, and D. S. Dizon. A Phase II evaluation of ixabepilone (IND #59699, NSC #710428) in the treatment of recurrent or persistent leiomyosarcoma of the uterus: An NRG Oncology/Gynecologic Oncology Group Study. *Gynecologic Oncology*, 135 (1):44–48, 2014. ISSN 10956859. doi: 10.1016/j.ygyno.2014.07.101.

[55] M. L. Hensley, A. Miller, D. M. O'Malley, R. S. Mannel, K. Behbakht, J. N. Bakkum-Gamez, and H. Michael. Randomized phase III trial of gemcitabine plus docetaxel plus bevacizumab or placebo as first-line treatment

for metastatic uterine leiomyosarcoma: An NRG oncology/ gynecologic oncology group study. *Journal of Clinical Oncology*, 33(10):1180–1185, 2015. ISSN 15277755. doi: 10.1200/JCO.2014.58.3781.

[56] E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer*, 45(2):228–247, 2009. ISSN 09598049. doi: 10.1016/j.ejca.2008.10.026. URL http://dx.doi.org/10.1016/j.ejca.2008.10.026.

[57] S. Nakagawa, D. W. A. Noble, A. M. Senior, and M. Lagisz. Meta-evaluation of meta-analysis: Ten appraisal questions for biologists. *BMC Biology*, 15(1):1–14, 2017. ISSN 17417007. doi: 10.1186/s12915-017-0357-7.

[58] M. Savina, S. Litière, A. Italiano, T. Burzykowski, F. Bonnetain, S. Gourgou, V. Rondeau, J. Y. Blay, S. Cousin, F. Duffaud, H. Gelderblom, A. Gronchi, I. Judson, A. Le Cesne, P. Lorigan, J. Maurel, W. T. A. Van Der Graaf, J. Verweij, S. Mathoulin-Pélissier, and C. Bellera. Surrogate endpoints in advanced sarcoma trials: A meta-analysis. *Oncotarget*, 9(77):34617–34627, 2018. ISSN 19492553. doi: 10.18632/oncotarget.26166.

[59] K. Tanaka, M. Kawano, T. Iwasaki, I. Itonaga, and H. Tsumura. Surrogacy of intermediate endpoints for overall survival in randomized controlled trials of first-line treatment for advanced soft tissue sarcoma in the pre- and post-pazopanib era: A meta-analytic evaluation. *BMC Cancer*, 19(1):1–9, 2019. ISSN 14712407. doi: 10.1186/s12885-019-5268-2.

3

New benchmarks to design clinical trials with advanced or metastatic liposarcoma or synovial sarcoma patients: A EORTC - Soft Tissue and Bone Sarcoma Group (STBSG) meta-analysis based on a literature review for soft-tissue sarcomas

This chapter is based on: G. Kantidakis, S. Litière, A. Neven, M. Vinches, I. Judson, J. Y. Blay, E. Wardelmann, S. Stacchiotti, L. D'Ambrosio, S. Marréaud, W. T. A. van der Graaf, B. Kasper, M. Fiocco, and H. Gelderblom. New benchmarks to design clinical trials with advanced or metastatic liposarcoma or synovial sarcoma patients: A EORTC - Soft Tissue and Bone Sarcoma Group (STBSG) meta-analysis based on a literature review for soft-tissue sarcomas. *European Journal of Cancer*, 174:261-276, 2022.

Abstract

Background: Recently, we performed a meta-analysis based on a literature review for STS trials (published 2003–2018, ≥ 10 adult patients) to update long-standing reference values for leiomyosarcomas. This work is extended for liposarcomas (LPS) and synovial sarcomas (SS).

Materials and methods: Study endpoints were progression-free survival rates (PFSRs) at 3 and 6 months. Trial-specific estimates were pooled per treatment line (first-line or pre-treated) with random effects meta-analyses. The choice of the therapeutic benefit to target in future trials was guided by the European Society for Medical Oncology Magnitude of Clinical Benefit Scale (ESMO-MCBS).

Results: Information was acquired for 1030 LPS (25 trials; 7 first-line, 17 pre-treated, 1 both) and 348 SS patients (13 trials; 3 first-line, 10 pre-treated). For LPS, the overall pooled first-line PFSRs were 69% (95%-CI 60-77%) and 56% (95%-CI 45-67%) at 3 and 6 months, respectively. These rates were 49% (95%-CI 40-57%) / 28% (95%-CI 22-34%) for >1 lines. For SS, first-line PFSRs were 74% (95%-CI 58-86%) / 56% (95%-CI 31-78%) at 3 and 6 months and pre-treated rates were 45% (95%-CI 34-57%) / 25% (95%-CI 16-36%). Following ESMO-MCBS guidelines, the minimum values to target are 79% and 69% for first-line LPS (82% and 69% for SS) at 3 and 6 months. For pre-treated LPS, recommended PFSRs at 3 and 6 months suggesting drug activity are 63% and 44% (60% and 41% for SS).

Conclusions: New benchmarks are proposed for advanced/metastatic LPS or SS to design future histology-specific phase II trials. More data are needed to provide definitive thresholds for the different LPS subtypes.

3.1 Introduction

Soft-tissue sarcomas (STS) are very heterogeneous rare mesenchymal malignancies that account for about 1% of all adult tumours. In general, over the years more than 100 histologic subtypes have been recognised with a widely varying presentation, sensitivity to treatment and long-term outcomes [1]. The prognosis for advanced STS is poor with median overall survival (OS) now ranging from 12 to 18 months [2]. The most common site of metastasis is the lungs but other (intraabdominal, bone) locations are not uncommon [1–3]. Systemic treatment represents the mainstay for the management of the locally advanced or metastatic disease. For first-line treatment of STS, doxorubicin alone or in combination with ifosfamide has been considered the most active drug (combination) for several decades [2]. After first-line drugs, subsequent treatments depend on subtype. Among the most used ones in second- and further-line 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 [4]. The combination of olaratumab + doxorubicin appeared to show a survival benefit compared with doxorubicin alone in a randomised phase II study [5], but eribulin is the only drug to have shown a survival benefit although curiously no benefit in PFS.

Liposarcomas (LPS), one of the most common STS types (15-20% of all STS), are complex and diverse neoplasms [6]. These tumours can be separated into three biological subtypes based on specific genetic alterations: well-differentiated/dedifferentiated LPS that is the most common (∼70%), myxoid LPS (∼20%), and pleomorphic LPS (∼5%) which has the worst prognosis [7, 8]. Currently, available systemic therapies include anthracycline-based treatment for first-line typically with doxorubicin (with/without ifosfamide), and trabectedin or eribulin after anthracycline failure.

A somewhat less common STS type with varying clinical behaviour and response to treatment is synovial sarcoma (SS; 5-10% of STS) [9, 10]. Patients with SS have a relatively young age at diagnosis (mean 39 years) [10]. These tumours are either monophasic (pure sarcomas), biphasic (epithelioid and sarcomatous components combined), or poorly differentiated and have a unique biology among STS characterised by SYT-SSX1, 2 or 4 translocations [11]. In the advanced/metastatic setting, SS usually shows a higher chemosensitivity compared to other STS histotypes. SS is commonly treated with anthracyclines and/or ifosfamide in first-line, while high-dose continuous infusion ifosfamide, pazopanib, and trabectedin represent the most used agents in pre-treated patients [10, 11].

In 2002, Van Glabbeke *et al.* [12], published a pooled analysis with independent patient data calculating progression-free rates for first-line or pre-treated STS patients who had been included in phase II trials of the European Organisation for Research and Treatment of Cancer (EORTC) - Soft Tissue and Bone Sarcoma Group (STBSG) database. Efficacy thresholds were estimated in order to make a distinction between active and inactive antineoplastic agents. In first-line, a 6-month rate of 30-56% was considered as a reference value for drug activity depending on histology. For the pre-treated population, a 3-month rate $\geq 40\%$ suggested drug activity and $\leq 20\%$ inactivity for any histologic subtype. These values have been applied extensively (> 420 citations) to design new studies for all STS.

In a previous study by our group (Kantidakis *et al.*, 2021) [13], we collected summary estimates from an extensive literature review of phase II, III or IV studies published between 2003 and 2018 on advanced or metastatic STS to provide an update for leiomyosarcomas (LMS) – the most frequently appearing histologic type in these publications. The primary endpoint was defined as progression-free survival rate at 3 or 6 months (PFSR; counting any death as an event) which is nowadays a preferred and more popular endpoint than progression-free rate (censoring deaths not related to disease). Drugs were classified as recommended or not based on the European Society for Medical Oncology (ESMO) 2018 guidelines [14]. Since the differences between recommended and non-recommended agents were not significant, the overall pooled PFSR was used as a reference. The ESMO Magnitude of Clinical Benefit Scale (ESMO-MCBS) [15] pinpointed the treatment effects to target for a clinically relevant benefit in future phase II trials. For first-line LMS, a PFSR at 6 months ≥70%, and for pre-treated population, a 3-month PFSR ≥62% or at 6-month PFSR ≥44% would suggest drug activity.

Historically, the majority of STS trials have been designed with a one-size-fits-all principle mixing several histologic types. However, our recent study is in accordance with a trend towards histology-specific tailored research [1, 4]. Importantly, the 2002 efficacy thresholds should be updated and recalibrated for prevalent advanced/metastatic STS types to reflect modern clinical practice, as future agents should perform better than currently available standards of care. Here, the aim is to extend our 2021 study for advanced/metastatic LPS or SS, the second and third most common types in our literature review (2003-2018), which differ from real-life incidence [16], to provide benchmarks to design new phase II studies with PFSR as the primary endpoint.

3.2 Methods

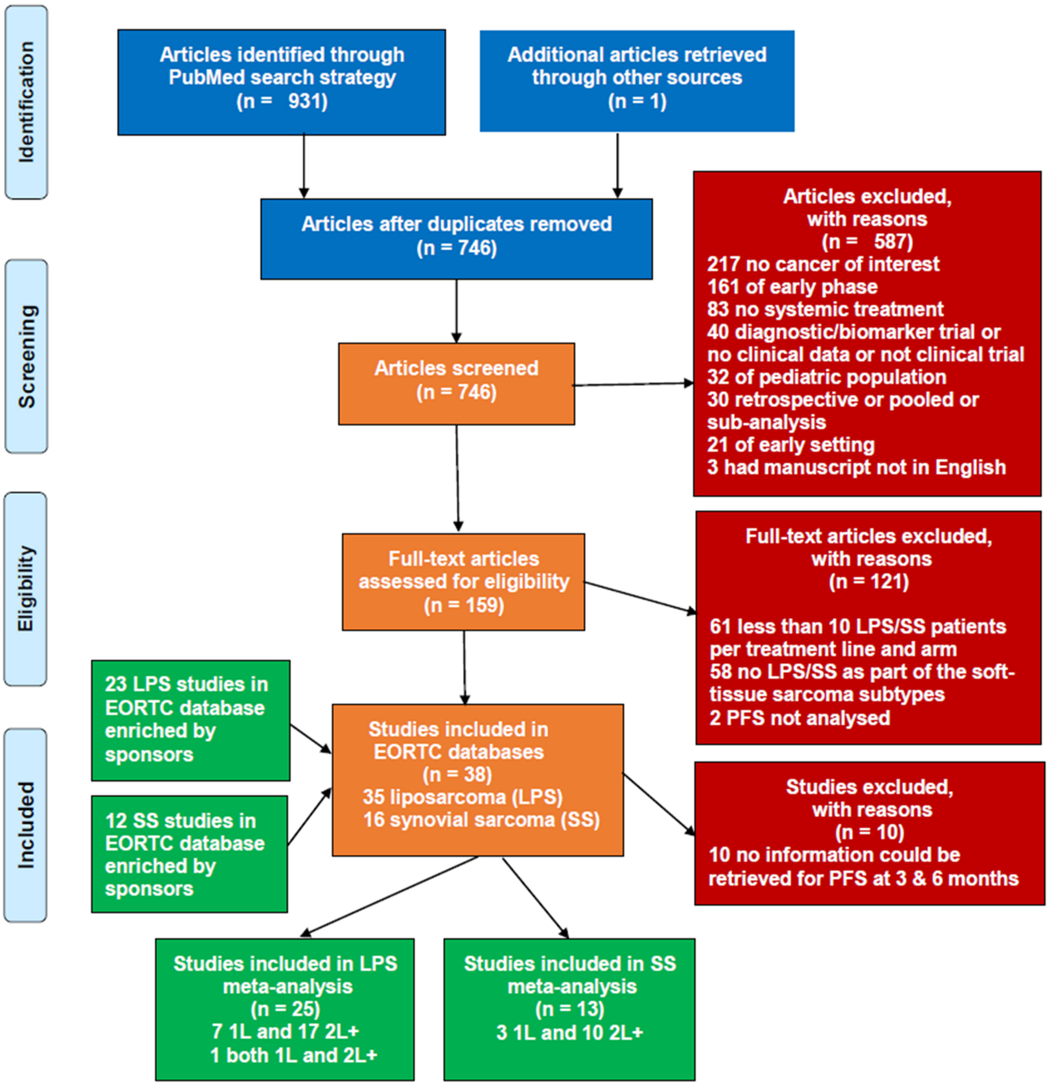

Identification
Screening
Eligibility
Included
Articles identified through PubMed search strategy (n = 931)
Additional articles retrieved through other sources (n = 1)
Articles after duplicates removed (n = 746)
Articles excluded, with reasons (n = 587)
217 no cancer of interest
161 of early phase
83 no systemic treatment
40 diagnostic/biomarker trial or no clinical data or not clinical trial
32 of pediatric population
30 retrospective or pooled or sub-analysis
21 of early setting
3 had manuscript not in English
Articles screened (n = 746)
Full-text articles assessed for eligibility (n = 159)
Full-text articles excluded, with reasons (n = 121)
61 less than 10 LPS/SS patients per treatment line and arm
58 no LPS/SS as part of the soft-tissue sarcoma subtypes
2 PFS not analysed
23 LPS studies in EORTC database enriched by sponsors
12 SS studies in EORTC database enriched by sponsors
Studies included in EORTC databases (n = 38)
35 liposarcoma (LPS)
16 synovial sarcoma (SS)
Studies excluded, with reasons (n = 10)
10 no information could be retrieved for PFS at 3 & 6 months
Studies included in LPS meta-analysis (n = 25)
7 1L and 17 2L+
1 both 1L and 2L+
Studies included in SS meta-analysis (n = 13)
3 1L and 10 2L+

Figure 3.1: **Study selection**. There were 38 potentially relevant studies for LPS or SS patients in the EORTC databases; 35 for LPS and 16 for SS. A total of 13 of these trials included LPS and SS patients, whereas 3 trials included SS patients only. Collecting extra information (PFS estimates at 3-6 months) was of paramount importance because of the very limited data availability before the enrichment of the databases by the sponsors (PFS estimates could only be recovered by the publications for two studies with LPS, and one study with SS patients).

3.2.1 Search strategy and selection criteria for the literature review

The literature review and meta-analyses were conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [17]. An electronic search was performed in PubMed for phase II, III, or IV clinical trials with advanced/metastatic or non-operable STS patients. Studies were published in English between January 1, 2003 and December 31, 2018. Eligible study designs included non-randomised trials, randomised controlled trials, and prospective real-life studies. Study domain included first, second, or later line systemic therapy. Papers with retrospective clinical data, case-control studies, early phase trials, pooled analyses, and reports were excluded as well as those devoted to bone sarcomas, GIST, or paediatric population.

A two-step procedure was performed by three authors (G.K., A.N., and M.V.) to construct the database. More information about the trial selection can be found in Ref. [13] and in Appendix pp 3-4.

3.2.2 Extracting information for the meta-analyses

For this work, the focus was on the second and third most prevalent STS in the database; LPS and SS. Two meta-analysis databases were designed with a row per treatment arm and treatment line (first-line versus pre-treated population). For each of them, G.K. extracted the number of evaluable LPS or SS patients for PFSR (those included in the efficacy dataset based on the statistical plan's criteria), the PFSR at 3 and 6 months together with the 95% confidence intervals (95%-CIs), and the year of study activation. Placebo / best supportive care arms, arms with <10 patients, mixed treatment lines, or studies activated before 2000 were removed from the database. When summary PFS estimates (at 3 and 6 months) could not be retrieved from a publication, they were requested from first authors and/or study sponsors.

3.2.3 Statistical methods

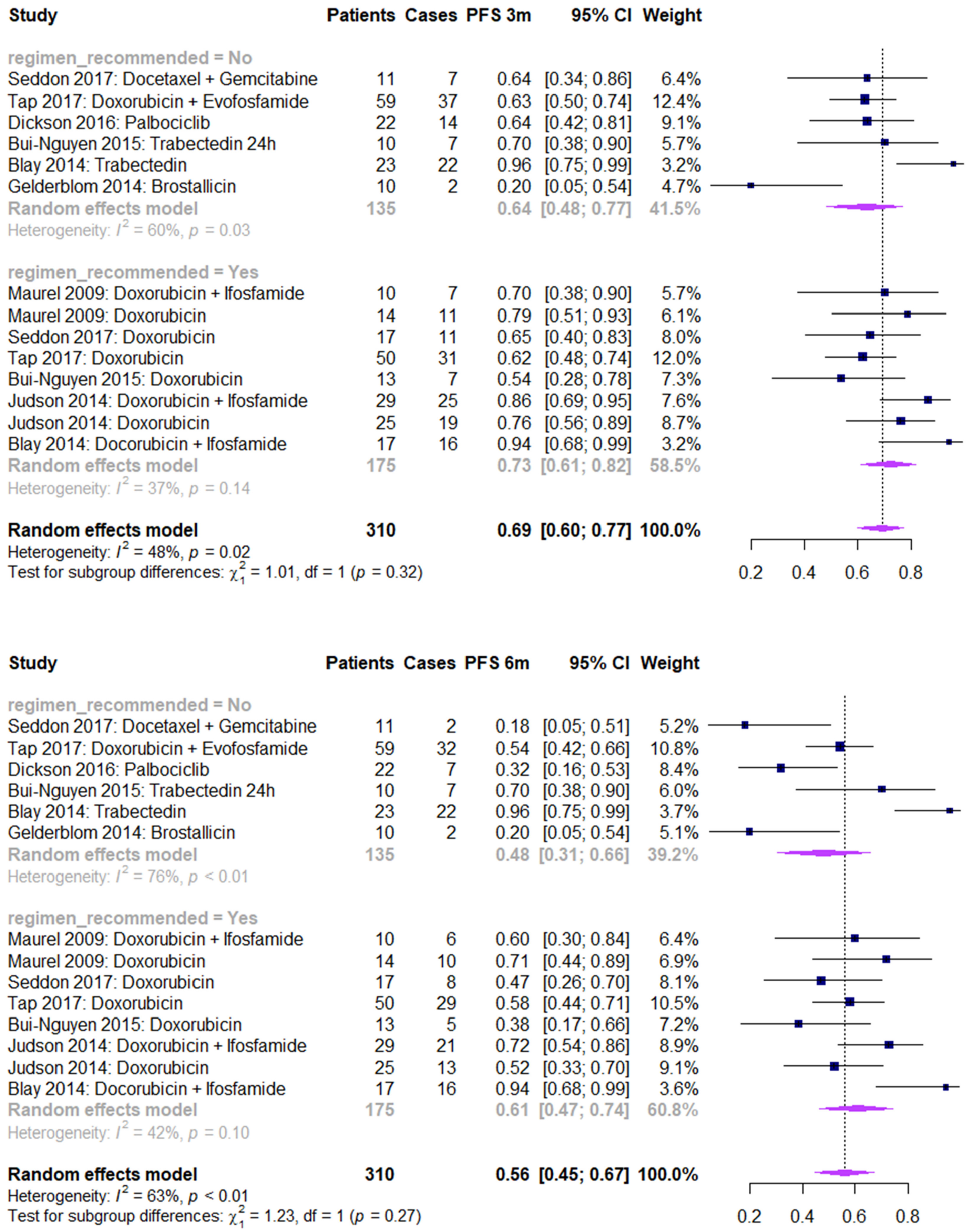
In both databases (for LPS and SS), a random-effects model was employed to estimate the overall PFSR at 3 and 6 months per line of treatment (first-line versus pre-treated). The DerSimonian and Laird method was used to estimate the between-study variance in clinical trials [18, 19]. The inverse variance method was used to pool treatment-specific PFS estimates (more weight is given to larger studies). For each treatment arm, the number of cases (patients alive and progression-free) was approximated based on the total number of evaluable patients and the recorded PFS estimate; the equivalent PFS proportion is defined as cases/evaluable patients. The calculated number of cases was employed under a binomial distribution to estimate the variance (unknown quantity) for each drug or combination and the 95%-CIs [20, 21]. The treatment-specific PFS estimates are presented on forest plots. 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) [22].

The ESMO 2021 guidelines [23] were used to classify each drug (or drug combination) as recommended treatment (R-T) or non-recommended treatment (NR-T) per treatment line and histologic subtype. The difference in PFS between the two groups of drugs (R-T versus NR-T) was formally compared using meta-regression (subgroup meta-analysis) with a chi-square statistic. The effect of other predictors on PFS (phase of the trial, study design, year of activation, sample size) was also tested in univariate models to address if they can explain part of the residual heterogeneity. Funnel plots and formal regression tests were used to assess the risk of publication bias [24–26]. Potentially influential studies and studies contributing to heterogeneity were detected with Baujat plots [27]. The choice of the therapeutic benefit to target in future trials was guided by the ESMO–MCBS [15]. Analyses were performed using packages *metafor* and *meta* in R version 4.1.2 [28, 29]. Reported p-values are two-sided. Further methodological details can be found in Appendix pp 12-14.

3.3 Results

3.3.1 Clinical trials included

The study selection is provided in Fig. 3.1. In total, 38 studies were potentially relevant for the meta-analyses (35 for LPS, 16 for SS): 25 trials were included in the LPS meta-analyses [30–54] and 13 trials [30, 32, 38, 41, 42, 46–48, 52, 53, 55–57] in the SS meta-analyses.



Study	Patients	Cases	PFS 3m	95% CI	Weight
regimen_recommended = No					
Seddon 2017: Docetaxel + Gemcitabine	11	7	0.64	[0.34; 0.86]	6.4%
Tap 2017: Doxorubicin + Evofosfamide	59	37	0.63	[0.50; 0.74]	12.4%
Dickson 2016: Palbociclib	22	14	0.64	[0.42; 0.81]	9.1%
Bui-Nguyen 2015: Trabectedin 24h	10	7	0.70	[0.38; 0.90]	5.7%
Blay 2014: Trabectedin	23	22	0.96	[0.75; 0.99]	3.2%
Gelderblom 2014: Brostallicin	10	2	0.20	[0.05; 0.54]	4.7%
Random effects model	135		0.64	[0.48; 0.77]	41.5%
Heterogeneity: $I^2 = 60\%$, $p = 0.03$					
regimen_recommended = Yes					
Maurel 2009: Doxorubicin + Ifosfamide	10	7	0.70	[0.38; 0.90]	5.7%
Maurel 2009: Doxorubicin	14	11	0.79	[0.51; 0.93]	6.1%
Seddon 2017: Doxorubicin	17	11	0.65	[0.40; 0.83]	8.0%
Tap 2017: Doxorubicin	50	31	0.62	[0.48; 0.74]	12.0%
Bui-Nguyen 2015: Doxorubicin	13	7	0.54	[0.28; 0.78]	7.3%
Judson 2014: Doxorubicin + Ifosfamide	29	25	0.86	[0.69; 0.95]	7.6%
Judson 2014: Doxorubicin	25	19	0.76	[0.56; 0.89]	8.7%
Blay 2014: Docorubicin + Ifosfamide	17	16	0.94	[0.68; 0.99]	3.2%
Random effects model	175		0.73	[0.61; 0.82]	58.5%
Heterogeneity: $I^2 = 37\%$, $p = 0.14$					
Random effects model	**310**		**0.69**	**[0.60; 0.77]**	**100.0%**
Heterogeneity: $I^2 = 48\%$, $p = 0.02$					
Test for subgroup differences: $\chi^2_1 = 1.01$, df = 1 ($p = 0.32$)					

Study	Patients	Cases	PFS 6m	95% CI	Weight
regimen_recommended = No					
Seddon 2017: Docetaxel + Gemcitabine	11	2	0.18	[0.05; 0.51]	5.2%
Tap 2017: Doxorubicin + Evofosfamide	59	32	0.54	[0.42; 0.66]	10.8%
Dickson 2016: Palbociclib	22	7	0.32	[0.16; 0.53]	8.4%
Bui-Nguyen 2015: Trabectedin 24h	10	7	0.70	[0.38; 0.90]	6.0%
Blay 2014: Trabectedin	23	22	0.96	[0.75; 0.99]	3.7%
Gelderblom 2014: Brostallicin	10	2	0.20	[0.05; 0.54]	5.1%
Random effects model	135		0.48	[0.31; 0.66]	39.2%
Heterogeneity: $I^2 = 76\%$, $p < 0.01$					
regimen_recommended = Yes					
Maurel 2009: Doxorubicin + Ifosfamide	10	6	0.60	[0.30; 0.84]	6.4%
Maurel 2009: Doxorubicin	14	10	0.71	[0.44; 0.89]	6.9%
Seddon 2017: Doxorubicin	17	8	0.47	[0.26; 0.70]	8.1%
Tap 2017: Doxorubicin	50	29	0.58	[0.44; 0.71]	10.5%
Bui-Nguyen 2015: Doxorubicin	13	5	0.38	[0.17; 0.66]	7.2%
Judson 2014: Doxorubicin + Ifosfamide	29	21	0.72	[0.54; 0.86]	8.9%
Judson 2014: Doxorubicin	25	13	0.52	[0.33; 0.70]	9.1%
Blay 2014: Docorubicin + Ifosfamide	17	16	0.94	[0.68; 0.99]	3.6%
Random effects model	175		0.61	[0.47; 0.74]	60.8%
Heterogeneity: $I^2 = 42\%$, $p = 0.10$					
Random effects model	**310**		**0.56**	**[0.45; 0.67]**	**100.0%**
Heterogeneity: $I^2 = 63\%$, $p < 0.01$					
Test for subgroup differences: $\chi^2_1 = 1.23$, df = 1 ($p = 0.27$)					

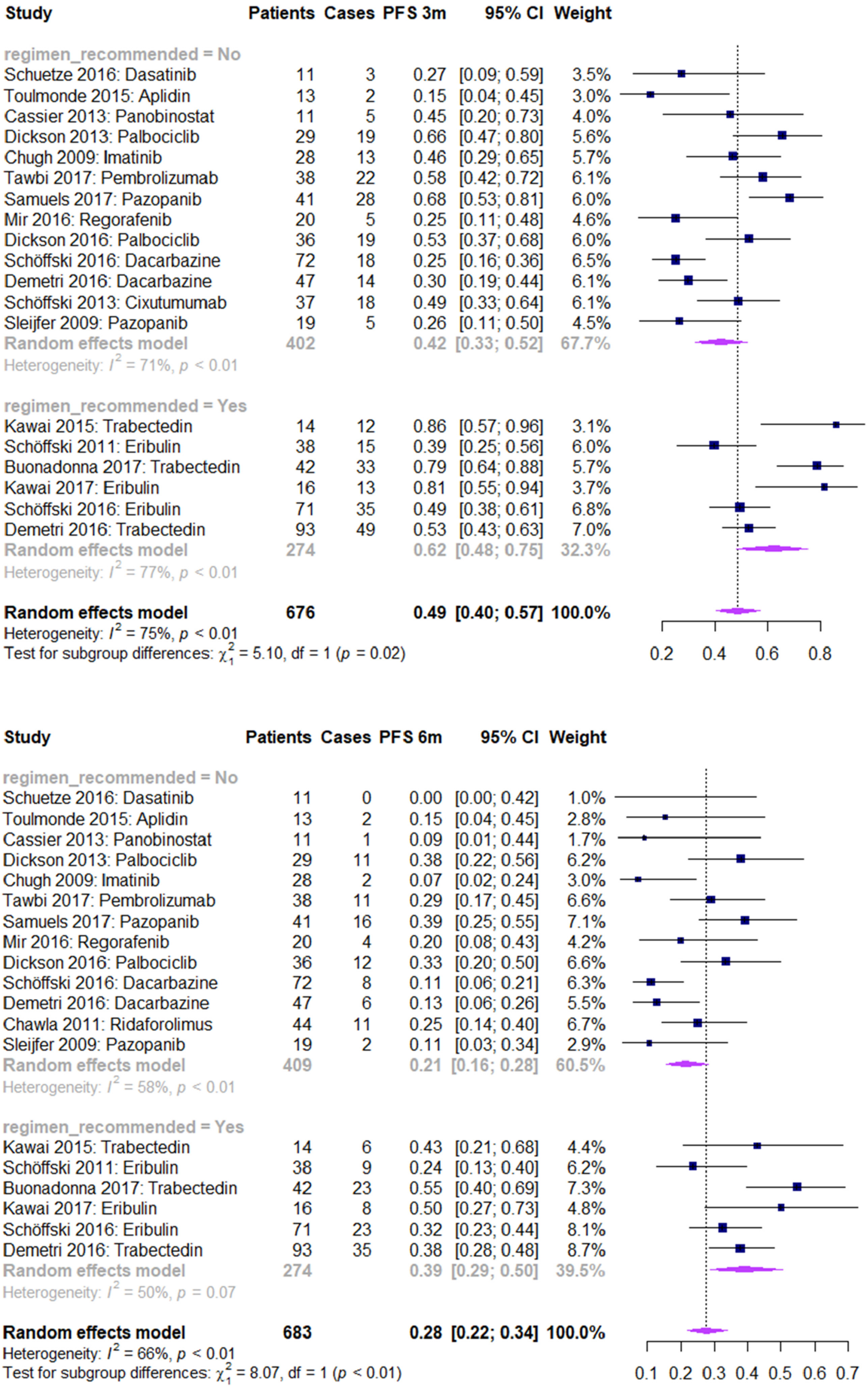
Figure 3.2: Forest plots of PFS at 3 (upper panel) and 6 (low panel) months for first-line LPS patients. PFS proportion at 3 or 6 months was defined as the approximate proportion of patients alive and without disease progression (: number of cases) at 3 or 6 months. Treatments were classified as recommended or non-recommended according to the ESMO 2021 guidelines [23]. Heterogeneity refers to variability in outcomes (PFS proportions) between the studies that cannot be attributed to random variation. A PFSR is the proportion*100.

3.3.2 Characteristics of trials

First author (year of publication)	Study period	Study type	Phase	Total patients registered	Histologic type	Drug or drug combination	Treatment line	Recommended	Evaluable patients analysed for PFS
Ray-Coquard et al. (2008)	2002 - 2005	Non-randomised	2	48	SS	Gefitinib	2+	No	46 (95.83%)
Chugh et al. (2009)	2001 - 2005	Non-randomised	2	185	LPS	Imatinib	2+	No	28 (15.14%)
					SS	Imatinib	2+	No	21 (11.35%)
Maurel et al. (2009)	2003 - 2007	Randomised	2	132	LPS	Doxorubicin	1	Yes	14 (10.61%)
					LPS	Doxorubicin + ifosfamide	1	Yes	10 (7.58%)
Sleijfer et al. (2009)	2005 - 2007	Non-randomised	2	142	LPS	Pazopanib	2+	No	19 (13.38%)
					SS	Pazopanib	2+	Yes	37 (26.06%)
Schöffski et al. (2011)	2007 - 2009	Non-randomised	2	128	LPS	Eribulin	2+	Yes	38 (29.69%)
					SS	Eribulin	2+	No	27 (21.09%)
Chawla et al. (2012)	2004 - 2005	Non-randomised	2	216	LPS	Ridaforolimus	2+	No	44 (20.37%)
van der Graaf et al. (2012)	2008 - 2010	Randomised	3	372	SS	Pazopanib	2+	Yes	29 (7.80%)
Cassier et al. (2013)	2010	Non-randomised	2	47	LPS	Panobinostat	2+	No	11 (23.40%)
Dickson et al. (2013)	2010 - 2011	Non-randomised	2	30	LPS	Palbociclib	2+	No	29 (96.67%)
Schöffski et al. (2013)	2008 - 2012	Non-randomised	2	113	LPS	Cixutumumab	2+	No	37 (32.74%)
					SS	Cixutumumab	2+	No	17 (15.04%)
Blay et al. (2014)	2008 - 2012	Randomised	3	121	LPS	Doxorubicin + ifosfamide	1	Yes	17 (14.05%)
					LPS	Trabectedin	1	No	23 (19.01%)
					SS	Trabectedin	1	No	15 (12.40%)
Gelderblom et al. (2014)	2006 - 2008	Randomised	2	118	LPS	Brostallicin	1	No	10 (8.47%)
Judson et al. (2014)	2003 - 2010	Randomised	3	455	LPS	Doxorubicin	1	Yes	25 (5.49%)
					LPS	Doxorubicin + ifosfamide	1	Yes	29 (6.37%)
					SS	Doxorubicin	1	Yes	37 (8.13%)
					SS	Doxorubicin + ifosfamide	1	Yes	26 (5.71%)
Bui-Nguyen et al. (2015)	2011 - 2012	Randomised	2|3	133	LPS	Doxorubicin	1	Yes	13 (9.77%)
					LPS	Trabectedin 24h	1	No	10 (7.52%)
Kawai et al. (2015)	2012 - 2014	Randomised	2	76	LPS	Trabectedin	2+	Yes	14 (18.42%)
Robbins et al. (2015)	2008 - 2012	Non-randomised	2	38	SS	Cyclophosphamide + fludarabine + TCR transduced cells	2+	No	19 (50.00%)
Toulmonde et al.(2015)	2012 - 2013	Non-randomised	2	24	LPS	Aplidin	2+	No	13 (54.17%)
Demetri et al. (2016)	2011 - 2013	Randomised	3	518	LPS	Dacarbazine	2+	No	47 (9.07%)
					LPS	Trabectedin	2+	Yes	93 (17.95%)
Dickson et al. (2016)	2011 - 2014	Non-randomised	2	60	LPS	Palbociclib	1	No	22 (36.67%)
					LPS	Palbociclib	2+	No	36 (60.00%)
Mir et al. (2016)	2013 - 2014	Randomised	2	182	LPS	Regorafenib	2+	No	20 (10.99%)
					SS	Regorafenib	2+	No	13 (7.14%)
Schöffski et al. (2016)	2011 - 2013	Randomised	3	452	LPS	Dacarbazine	2+	No	72 (15.93%)
					LPS	Eribulin	2+	Yes	71 (15.71%)
Schuetze et al. (2016)	2007 - 2009	Non-randomised	2	196	LPS	Dasatinib	2+	No	11 (5.61%)
Buonadonna et al. (2017)	2012 - 2014	Non-randomised	4	218	LPS	Trabectedin	2+	Yes	42 (19.27%)
					SS	Trabectedin	2+	Yes	23 (10.55%)
Kawai et al. (2017)	2011 - 2014	Non-randomised	2	52	LPS	Eribulin	2+	Yes	16 (30.77%)
Samuels et al. (2017)	2012 - 2015	Non-randomised	2	41	LPS	Pazopanib	2+	No	41 (100.00%)
Seddon et al. (2017)	2010 - 2014	Randomised	3	257	LPS	Doxorubicin	1	Yes	17 (6.61%)
					LPS	Docetaxel + gemcitabine	1	No	11 (4.28%)
Tap et al. (2017)	2011 - 2014	Randomised	3	640	LPS	Doxorubicin	1	Yes	50 (7.81%)
					LPS	Doxorubicin + evofosfamide	1	No	59 (9.22%)
					SS	Doxorubicin	1	Yes	11 (1.72%)
					SS	Doxorubicin + evofosfamide	1	No	17 (2.66%)
Tawbi et al. (2017)	2015 - 2016	Non-randomised	2	86	LPS	Pembrolizumab	2+	No	38 (44.19%)
					SS	Pembrolizumab	2+	No	10 (11.63%)

Table 3.1: **Main characteristics of all studies included in the LPS or SS meta-analyses**. Studies in the SS database are presented in shade. Treatments were classified as recommended (yes or no) according to ESMO 2021 guidelines [23]. Study period = period of first to last patient accrual. Evaluable patients were those who satisfied the study's statistical plan criteria for inclusion in efficacy data sets. Trabectedin 24h = trabectedin 24-h infusion treatment schedule. The 3-h infusion treatment arm was excluded from the LPS meta-analysis due to limited number of patients (n = 6 <10). The Gelderblom study (2014) contained two treatment arms: doxorubicin and brostallicin. The doxorubicin arm was excluded from the LPS meta-analysis because it did not reach the predetermined number of patients (n = 9 <10). In the Blay study (2014), the control arm: doxorubicin + ifosfamide was removed from SS meta-analysis as it did not reach the required sample size (n = 9 <10). Placebo / best supportive care arms were also not included (van der Graaf et al. (2012), Kawai et al. (2015), Mir et al. (2016)).

A total of 1030 patients were evaluable for the LPS meta-analysis (range 10 to 93 patients per trial, table 3.1) and 348 for the SS meta-analysis (range 10 to 46, table 3.1). In first-line, the most common regimens were doxorubicin alone or in combination with ifosfamide (eight times) for LPS and doxorubicin monotherapy or in combination with evofosfamide or ifosfamide (four times) for SS. In pre-treated population, eribulin and trabectedin were the most common drugs for LPS (three times) and pazopanib for SS (two times).



Study	Patients	Cases	PFS 3m	95% CI	Weight
regimen_recommended = No					
Schuetze 2016: Dasatinib	11	3	0.27	[0.09; 0.59]	3.5%
Toulmonde 2015: Aplidin	13	2	0.15	[0.04; 0.45]	3.0%
Cassier 2013: Panobinostat	11	5	0.45	[0.20; 0.73]	4.0%
Dickson 2013: Palbociclib	29	19	0.66	[0.47; 0.80]	5.6%
Chugh 2009: Imatinib	28	13	0.46	[0.29; 0.65]	5.7%
Tawbi 2017: Pembrolizumab	38	22	0.58	[0.42; 0.72]	6.1%
Samuels 2017: Pazopanib	41	28	0.68	[0.53; 0.81]	6.0%
Mir 2016: Regorafenib	20	5	0.25	[0.11; 0.48]	4.6%
Dickson 2016: Palbociclib	36	19	0.53	[0.37; 0.68]	6.0%
Schöffski 2016: Dacarbazine	72	18	0.25	[0.16; 0.36]	6.5%
Demetri 2016: Dacarbazine	47	14	0.30	[0.19; 0.44]	6.1%
Schöffski 2013: Cixutumumab	37	18	0.49	[0.33; 0.64]	6.1%
Sleijfer 2009: Pazopanib	19	5	0.26	[0.11; 0.50]	4.5%
Random effects model	402		0.42	[0.33; 0.52]	67.7%
Heterogeneity: $I^2 = 71\%$, $p < 0.01$					
regimen_recommended = Yes					
Kawai 2015: Trabectedin	14	12	0.86	[0.57; 0.96]	3.1%
Schöffski 2011: Eribulin	38	15	0.39	[0.25; 0.56]	6.0%
Buonadonna 2017: Trabectedin	42	33	0.79	[0.64; 0.88]	5.7%
Kawai 2017: Eribulin	16	13	0.81	[0.55; 0.94]	3.7%
Schöffski 2016: Eribulin	71	35	0.49	[0.38; 0.61]	6.8%
Demetri 2016: Trabectedin	93	49	0.53	[0.43; 0.63]	7.0%
Random effects model	274		0.62	[0.48; 0.75]	32.3%
Heterogeneity: $I^2 = 77\%$, $p < 0.01$					
Random effects model	**676**		**0.49**	**[0.40; 0.57]**	**100.0%**
Heterogeneity: $I^2 = 75\%$, $p < 0.01$					
Test for subgroup differences: $\chi^2_1 = 5.10$, df = 1 ($p = 0.02$)					

Study	Patients	Cases	PFS 6m	95% CI	Weight
regimen_recommended = No					
Schuetze 2016: Dasatinib	11	0	0.00	[0.00; 0.42]	1.0%
Toulmonde 2015: Aplidin	13	2	0.15	[0.04; 0.45]	2.8%
Cassier 2013: Panobinostat	11	1	0.09	[0.01; 0.44]	1.7%
Dickson 2013: Palbociclib	29	11	0.38	[0.22; 0.56]	6.2%
Chugh 2009: Imatinib	28	2	0.07	[0.02; 0.24]	3.0%
Tawbi 2017: Pembrolizumab	38	11	0.29	[0.17; 0.45]	6.6%
Samuels 2017: Pazopanib	41	16	0.39	[0.25; 0.55]	7.1%
Mir 2016: Regorafenib	20	4	0.20	[0.08; 0.43]	4.2%
Dickson 2016: Palbociclib	36	12	0.33	[0.20; 0.50]	6.6%
Schöffski 2016: Dacarbazine	72	8	0.11	[0.06; 0.21]	6.3%
Demetri 2016: Dacarbazine	47	6	0.13	[0.06; 0.26]	5.5%
Chawla 2011: Ridaforolimus	44	11	0.25	[0.14; 0.40]	6.7%
Sleijfer 2009: Pazopanib	19	2	0.11	[0.03; 0.34]	2.9%
Random effects model	409		0.21	[0.16; 0.28]	60.5%
Heterogeneity: $I^2 = 58\%$, $p < 0.01$					
regimen_recommended = Yes					
Kawai 2015: Trabectedin	14	6	0.43	[0.21; 0.68]	4.4%
Schöffski 2011: Eribulin	38	9	0.24	[0.13; 0.40]	6.2%
Buonadonna 2017: Trabectedin	42	23	0.55	[0.40; 0.69]	7.3%
Kawai 2017: Eribulin	16	8	0.50	[0.27; 0.73]	4.8%
Schöffski 2016: Eribulin	71	23	0.32	[0.23; 0.44]	8.1%
Demetri 2016: Trabectedin	93	35	0.38	[0.28; 0.48]	8.7%
Random effects model	274		0.39	[0.29; 0.50]	39.5%
Heterogeneity: $I^2 = 50\%$, $p = 0.07$					
Random effects model	**683**		**0.28**	**[0.22; 0.34]**	**100.0%**
Heterogeneity: $I^2 = 66\%$, $p < 0.01$					
Test for subgroup differences: $\chi^2_1 = 8.07$, df = 1 ($p < 0.01$)					

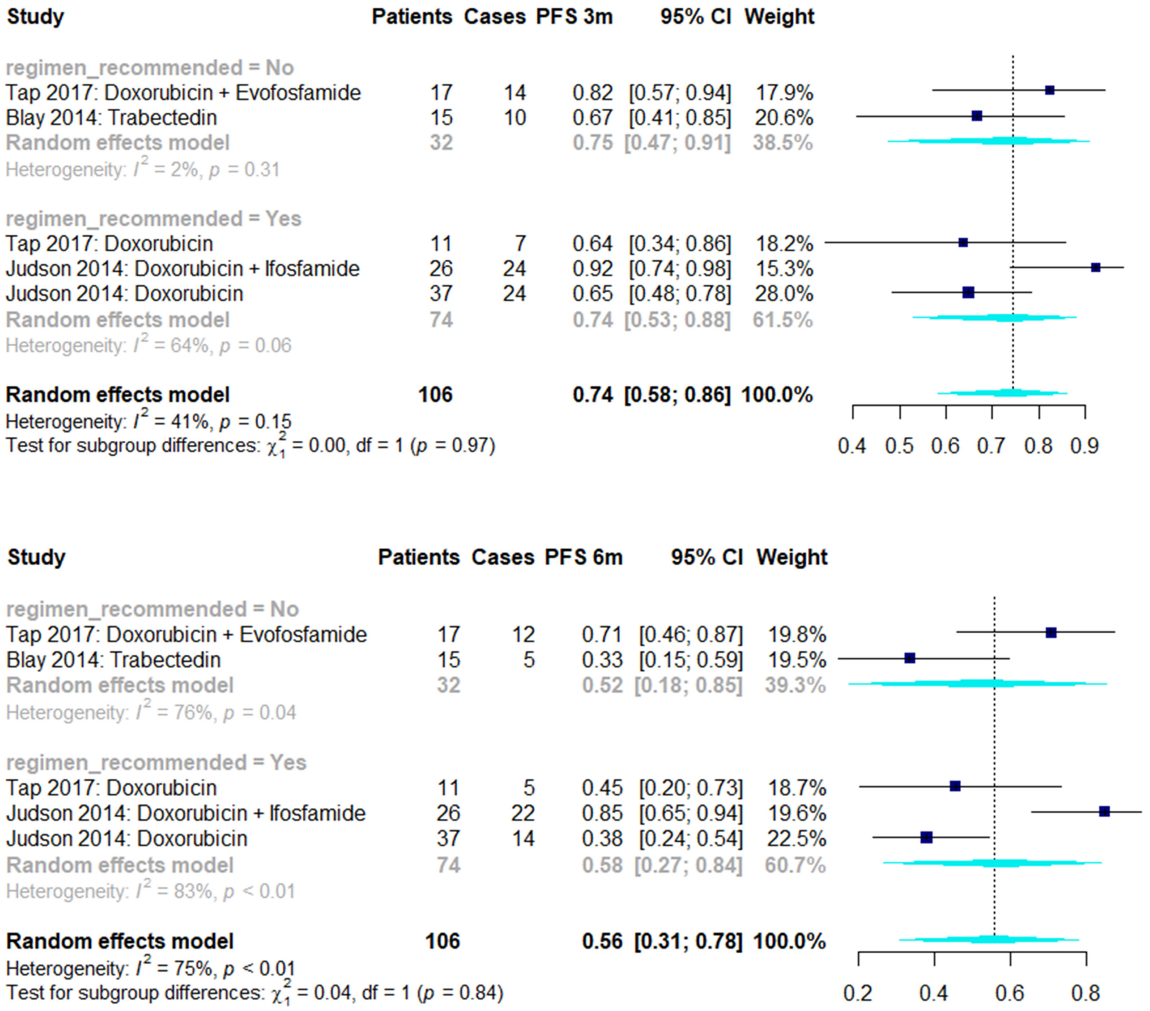
Figure 3.3: Forest plots of PFS at 3 (upper panel) and 6 (low panel) months for pre-treated LPS patients.

3.3.3 Risk of publication bias

The contour enhanced funnel plots did not indicate systematic asymmetry between the studies included for LPS or SS meta-analyses with the exception of pre-treated LPS population at 6 months. Tests for funnel plot asymmetry indicated low risk of publication bias in the databases for SS and first-line LPS, as well as high risk of bias for pre-treated LPS at 6 months (see Appendix sections 2.3 and 2.4). However, publication bias cannot be excluded for first-line SS patients because of the very limited number of studies (three trials, five treatment regimens).

3.3.4 LPS meta-analyses

Forest plots for first-line and pre-treated patients are illustrated in Fig. 3.2 and 3.3. For first-line, the pooled PFSRs at 3 months were 73% (95%-CI 61-82%) and 64% (95%-CI 48-77%) for R-T / NR-T, respectively. At 6 months PFSRs were 61% (95%-CI 47-74%) and 48% (95%-CI 31-66%). There was no statistically significant difference between the two groups of drugs (p-values 0.32 and 0.27). Overall heterogeneity was high (I^2 = 48% with p = 0.02 at 3 months, I^2 = 63% with p < 0.01 at 6 months). Regarding the pre-treated population, the pooled PFSRs were 62% (95%-CI 48-75%) or 39% (95%-CI 29-50%) at 3 and 6 months for R-T, and 42% (95%-CI 33-52%) or 21% (95%-CI 16-28%) for NR-T, respectively. Differences between R-T/NR-T were found to be significant at both 3 and 6 months (p-value < 0.05). Overall heterogeneity was very high at 3 and 6 months (I^2 > 65%, p < 0.01). Univariate meta-regressions did not identify any prognostic factors which can explain part of the heterogeneity at 3 and 6 months (amongst phase, study design, year of activation, sample size; see Appendix).



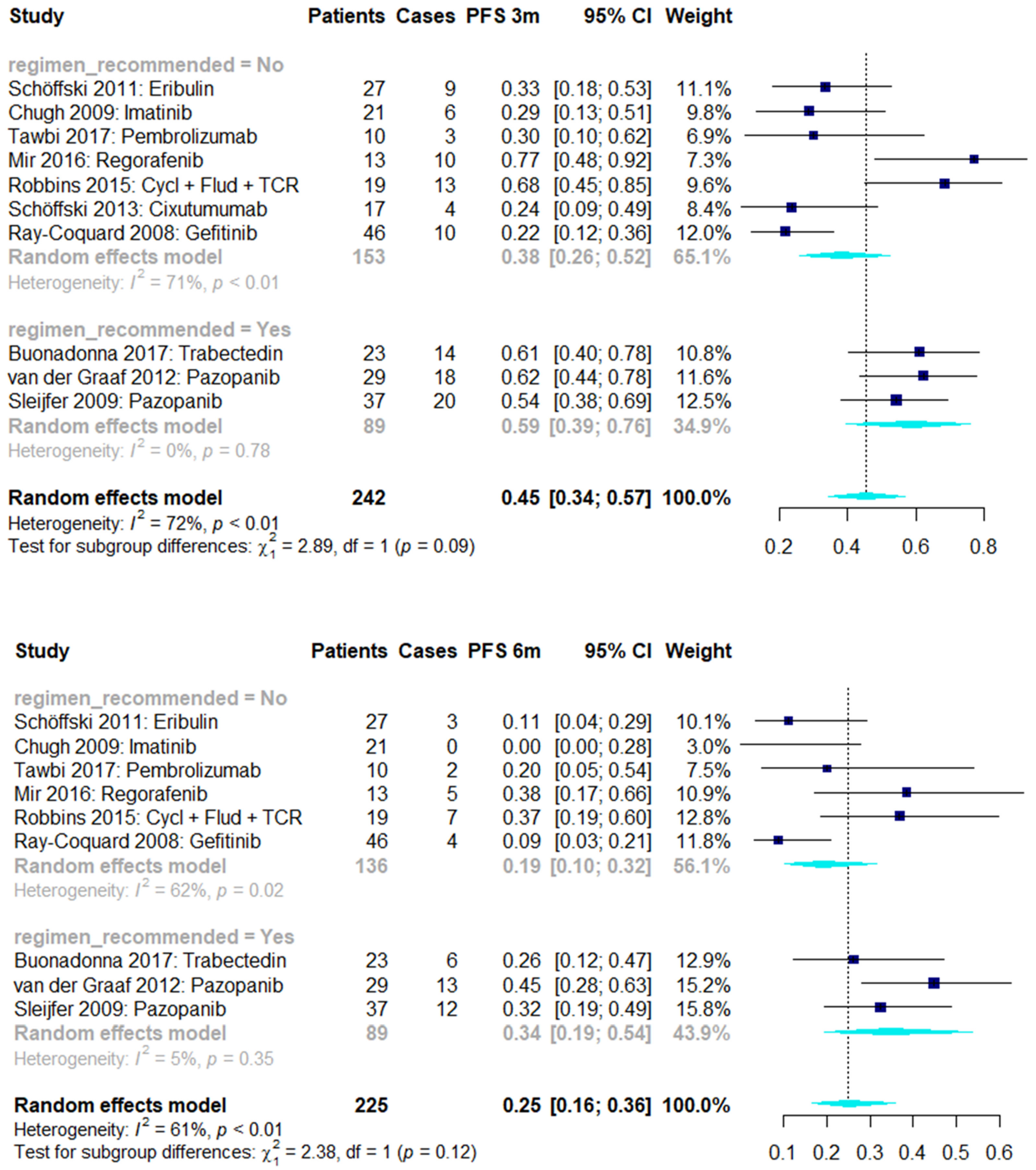
Study	Patients	Cases	PFS 3m	95% CI	Weight
regimen_recommended = No					
Tap 2017: Doxorubicin + Evofosfamide	17	14	0.82	[0.57; 0.94]	17.9%
Blay 2014: Trabectedin	15	10	0.67	[0.41; 0.85]	20.6%
Random effects model	32		0.75	[0.47; 0.91]	38.5%
Heterogeneity: I^2 = 2%, p = 0.31					
regimen_recommended = Yes					
Tap 2017: Doxorubicin	11	7	0.64	[0.34; 0.86]	18.2%
Judson 2014: Doxorubicin + Ifosfamide	26	24	0.92	[0.74; 0.98]	15.3%
Judson 2014: Doxorubicin	37	24	0.65	[0.48; 0.78]	28.0%
Random effects model	74		0.74	[0.53; 0.88]	61.5%
Heterogeneity: I^2 = 64%, p = 0.06					
Random effects model	**106**		**0.74**	**[0.58; 0.86]**	**100.0%**
Heterogeneity: I^2 = 41%, p = 0.15					
Test for subgroup differences: χ^2_1 = 0.00, df = 1 (p = 0.97)					

Study	Patients	Cases	PFS 6m	95% CI	Weight
regimen_recommended = No					
Tap 2017: Doxorubicin + Evofosfamide	17	12	0.71	[0.46; 0.87]	19.8%
Blay 2014: Trabectedin	15	5	0.33	[0.15; 0.59]	19.5%
Random effects model	32		0.52	[0.18; 0.85]	39.3%
Heterogeneity: I^2 = 76%, p = 0.04					
regimen_recommended = Yes					
Tap 2017: Doxorubicin	11	5	0.45	[0.20; 0.73]	18.7%
Judson 2014: Doxorubicin + Ifosfamide	26	22	0.85	[0.65; 0.94]	19.6%
Judson 2014: Doxorubicin	37	14	0.38	[0.24; 0.54]	22.5%
Random effects model	74		0.58	[0.27; 0.84]	60.7%
Heterogeneity: I^2 = 83%, p < 0.01					
Random effects model	**106**		**0.56**	**[0.31; 0.78]**	**100.0%**
Heterogeneity: I^2 = 75%, p < 0.01					
Test for subgroup differences: χ^2_1 = 0.04, df = 1 (p = 0.84)					

Figure 3.4: Forest plots of PFS at 3 (upper panel) and 6 (low panel) months for first-line SS patients.

3.3.5 SS meta-analyses

In the first-line setting, the pooled 3-month PFSR for R-T / NR-T, respectively was 74% (95%-CI 53-88%) and 75% (95%-CI 47-91%) and the pooled 6-month PFSR 58% (95%-CI 27-84%) and 52% (95%-CI 18-85%) (see Fig. 3.4). Differences between recommended or non-recommended drugs were non-significant (p-values 0.97 and 0.84). Overall variation was moderate at 3 months ($I^2 = 41\%$, $p = 0.15$) and high at 6 months ($I^2 = 75\%$, $p < 0.01$). The pooled PFSR for pre-treated patients reduced to 59% (95%-CI 39-76%) and 38% (95%-CI 26-52%) for R-T / NR-T at 3 months and 34% (95%-CI 19-54%) / 19% (95%-CI 10-32%) at 6 months (Fig. 3.5). Nevertheless, differences between the classified drugs were not significant (3- and 6-month p-values 0.09 and 0.12). Overall heterogeneity at both time points was estimated to be high ($I^2 > 60\%$, $p < 0.01$). Meta-regressions were not performed for first-line due to very limited number of studies. For the pre-treated patients, meta-regressions did not identify any prognostic covariate at 3 and 6 months.



Study	Patients	Cases	PFS 3m	95% CI	Weight
regimen_recommended = No					
Schöffski 2011: Eribulin	27	9	0.33	[0.18; 0.53]	11.1%
Chugh 2009: Imatinib	21	6	0.29	[0.13; 0.51]	9.8%
Tawbi 2017: Pembrolizumab	10	3	0.30	[0.10; 0.62]	6.9%
Mir 2016: Regorafenib	13	10	0.77	[0.48; 0.92]	7.3%
Robbins 2015: Cycl + Flud + TCR	19	13	0.68	[0.45; 0.85]	9.6%
Schöffski 2013: Cixutumumab	17	4	0.24	[0.09; 0.49]	8.4%
Ray-Coquard 2008: Gefitinib	46	10	0.22	[0.12; 0.36]	12.0%
Random effects model	153		0.38	[0.26; 0.52]	65.1%
Heterogeneity: $I^2 = 71\%$, $p < 0.01$					
regimen_recommended = Yes					
Buonadonna 2017: Trabectedin	23	14	0.61	[0.40; 0.78]	10.8%
van der Graaf 2012: Pazopanib	29	18	0.62	[0.44; 0.78]	11.6%
Sleijfer 2009: Pazopanib	37	20	0.54	[0.38; 0.69]	12.5%
Random effects model	89		0.59	[0.39; 0.76]	34.9%
Heterogeneity: $I^2 = 0\%$, $p = 0.78$					
Random effects model	**242**		**0.45**	**[0.34; 0.57]**	**100.0%**
Heterogeneity: $I^2 = 72\%$, $p < 0.01$					
Test for subgroup differences: $\chi^2_1 = 2.89$, df = 1 ($p = 0.09$)					

Study	Patients	Cases	PFS 6m	95% CI	Weight
regimen_recommended = No					
Schöffski 2011: Eribulin	27	3	0.11	[0.04; 0.29]	10.1%
Chugh 2009: Imatinib	21	0	0.00	[0.00; 0.28]	3.0%
Tawbi 2017: Pembrolizumab	10	2	0.20	[0.05; 0.54]	7.5%
Mir 2016: Regorafenib	13	5	0.38	[0.17; 0.66]	10.9%
Robbins 2015: Cycl + Flud + TCR	19	7	0.37	[0.19; 0.60]	12.8%
Ray-Coquard 2008: Gefitinib	46	4	0.09	[0.03; 0.21]	11.8%
Random effects model	136		0.19	[0.10; 0.32]	56.1%
Heterogeneity: $I^2 = 62\%$, $p = 0.02$					
regimen_recommended = Yes					
Buonadonna 2017: Trabectedin	23	6	0.26	[0.12; 0.47]	12.9%
van der Graaf 2012: Pazopanib	29	13	0.45	[0.28; 0.63]	15.2%
Sleijfer 2009: Pazopanib	37	12	0.32	[0.19; 0.49]	15.8%
Random effects model	89		0.34	[0.19; 0.54]	43.9%
Heterogeneity: $I^2 = 5\%$, $p = 0.35$					
Random effects model	**225**		**0.25**	**[0.16; 0.36]**	**100.0%**
Heterogeneity: $I^2 = 61\%$, $p < 0.01$					
Test for subgroup differences: $\chi^2_1 = 2.38$, df = 1 ($p = 0.12$)					

Figure 3.5: Forest plots of PFS at 3 (upper panel) and 6 (low panel) months for pre-treated SS patients. Cycl + Flud + TCR = Cyclophosphamide + fludarabine + TCR transduced cells.

3.3.6 Sensitivity meta-analyses

Regarding LPS (see Appendix section 2.3), Baujat plots detected 'Blay 2014: Trabectedin' [53] as a potentially influential treatment regimen for first-line at 3 and 6 months (overall pooled PFSR decreased 2% and 3% after the exclusion of this treatment regimen). Overall heterogeneity slightly decreased. For patients previously treated with systemic therapy, 'Samuels 2017: pazopanib' [44] was identified by Baujat plots and diagnostics (overall PFSR decreased 2% and 1% at 3 and 6 months but heterogeneity did not go down). Results were robust to the candidate outlier in the pre-treated setting and less robust in the first-line setting.

Secondly for SS (see Appendix section 2.4), the plots and diagnostics for first-line agents pointed out 'Judson 2014: doxorubicin+ifosfamide' [32] as the most influential study (overall pooled PFSR decreased 6% at 3 months and 10% at 6 months after removing it from the database). Overall heterogeneity dropped substantially, which could be expected because of the limited studies here (three clinical trials, five regimens). For pre-treated population, the treatment regimen of 'Robbins 2015: cyclophosphamide + fludarabine + TCR transduced cells' [57] was detected as outlier (overall rate decreased 2% and 1% but heterogeneity did not change substantially). Findings showed that meta-analyses were robust in the pre-treated but not robust in first-line setting (because of the only five treatment regimens in total).

3.3.7 New benchmarks

Similar to our previous LMS meta-analysis [13], the overall pooled PFSRs at 3 or 6 months are used as the reference values for the parameter P_0 (null hypothesis). To elaborate on this, for all LMS, PFS rates did not differ significantly between the two groups of drugs (R-T, NR-T) for first or further lines of treatment. Here, results for LPS and SS were concordant with the exception of previously treated LPS patients where differences between R-T and NR-T were significant. For the sake of consistency, it was decided to use the overall pooled rates to guide P_0. To calculate the reference values of the parameter P_1 (alternative hypothesis), the ESMO-MCBS suggestions [15] in an advanced/metastatic setting were employed assuming an exponential PFS curve. The tool recommends a hazard ratio (HR) ≤ 0.65 (scale evaluation form 2b).

Parameters P_0 and P_1 are provided in table 2 per treatment line and analysed group. For LPS, the minimum values to reach for suggesting drug activity in first-line patients are 79% and 69% (82% and 69% for SS) at 3 and 6 months. For pre-treated patients, recommended rates are 63% and 44% (60% and 41% for SS), respectively. Owing to the limited numbers of studies and the differences between primary and sensitivity analyses, benchmarks for first-line SS patients have to be interpreted with caution. Please see fig. 6 of Ref. [13] for further details on how to use these benchmarks (P_0, P_1) to aid the design of new phase II studies.

	3 months		**6 months**	
Treatment line and analysed group	**Ref (P_0)**	**Min target (P_1)**	**Ref (P_0)**	**Min target (P_1)**
First-line LPS	69%	79%	56%	69%
First-line SS	74%	82%	56%	69%
Pre-treated LPS	49%	63%	28%	44%
Pre-treated SS	45%	60%	25%	41%

Table 3.2: **Treatment effect (PFSR) for the null hypothesis (H_0) parameter P_0 and the alternative hypothesis (H_1) parameter P_1 of a study for LPS or SS patients.**. Note: LPS, liposarcoma. SS, synovial sarcoma. PFSRs for SS are presented in shade. The overall pooled PFSRs at 3 and 6 months were used to provide reference values for P_0. Using the recommended treatment effect for PFS by the ESMO Magnitude of Clinical Benefit Scale (ESMO-MCBS), minimum values to target for P_1 were calculated.

Suppose that we would like to calculate the sample size for a new phase II trial with pre-treated LPS patients and a single-stage A'Hern design [58], given the new thresholds, assuming that the primary endpoint is PFSR at 3 months (i.e. $P_0 = 49\%$, $P_1 = 63\%$). The power and sample size are computed under the alternative hypothesis that P = P_1. For a type I error 10% (α = 0.10) and 80% power (β = 0.20), a total of 60 eligible patients will need to be treated and followed for the assessment of the primary endpoint. This design would then require 35 patients alive and progression-free to justify further drug investigation.

3.4 Discussion

This research project yielded efficacy thresholds to design new phase II clinical trials for advanced/metastatic LPS or SS patients with PFSR at 3 or 6 months as the primary endpoint, based on meta-analyses of summary data collected from sponsors and published papers (2003 – 2018). Reference values were estimated for the parameter of null hypothesis (P_0) as the overall pooled PFSR per treatment line, and new values were calculated for the parameter of alternative hypothesis (P_1) using the recommended treatment effects to target by the ESMO-MCBS recommendations [15].

Two decades ago, the Van Glabbeke study [12] suggested benchmarks for various STS patients who participated in phase II clinical trials of the EORTC - STBSG database for treatment with inactive (used for P_0) or active agents (used for P_1). Hereto, the authors performed an individual patient data (IPD) meta-analysis using progression-free rate as the principal endpoint. In first-line setting with anthracycline-containing regimens, a rate of 64% or 55% suggested drug activity for LPS (n = 110), whereas for SS (n = 115), a rate of 77% or 56% at 3 or 6 months, respectively. On the other hand, in the pre-treated setting, reference values for activity were calculated based on 146 patients from all STS subgroups (39% or 14% at 3 and 6 months, respectively). These values have now been updated and re-evaluated for LPS and SS, per treatment line, to reflect current practice (see table 3.2). However, a direct comparison is not meaningful since here we used summary estimates (and not IPD) of a larger number of patients per histotype: 1030 LPS (310 first-line), 348 SS (106 first-line) from phase II, III, or even IV clinical trials, defined benchmarks separately for first-line or pre-treated LPS or SS patients by employing the overall pooled PFSR as P_0 (based on inactive and active agents) and the ESMO-MCBS tool to target P_1, and used PFSR (any death counted as an event) instead of progression-free rate as the primary endpoint.

Based on our sensitivity meta-analyses, the new thresholds were shown to be robust (stable) in pre-treated LPS and SS patients. However, values were less robust for first-line LPS, and not robust for first-line SS. Removing one outlier decreases the 6-month PFSR by 10% for first-line SS. This indicates an inconsistent estimate (there), which was expected due to the very limited number of studies (three clinical trials – five treatment regimens). Publication bias was not observed based on the tests except for pre-treated LPS patients at 6 months. A high risk of publication bias could lead to a biased estimate of the summary effect. This is a further reason to push publication of trials regardless of their results. Heterogeneity between studies was moderate to high for first-line LPS or SS patients ($I^2 > 40\%$), as well as high for pre-treated LPS or SS ($I^2 > 60\%$). Note that a (very) high overall heterogeneity (I^2) indicates a large variation between-study-specific effect sizes which could challenge the validity of the meta-analyses. In particular, results for pre-treated subjects should be interpreted with caution due to substantial variability. Heterogeneity could not be explained by meta-regressions (subgroup meta-analyses). Findings of excessive heterogeneity are consistent with our previous work for all LMS (Kantidakis et al. 2021 [13]). Further research is needed to better address this heterogeneity.

Benchmarks provided in this manuscript are directly comparable with those for LMS [13] since they are based on the same literature review for STS and estimated using the same methodology. For first-line treatment, to suggest drug activity, the proposed 3-month PFSRs are slightly higher for all LMS / SS (82% for both) versus LPS (79%). Differences at 6 months are minimal (70% for all LMS, versus 69% for LPS / SS). For second or later lines, values to reach for LPS (63% and 44% at 3 and 6 months, respectively) and all LMS (62% and 44%) are a bit higher than those recommended for SS patients (60% and 41%). Thus, a need to raise the bar of thresholds for the commonest STS types in future phase II trials is indicated by both of our studies, which aligns with the perspective of the American Society of Clinical Oncology [59]. The cost-benefit of new systemic therapies for cancer should be balanced against the societal resources in this era of rapidly rising healthcare costs.

These manuscripts share a number of limitations. First, the large majority of the trials were designed for several STS types and are therefore underpowered for specific subgroup analyses (i.e. here for LPS and SS). This could explain the non-significant difference between recommended and non-recommended treatments based on the standard ESMO guidelines [23] for first-line LPS/SS and pre-treated SS patients (and also for all LMS in the previous study). Secondly, PFSRs were calculated based on summary estimates per treatment arm and treatment line, which are less reliable than IPD data but require a smaller amount of time to be collected from the different study sponsors. Thirdly, LPS were addressed as a single disease while it is known that there are three different LPS histologic subtypes (e.g. well differentiated/dedifferentiated, myxoid, or pleomorphic) that exhibit different clinical behaviour and sensitivity to treatments. Yet, in older studies, such information might not have been collected at the subtype level. Moreover, the condition of any meta-analysis that the effect sizes between drugs of the same trial are independent may be violated in the randomised studies, as a random-effect model was used for each treatment regimen. We observed a high unexplained overall heterogeneity indicative of a large variation between effect sizes, which may limit our meta-analytic results. Finally, as emphasised in our previous meta-analysis for LMS [13], strong surrogacy properties between PFS and OS are questionable based on two meta-analyses of randomised studies with advanced STS [60, 61]. Thus, PFS might lead to exaggerated enthusiasm for a new anticancer therapy (see Refs. [5, 62]). As such, PFS can be used as the primary endpoint in phase II trials or as futility endpoint in phase III trials, but OS should remain the optimal primary endpoint in phase III trials.

For instance, the sample size of EORTC 1202 study for second-line patients with metastatic or inoperable locally advanced dedifferentiated LPS with cabazitaxel [63] was calculated based on a Simon two-stage optimal design ($\alpha = \beta = 0.10$) [64] and the Van Glabbeke rules (P_0 = 20%, P_1 = 40%). Stage one required 4/17 eligible patients progression-free, and stage two required 11/37 eligible patients progression-free at 12 weeks. Hence, according to these rules, the 1202 study has met its primary endpoint (21/38 or 55.3% of patients progression-free at 12 weeks) indicating activity of cabazitaxel. Nevertheless, according to the new values (i.e. P_0 = 49%, P_1 = 63%, see table 3.2), it may be challenging to obtain a significant and relevant improvement over a standard of care in a prospective randomised phase III trial. Note that our new benchmarks might require relatively large sample sizes for new phase II studies because of the smaller target difference between P_0 and P_1 compared to the ones previously proposed. This could be overcome by targeting a larger treatment difference, e.g. P_1 = 69% instead of 63% for a P_0 = 49%, or to choose PFSR at 6 months as the endpoint where the differences between P_0 and P_1 are larger. Our analyses clearly show that the cut-offs provided by Van Glabbeke et al. are suboptimal (3- and 6-month rates of P_1 = 64%, 55% for LPS, and 77%, 56% for SS in first-line, 40% and 20% for any histologic subtype in pre-treated setting), they can no longer pave the way to new standard of care. Our benchmarks are setting the bar higher, aiming to identify earlier in the drug development process compounds which have a higher chance to impact clinical practice. If traditional clinical trial designs are deemed unfeasible, more complex and flexible options (e.g. adaptive designs) could be considered. Especially in ultra-rare sarcomas or when accrual is particularly demanding in terms of numbers or timeframe, recruitment challenges could be overcome through international, multi-centre trials.

There are certain LPS subtypes that could benefit from non-licensed agents. For instance, trabectedin was shown to be highly active for first-line myxoid LPS in the Blay 2014 study [53] (3- and 6-month PFSR of 96%), and pembrolizumab is currently used on an individual basis for dedifferentiated LPS, but as they are not formally approved for front-line treatment of STS, they are not recommended for first-line treatment of LPS according to the ESMO 2021 clinical practice guidelines. Prospective data to support emerging agents are currently lacking, and this is preventing their adoption in practice. Even if randomised controlled trials are the golden standard, real-world evidence or single-arm phase I/II trials can be helpful for cancer types with rare/ultra-rare indications – including many STS – to accelerate the development and approval of new anticancer treatments [65].

Mesenchymal tumours (i.e. STS) are regarded as one of the most challenging fields of diagnostic pathology [66]. An accurate diagnosis is laborious for non-specialised pathologists. Data have indicated a proportion of diagnostic error 25-40% in STS [6, 67]. It may also be challenging to obtain the correct classification within a histological type (e.g. well differentiated could be re-graded as dedifferentiated LPS) [68]. Patients should have computed tomography (CT) scans performed within reference sarcoma centres to improve diagnosis and tailoring treatment allocation [11]. Furthermore, STS have demonstrated a tremendous heterogeneity (genetic and histologic diversity, clinical prognosis, metastatic patterns, etc.) [69]. Therefore, the management of adult STS requires a multidisciplinary approach where collaboration is key to allow sufficiently large studies [4].

In advanced/metastatic STS, therapeutic options beyond first-line (anthracyclines) are increasingly driven by histology. An urgent need remains for the development of individualised treatment plans such as targeted therapy to move away from the conventional chemotherapy options. This work provides modern thresholds for suggesting drug activity, this time for LPS and SS patients, to aid the design of new histology-tailored phase II trials using PFSR at 3 or 6 months as endpoint. We hope that phase II studies which meet the updated thresholds for these histotypes will then lead to higher success rates in new prospective phase III trials to avoid the large costs associated with their failure.

Declarations

Role of the funding source

This work was supported by the European Organisation for Research and Treatment of Cancer – Soft Tissue and Bone Sarcoma Group (EORTC - STBSG). The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Acknowledgements

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 as well as from the EORTC Cancer Research Fund (ECRF). This publication was supported by a donation from Kom Op Tegen Kanker (Stand up to Cancer), the Flemish cancer society from Belgium. The authors would like to express their appreciation to all the primary investigators (PI) who shared valuable PFS estimates at 3 and 6 months for liposarcoma (LPS) and/or synovial sarcoma (SS) patients, namely, Dr Akira Kawai (1 study; LPS), Dr PA Cassier (1 study; LPS), Dr Mark A. Dickson (2 studies; LPS), Dr Joan Maurel (1 study; LPS), Dr Beatrice Seddon (1 study; LPS), Dr. Paul F. Robbins (1 study; SS), and Prof Antoine Italiano (1 study; LPS). In addition, we would like to sincerely thank the following sponsors SARC (4 studies: PI Prof Scott M. Schuetze; LPS, PI Prof Rashmi Chugh; LPS and SS, PI Prof Hussein A Tawbi; LPS and SS, PI Dr William D Tap; LPS and SS), PharmaMar (2 studies: PI Prof Nicolas Penel; LPS and SS, PI Prof Jean-Yves Blay; LPS and SS), Eisai (1 study: PI Prof Patrick Schöffski; LPS), and Taiho Pharmaceutical (1 study: PI Dr Akira Kawai; LPS) for the provision of summary PFS estimates at 3 and 6 months. All analyses and conclusions in this article are the sole responsibility of the authors and do not necessarily reflect the opinions or views of these primary investigators, SARC, PharmaMar, Eisai, or Taiho Pharmaceutical.

Online supplementary materials

The Appendix of this Chapter is available online at https://github.com/GKantidakis/Thesis_supplementary_materials/blob/main/Chapter3/Appendix.docx.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] W. D. Tap, R. L. Jones, B. A. Van Tine, B. Chmielowski, A. D. Elias, D. Adkins, M. Agulnik, M. M. Cooney, M. B. Livingston, G. Pennock, M. R. Hameed, G. D. Shah, A. Qin, A. Shahir, D. M. Cronier, R. Ilaria, I. Conti, J. Cosaert, and G. K. Schwartz. Olaratumab and doxorubicin versus doxorubicin alone for treatment of soft-tissue sarcoma: an open-label phase 1b and randomised phase 2 trial. *The Lancet*, 388(10043):488–497, jul 2016. ISSN 1474-547X. doi: 10.1016/S0140-6736(16)30587-6. URL https://pubmed.ncbi.nlm.nih.gov/27291997/.

[6] F. Chamberlain, C. Benson, K. Thway, P. Huang, R. L. Jones, and S. Gennatas. Pharmacotherapy for liposarcoma: Current and emerging synthetic treatments. *Future Oncology*, 17(20):2659–2670, 2021. ISSN 17448301. doi: 10.2217/fon-2020-1092.

[7] L. Yang, S. Chen, P. Luo, W. Yan, and C. Wang. Liposarcoma: Advances in cellular and molecular genetics alterations and corresponding clinical treatment. *Journal of Cancer*, 11(1):100–107, 2020. ISSN 18379664. doi: 10.7150/jca.36380.

[8] E. Z. Keung and N. Somaiah. Overview of liposarcomas and their genomic landscape. *Journal of Translational Genetics and Genomics*, 3:8, 2019. doi: 10.20517/jtgg.2019.03.

[9] M. N. Aytekin, R. Öztürk, K. Amer, and A. Yapar. Epidemiology, incidence, and survival of synovial sarcoma subtypes: SEER database analysis. *Journal of Orthopaedic Surgery*, 28(2):1–12, 2020. ISSN 23094990. doi: 10.1177/2309499020936009.

[10] A. M. Gazendam, S. Popovic, S. Munir, N. Parasu, D. Wilson, and M. Ghert. Synovial sarcoma: A clinical review. *Current Oncology*, 28(3):1909–1920, 2021. ISSN 17187729. doi: 10.3390/curroncol28030177.

[11] S. Stacchiotti and B. A. Van Tine. Synovial sarcoma: Current concepts and future perspectives. *Journal of Clinical Oncology*, 36(2):180–187, 2018. ISSN 15277755. doi: 10.1200/JCO.2017.75.1941.

[12] 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.

[13] 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.

[14] P. G. Casali, N. Abecassis, H. T. Aro, S. Bauer, R. Biagini, S. Bielack, S. Bonvalot, I. Boukovinas, J. V. M. G. Bovee, T. Brodowicz, J. M. Broto, A. Buonadonna, E. De Álava, A. P. Dei Tos, X. G. Del Muro, P. Dileo, M. Eriksson, A. Fedenko, V. Ferraresi, A. Ferrari, S. Ferrari, A. M. Frezza, S. Gasperoni, H. Gelderblom, T. Gil, G. Grignani, A. Gronchi, R. L. Haas, B. Hassan, P. Hohenberger, R. Issels, H. Joensuu, R. L. Jones,

I. Judson, P. Jutte, S. Kaal, B. Kasper, K. Kopeckova, D. A. Krákorová, A. Le Cesne, I. Lugowska, O. Merimsky, M. Montemurro, M. A. Pantaleo, R. Piana, P. Picci, S. Piperno-Neumann, A. L. Pousa, P. Reichardt, M. H. Robinson, P. Rutkowski, A. A. Safwat, P. Schöffski, S. Sleijfer, S. Stacchiotti, K. Sundby Hall, M. Unk, F. Van Coevorden, W. T. A. Van Der Graaf, J. Whelan, E. Wardelmann, O. Zaikova, and J. Y. Blay. Soft tissue and visceral sarcomas: ESMO-EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 29(Supplement_4):iv51—-iv67, 2018. ISSN 15698041. doi: 10.1093/annonc/mdy321.

[15] N. I. Cherny, U. Dafni, J. Bogaerts, N. J. Latino, G. Pentheroudakis, J. Y. Douillard, J. Tabernero, C. Zielinski, M. J. Piccart, and E. G. E. de Vries. ESMO-Magnitude of Clinical Benefit Scale version 1.1. *Annals of Oncology*, 28(10):2340–2366, 2017. doi: 10.1093/annonc/mdx310.

[16] G. de Pinieux, M. Karanian, F. Le Loarer, S. Le Guellec, S. Chabaud, P. Terrier, C. Bouvier, M. Batistella, A. Neuville, Y. M. Robin, J. F. Emile, A. Moreau, F. Larousserie, A. Leroux, N. Stock, M. Lae, F. Collin, N. Weinbreck, S. Aubert, F. Mishellany, C. Charon-Barra, S. Croce, L. Doucet, I. Quintin-Rouet, M. C. Chateau, C. Bazille, I. Valo, B. Chetaille, N. Ortonne, A. Brouchet, P. Rochaix, A. Demuret, J. P. Ghnassia, L. Mescam, N. Macagno, I. Birtwisle-Peyrottes, C. Delfour, E. Angot, I. Pommepuy, D. Ranchere, C. Chemin-Airiau, M. Jean-Denis, Y. Fayet, J. B. Courrèges, N. Mesli, J. Berchoud, M. Toulmonde, A. Italiano, A. Le Cesne, N. Penel, F. Ducimetiere, F. Gouin, J. M. Coindre, and J. Y. Blay. Nationwide incidence of sarcomas and connective tissue tumors of intermediate malignancy over four years using an expert pathology review network. *PLoS ONE*, 16(2):e0246958, 2021. ISSN 19326203. doi: 10.1371/journal.pone.0246958.

[17] A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. A. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen, and D. Moher. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of clinical epidemiology*, 62(10):e1—-e34, 2009. ISSN 18785921. doi: 10.1016/j.jclinepi.2009.06.006.

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

[19] 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.

[20] W. Feller. On the Normal Approximation to the Binomial Distribution. *The Annals of Mathematical Statistics*, 16(4):319–329, 1945. ISSN 0003-4851. doi: 10.1214/aoms/1177731058. URL https://projecteuclid.org/euclid.aoms/1177731058.

[21] N. Wang. How to Conduct a Meta-Analysis of Proportions in R : A Comprehensive Tutorial Conducting Meta-Analyses of Proportions in R. *John Jay Coll Crim Justice*, pages 1–62, 2018. doi: 10.13140/RG.2.2.27199.00161.

[22] 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.

[23] A. Gronchi, A. B. Miah, A. P. Dei Tos, N. Abecassis, J. Bajpai, S. Bauer, R. Biagini, S. Bielack, J. Y. Blay, S. Bolle, S. Bonvalot, I. Boukovinas, J. V. M. G. Bovee, K. Boye, B. Brennan, T. Brodowicz, A. Buonadonna, E. De Álava, X. G. Del Muro, A. Dufresne, M. Eriksson, F. Fagioli, A. Fedenko, V. Ferraresi, A. Ferrari, A. M. Frezza, S. Gasperoni, H. Gelderblom, F. Gouin, G. Grignani, R. Haas, A. B. Hassan, S. Hecker-Nolting, N. Hindi, P. Hohenberger, H. Joensuu, R. L. Jones, C. Jungels, P. Jutte, L. Kager, B. Kasper, A. Kawai, K. Kopeckova, D. A. Krákorová, A. Le Cesne, F. Le Grange, E. Legius, A. Leithner, A. Lopez-Pousa, J. Martin-Broto, O. Merimsky, C. Messiou, O. Mir, M. Montemurro, B. Morland, C. Morosi, E. Palmerini, M. A. Pantaleo, R. Piana, S. Piperno-Neumann, P. Reichardt, P. Rutkowski, A. A. Safwat, C. Sangalli, M. Sbaraglia, S. Scheipl, P. Schöffski, S. Sleijfer, D. Strauss, S. Strauss, K. Sundby

Hall, A. Trama, M. Unk, M. A. J. van de Sande, W. T. A. van der Graaf, W. J. van Houdt, T. Frebourg, P. G. Casali, and S. Stacchiotti. Soft tissue and visceral sarcomas: ESMO–EURACAN–GENTURIS Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 32(11):1348–1365, 2021. ISSN 15698041. doi: 10.1016/j.annonc.2021.07.006. URL https://doi.org/10.1016/j.annonc.2021.07.006.

[24] J. L. Peters, A. J. Sutton, D. R. Jones, K. R. Abrams, and L. Rushton. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology*, 61(10):991–996, oct 2008. ISSN 08954356. doi: 10.1016/j.jclinepi.2007.11.010.

[25] M. Egger, G. D. Smith, M. Schneider, and C. Minder. Bias in meta-analysis detected by a simple, graphical test measures of funnel plot asymmetry. *BMJ*, 315(7109):629–634, 1997. doi: 10.1136/bmj.315.7109.629.

[26] C. B. Begg and M. Mazumdar. Operating Characteristics of a Rank Correlation Test for Publication Bias. *Biometrics*, 50(4):1088–1101, 1994. doi: 10.2307/2533446. URL https://www.jstor.org/stable/pdf/2533446.pdf.

[27] B. Baujat, C. Mahé, J. P. Pignon, and C. Hill. A graphical method for exploring heterogeneity in meta-analyses: Application to a meta-analysis of 65 trials. *Statistics in Medicine*, 21(18):2641–2652, 2002. doi: 10.1002/sim.1221.

[28] W. Viechtbauer. Conducting meta-analyses in R with the metafor. *Journal of Statistical Software*, 36(3): 1–48, 2010. ISSN 15487660.

[29] G. Swarzer. meta: an R package for meta-analysis. *R News*, 7(3):40–45, 2007. ISSN 1609-3631. URL https://www.researchgate.net/publication/285729385_meta_An_R_Package_for_Meta-Analysis.

[30] R. Chugh, J. K. Wathen, R. G. Maki, R. S. Benjamin, S. R. Patel, P. A. Myers, D. A. Priebat, D. K. Reinke, D. G. Thomas, M. L. Keohan, B. L. Samuels, and L. H. Baker. Phase II multicenter trial of imatinib in 10 histologic subtypes of sarcoma using a bayesian hierarchical statistical model. *Journal of Clinical Oncology*, 27(19):3148–3153, 2009. ISSN 0732183X. doi: 10.1200/JCO.2008.20.5054.

[31] J. Maurel, A. López-Pousa, R. De Las Peñas, J. Fra, J. Martín, J. Cruz, A. Casado, A. Poveda, J. Martínez-Trufero, C. Balañá, M. A. Gómez, R. Cubedo, O. Gallego, B. Rubio-Viqueira, J. Rubió, R. Andrós, I. Sevilla, J. J. De La Cruz, X. G. Del Muro, and J. M. Buesa. Efficacy of sequential high-dose doxorubicin and ifosfamide compared with standard-dose doxorubicin in patients with advanced soft tissue sarcoma: An open-label randomized phase II study of the Spanish group for research on sarcomas. *Journal of Clinical Oncology*, 27(11):1893–1898, 2009. ISSN 0732183X. doi: 10.1200/JCO.2008.19.2930.

[32] I. Judson, J. Verweij, H. Gelderblom, J. T. Hartmann, P. Schöffski, J. Y. Blay, J. M. Kerst, J. Sufliarsky, J. Whelan, P. Hohenberger, A. Krarup-Hansen, T. Alcindor, S. Marréaud, S. Litière, C. Hermans, C. Fisher, P. C. W. Hogendoorn, A. P. Dei Tos, and W. T. A. van der Graaf. Doxorubicin alone versus intensified doxorubicin plus ifosfamide for first-line treatment of advanced or metastatic soft-tissue sarcoma: A randomised controlled phase 3 trial. *The Lancet Oncology*, 15(4):415–423, 2014. ISSN 14745488. doi: 10.1016/S1470-2045(14)70063-4.

[33] B. Bui-Nguyen, J. E. Butrynski, N. Penel, J. Y. Blay, N. Isambert, M. Milhem, J. M. Kerst, A. K. L. Reyners, S. Litière, S. Marréaud, F. Collin, and W. T. A. van der Graaf. A phase IIb multicentre study comparing the efficacy of trabectedin to doxorubicin in patients with advanced or metastatic untreated soft tissue sarcoma: The TRUSTS trial. *European Journal of Cancer*, 51(10):1312–1320, 2015. ISSN 18790852. doi: 10.1016/j.ejca.2015.03.023.

[34] A. Kawai, N. Araki, H. Sugiura, T. Ueda, T. Yonemoto, M. Takahashi, H. Morioka, H. Hiraga, T. Hiruma, T. Kunisada, A. Matsumine, T. Tanase, T. Hasegawa, and S. Takahashi. Trabectedin monotherapy after

standard chemotherapy versus best supportive care in patients with advanced, translocation-related sarcoma: A randomised, open-label, phase 2 study. *The Lancet Oncology*, 16(4):406–416, 2015. ISSN 14745488. doi: 10.1016/S1470-2045(15)70098-7. URL `http://dx.doi.org/10.1016/S1470-2045(15)70098-7`.

[35] M. Toulmonde, A. Le Cesne, S. Piperno-Neumann, N. Penel, C. Chevreau, F. Duffaud, C. Bellera, and A. Italiano. Aplidin in patients with advanced dedifferentiated liposarcomas: A French Sarcoma Group Single-Arm Phase II study. *Annals of Oncology*, 26(7):1465–1470, 2015. ISSN 15698041. doi: 10.1093/annonc/mdv195.

[36] G. D. Demetri, M. Von Mehren, R. L. Jones, M. L. Hensley, S. M. Schuetze, A. Staddon, M. Milhem, A. Elias, K. Ganjoo, H. Tawbi, B. A. Van Tine, A. Spira, A. Dean, N. Z. Khokhar, Y. C. Park, R. E. Knoblauch, T. V. Parekh, R. G. Maki, and S. R. Patel. Efficacy and safety of trabectedin or dacarbazine for metastatic liposarcoma or leiomyosarcoma after failure of conventional chemotherapy: Results of a phase III randomized multicenter clinical trial. *Journal of Clinical Oncology*, 34(8):786–793, 2016. ISSN 15277755. doi: 10.1200/JCO.2015.62.4734.

[37] M. A. Dickson, G. K. Schwartz, M. Louise Keohan, S. P. D'Angelo, M. M. Gounder, P. Chi, C. R. Antonescu, J. Landa, L. X. Qin, A. M. Crago, S. Singer, A. Koff, and W. D. Tap. Phase 2 Trial of the CDK4 inhibitor Palbociclib (PD0332991) at 125 mg dose in Well-Differentiated or Dedifferentiated Liposarcoma. *JAMA oncology*, 2(7):937–940, 2016. ISSN 23742445. doi: 10.1001/jamaoncol.2016.0264.

[38] 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, dec 2016. ISSN 14745488. doi: 10.1016/S1470-2045(16)30507-1.

[39] P. Schöffski, S. Chawla, R. G. Maki, A. Italiano, H. Gelderblom, E. Choy, G. Grignani, V. Camargo, S. Bauer, S. Y. Rha, J. Y. Blay, P. Hohenberger, D. D'Adamo, M. Guo, B. Chmielowski, A. Le Cesne, G. D. Demetri, and S. R. Patel. Eribulin versus dacarbazine in previously treated patients with advanced liposarcoma or leiomyosarcoma: A randomised, open-label, multicentre, phase 3 trial. *The Lancet*, 387(10028):1629–1637, 2016. ISSN 1474547X. doi: 10.1016/S0140-6736(15)01283-0.

[40] S. M. Schuetze, J. K. Wathen, D. R. Lucas, E. Choy, B. L. Samuels, A. P. Staddon, K. N. Ganjoo, M. Von Mehren, W. A. Chow, D. M. Loeb, H. A. Tawbi, D. A. Rushing, S. R. Patel, D. G. Thomas, R. Chugh, D. K. Reinke, and L. H. Baker. SARC009: Phase 2 study of dasatinib in patients with previously treated, high-grade, advanced sarcoma. *Cancer*, 122(6):868–874, 2016. ISSN 10970142. doi: 10.1002/cncr.29858.

[41] A. Buonadonna, C. Benson, J. Casanova, B. Kasper, A. López Pousa, F. Mazzeo, T. Brodowicz, and N. Penel. A noninterventional, multicenter, prospective phase IV study of trabectedin in patients with advanced soft tissue sarcoma. *Anti-Cancer Drugs*, 28(10):1157–1165, 2017. ISSN 14735741. doi: 10.1097/CAD.0000000000000560.

[42] S. Sleijfer, I. Ray-Coquard, Z. Papai, A. Le Cesne, M. Scurr, P. Schöffski, F. Collin, L. Pandite, S. Marréaud, A. De Brauwer, M. Van Glabbeke, J. Verweij, and J. Y. Blay. Pazopanib, a multikinase angiogenesis inhibitor, in patients with relapsed or refractory advanced soft tissue sarcoma: A phase II study from the European organisation for research and treatment of cancer-soft tissue and bone sarcoma group (EORTC Study 620. *Journal of Clinical Oncology*, 27(19):3126–3132, 2009. ISSN 0732183X. doi: 10.1200/JCO.2008.21.3223.

[43] A. Kawai, N. Araki, Y. Naito, T. Ozaki, H. Sugiura, Y. Yazawa, H. Morioka, A. Matsumine, K. Saito, S. Asami, and K. Isu. Phase 2 study of eribulin in patients with previously treated advanced or metastatic soft tissue sarcoma. *Japanese Journal of Clinical Oncology*, 47(2):137–144, 2017. ISSN 14653621. doi: 10.1093/jjco/hyw175.

[44] B. L. Samuels, S. P. Chawla, N. Somaiah, A. P. Staddon, K. M. Skubitz, M. M. Milhem, P. E. Kaiser, D. C. Portnoy, D. A. Priebat, M. S. Walker, and E. J. Stepanski. Results of a prospective phase 2 study of pazopanib in patients with advanced intermediate-grade or high-grade liposarcoma. *Cancer*, 123(23):4640–4647, 2017. ISSN 10970142. doi: 10.1002/cncr.30926.

[45] B. Seddon, S. J. Strauss, J. Whelan, M. Leahy, P. J. Woll, F. Cowie, C. Rothermundt, Z. Wood, C. Benson, N. Ali, M. Marples, G. J. Veal, D. Jamieson, K. Küver, R. Tirabosco, S. Forsyth, S. Nash, H. M. Dehbi, and S. Beare. Gemcitabine and docetaxel versus doxorubicin as first-line treatment in previously untreated advanced unresectable or metastatic soft-tissue sarcomas (GeDDiS): a randomised controlled phase 3 trial. *The Lancet Oncology*, 18(10):1397–1410, 2017. ISSN 14745488. doi: 10.1016/S1470-2045(17)30622-8.

[46] W. D. Tap, Z. Papai, B. A. Van Tine, S. Attia, K. N. Ganjoo, R. L. Jones, S. Schuetze, D. Reed, S. P. Chawla, R. F. Riedel, A. Krarup-Hansen, M. Toulmonde, I. Ray-Coquard, P. Hohenberger, G. Grignani, L. D. Cranmer, S. Okuno, M. Agulnik, W. Read, C. W. Ryan, T. Alcindor, X. F. G. del Muro, G. T. Budd, H. Tawbi, T. Pearce, S. Kroll, D. K. Reinke, and P. Schöffski. Doxorubicin plus evofosfamide versus doxorubicin alone in locally advanced, unresectable or metastatic soft-tissue sarcoma (TH CR-406/SARC021): an international, multicentre, open-label, randomised phase 3 trial. *The Lancet Oncology*, 18(8):1089–1103, aug 2017. ISSN 14745488. doi: 10.1016/S1470-2045(17)30381-9.

[47] H. A. Tawbi, M. Burgess, V. Bolejack, B. A. Van Tine, S. M. Schuetze, J. Hu, S. D'Angelo, S. Attia, R. F. Riedel, D. A. Priebat, S. Movva, L. E. Davis, S. H. Okuno, D. R. Reed, J. Crowley, L. H. Butterfield, R. Salazar, J. Rodriguez-Canales, A. J. Lazar, I. I. Wistuba, L. H. Baker, R. G. Maki, D. Reinke, and S. Patel. Pembrolizumab in advanced soft-tissue sarcoma and bone sarcoma (SARC028): a multicentre, two-cohort, single-arm, open-label, phase 2 trial. *The Lancet Oncology*, 18(11):1493–1501, 2017. ISSN 14745488. doi: 10.1016/S1470-2045(17)30624-1. URL http://dx.doi.org/10.1016/S1470-2045(17)30624-1.

[48] P. Schöffski, I. L. Ray-Coquard, A. Cioffi, N. B. Bui, S. Bauer, J. T. Hartmann, A. Krarup-Hansen, V. Grünwald, R. Sciot, H. Dumez, J. Y. Blay, A. Le Cesne, J. Wanders, C. Hayward, S. Marréaud, M. Ouali, and P. Hohenberger. Activity of eribulin mesylate in patients with soft-tissue sarcoma: A phase 2 study in four independent histological subtypes. *The Lancet Oncology*, 12(11):1045–1052, 2011. ISSN 14702045. doi: 10.1016/S1470-2045(11)70230-3.

[49] S. P. Chawla, A. P. Staddon, L. H. Baker, S. M. Schuetze, A. W. Tolcher, G. Z. D'Amato, J. Y. Blay, M. M. Mita, K. K. Sankhala, L. Berk, V. M. Rivera, T. Clackson, J. W. Loewy, F. G. Haluska, and G. D. Demetri. Phase II study of the mammalian target of rapamycin inhibitor ridaforolimus in patients with advanced bone and soft tissue sarcomas. *Journal of Clinical Oncology*, 30(1):78–84, 2012. ISSN 15277755. doi: 10.1200/JCO.2011.35.6329.

[50] P. A. Cassier, A. Lefranc, E. Y Amela, C. Chevreau, B. N. Bui, A. Lecesne, I. Ray-Coquard, S. Chabaud, N. Penel, Y. Berge, J. Dômont, A. Italiano, F. Duffaud, A. C. Cadore, V. Polivka, and J. Y. Blay. A phase II trial of panobinostat in patients with advanced pretreated soft tissue sarcoma. A study from the French Sarcoma Group. *British Journal of Cancer*, 109(4):909–914, 2013. ISSN 00070920. doi: 10.1038/bjc.2013.442.

[51] M. A. Dickson, W. D. Tap, M. L. Keohan, S. P. D'Angelo, M. M. Gounder, C. R. Antonescu, J. Landa, L. X. Qin, D. D. Rathbone, M. M. Condy, Y. Ustoyev, A. M. Crago, S. Singer, and G. K. Schwartz. Phase II trial of the CDK4 inhibitor PD0332991 in patients with advanced CDK4-amplified well-differentiated or dedifferentiated liposarcoma. *Journal of Clinical Oncology*, 31(16):2024–2028, 2013. ISSN 15277755. doi: 10.1200/JCO.2012.46.5476.

[52] P. Schöffski, D. Adkins, J. Y. Blay, T. Gil, A. D. Elias, P. Rutkowski, G. K. Pennock, H. Youssoufian, H. Gelderblom, R. Willey, and D. O. Grebennik. An open-label, phase 2 study evaluating the efficacy and safety of the anti-IGF-1R antibody cixutumumab in patients with previously treated advanced or metastatic soft-tissue sarcoma or Ewing family of tumours. *European Journal of Cancer*, 49(15):3219–3228, 2013. ISSN 09598049. doi: 10.1016/j.ejca.2013.06.010.

[53] J. Y. Blay, M. G. Leahy, B. B. Nguyen, S. R. Patel, P. Hohenberger, A. Santoro, A. P. Staddon, N. Penel, S. Piperno-Neumann, A. Hendifar, P. Lardelli, A. Nieto, V. Alfaro, and S. P. Chawla. Randomised phase III trial of trabectedin versus doxorubicin-based chemotherapy as first-line therapy in translocation-related sarcomas. *European Journal of Cancer*, 50(6):1137–1147, 2014. ISSN 18790852. doi: 10.1016/j.ejca.2014.01.012. URL http://dx.doi.org/10.1016/j.ejca.2014.01.012.

[54] H. Gelderblom, J. Y. Blay, B. M. Seddon, M. Leahy, I. Ray-Coquard, S. Sleijfer, J. M. Kerst, P. Rutkowski, S. Bauer, M. Ouali, S. Marréaud, R. J. H. M. Van Der Straaten, H. J. Guchelaar, S. D. Weitman, P. C.W. Hogendoorn, and P. Hohenberger. Brostallicin versus doxorubicin as first-line chemotherapy in patients with advanced or metastatic soft tissue sarcoma: An European Organisation for Research and Treatment of Cancer Soft Tissue and Bone Sarcoma Group randomised phase II and pharmacogeneti. *European Journal of Cancer*, 50(2):388–396, 2014. ISSN 09598049. doi: 10.1016/j.ejca.2013.10.002.

[55] I. Ray-Coquard, A. Le Cesne, J. S. Whelan, P. Schoffski, B. N. Bui, J. Verweij, S. Marréaud, M. van Glabbeke, P. Hogendoorn, and J. Y. Blay. A Phase II Study of Gefitinib for Patients with Advanced HER-1 Expressing Synovial Sarcoma Refractory to Doxorubicin-Containing Regimens. *The Oncologist*, 13(4): 467–473, 2008. ISSN 1083-7159. doi: 10.1634/theoncologist.2008-0065.

[56] W. T. A. van der Graaf, J. Y. Blay, S. P. Chawla, D. W. Kim, B. Bui-Nguyen, P. G. Casali, P. Schöffski, M. Aglietta, A. P. Staddon, Y. Beppu, A. Le Cesne, H. Gelderblom, I. R. Judson, N. Araki, M. Ouali, S. Marréaud, R. Hodge, M. R. Dewji, C. Coens, G. D. Demetri, C. D. Fletcher, A. P. Dei Tos, and P. Hohenberger. Pazopanib for metastatic soft-tissue sarcoma (PALETTE): A randomised, double-blind, placebo-controlled phase 3 trial. *The Lancet*, 379(9829):1879–1886, 2012. ISSN 1474547X. doi: 10.1016/S0140-6736(12)60651-5.

[57] P. F. Robbins, S. H. Kassim, T. L. N. Tran, J. S. Crystal, R. A. Morgan, S. A. Feldman, J. C. Yang, M. E. Dudley, J. R. Wunderlich, R. M. Sherry, U. S. Kammula, M. S. Hughes, N. P. Restifo, M. Raffeld, C. C. R. Lee, Y. F. Li, M. El-Gamil, and S. A. Rosenberg. A pilot trial using lymphocytes genetically engineered with an NY-ESO-1-reactive T-cell receptor: Long-term follow-up and correlates with response. *Clinical Cancer Research*, 21(5):1019–1027, 2015. ISSN 15573265. doi: 10.1158/1078-0432.CCR-14-2708.

[58] R. P. A'Hern. Sample size tables for exact single-stage phase II designs. *Statistics in Medicine*, 20(6): 859–866, 2001. ISSN 02776715. doi: 10.1002/sim.721.

[59] L. M. Ellis, D. S. Bernstein, E. E. Voest, J. D. Berlin, D. Sargent, P. Cortazar, E. Garrett-Mayer, R. S. Herbst, R. C. Lilenbaum, C. Sima, A. P. Venook, M. Gonen, R. L. Schilsky, N. J. Meropol, and L. E. Schnipper. American society of clinical oncology perspective: Raising the bar for clinical trials by defining clinically meaningful outcomes. *Journal of Clinical Oncology*, 32(12):1277–1280, apr 2014. ISSN 15277755. doi: 10.1200/JCO.2013.53.8009.

[60] M. Savina, S. Litière, A. Italiano, T. Burzykowski, F. Bonnetain, S. Gourgou, V. Rondeau, J. Y. Blay, S. Cousin, F. Duffaud, H. Gelderblom, A. Gronchi, I. Judson, A. Le Cesne, P. Lorigan, J. Maurel, W. van der Graaf, J. Verweij, S. Mathoulin-Pélissier, and C. Bellera. Surrogate endpoints in advanced sarcoma trials: A meta-analysis. *Oncotarget*, 9(77):34617–34627, 2018. ISSN 19492553. doi: 10.18632/oncotarget.26166.

[61] K. Tanaka, M. Kawano, T. Iwasaki, I. Itonaga, and H. Tsumura. Surrogacy of intermediate endpoints for overall survival in randomized controlled trials of first-line treatment for advanced soft tissue sarcoma in the pre- and post-pazopanib era: A meta-analytic evaluation. *BMC Cancer*, 19(1):1–9, 2019. ISSN 14712407. doi: 10.1186/s12885-019-5268-2.

[62] W. D. Tap, A. J. Wagner, P. Schöffski, J. Martin-Broto, A. Krarup-Hansen, K. N. Ganjoo, C. C. Yen, A. R. Abdul Razak, A. Spira, A. Kawai, A. Le Cesne, B. A. Van Tine, Y. Naito, S. H. Park, A. Fedenko, Z. Pápai, V. Soldatenkova, A. Shahir, G. Mo, J. Wright, and R. L. Jones. Effect of Doxorubicin Plus Olaratumab vs Doxorubicin Plus Placebo on Survival in Patients With Advanced Soft Tissue Sarcomas: The ANNOUNCE

Randomized Clinical Trial. *JAMA*, 323(13):1266–1276, apr 2020. ISSN 1538-3598. doi: 10.1001/JAMA.2020.1707. URL https://pubmed.ncbi.nlm.nih.gov/32259228/.

[63] R. Sanfilippo, R. L. Hayward, J. Musoro, C. Benson, M. G. Leahy, A. Brunello, J. Y. Blay, N. Steeghs, I. Desar, N. Ali, A. Hervieu, K. Thway, S. Marréaud, S. Litière, and B. Kasper.

[64] R. Simon. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*, 10(1):1–10, mar 1989. ISSN 0197-2456. doi: 10.1016/0197-2456(89)90015-9.

[65] M. G. P. Zuidgeest, I. Goetz, R. H. H. Groenwold, E. Irving, G. J. M. W. van Thiel, and D. E. Grobbee. Series: Pragmatic trials and real world evidence: Paper 1. Introduction. *Journal of Clinical Epidemiology*, 88:7–13, 2017. ISSN 18785921. doi: 10.1016/j.jclinepi.2016.12.023. URL http://dx.doi.org/10.1016/j.jclinepi.2016.12.023.

[66] M. Sbaraglia, E. Bellan, and A. P. Dei Tos. The 2020 WHO Classification of Soft Tissue Tumours: News and perspectives. *Pathologica*, 113(2):70–84, 2021. ISSN 1591951X. doi: 10.32074/1591-951X-213.

[67] I. Ray-Coquard, M. C. Montesco, J. M. Coindre, A. P. Dei tos, A. Lurkin, D. Ranchère-vince, A. Vecchiato, A. V. Decouvelaere, S. Mathoulin-pélissier, S. Albert, P. Cousin, D. Cellier, L. Toffolatti, C. R. Rossi, and J. Y. Blay. Sarcoma: Concordance between initial diagnosis and centralized expert review in a population-based study within three European regions. *Annals of Oncology*, 23(9):2442–2449, 2012. ISSN 15698041. doi: 10.1093/annonc/mdr610.

[68] K. Thway, J. Wang, T. Mubako, and C. Fisher. Histopathological diagnostic discrepancies in soft tissue tumours referred to a specialist centre: Reassessment in the era of ancillary molecular diagnosis. *Sarcoma*, 2014, 2014. ISSN 13691643. doi: 10.1155/2014/686902.

[69] X. H. Du, H. Wei, P. Zhang, W. T. Yao, and Q. Q. Cai. Heterogeneity of Soft Tissue Sarcomas and Its Implications in Targeted Therapy. *Frontiers in Oncology*, 10:1–3, 2020. ISSN 2234943X. doi: 10.3389/fonc.2020.564852. URL https://www.frontiersin.org/article/10.3389/fonc.2020.564852.

4

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

This chapter is based on: 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*, 2022:5815875, 2022.

Abstract

Background: Soft-tissue sarcomas (STS) constitute a rare group of heterogeneous mesenchymal tumours containing more than 100 histologic subtypes. Here, we investigate whether, and if so to what extent, skeletal metastases affect outcome of patients with advanced or metastatic disease.

Materials and methods: Selected patients participated in five clinical trials of EORTC - STBSG. Individuals were included if they started treatment with an active drug and had advanced/metastatic STS. The endpoints of interest were overall survival (OS) and progression-free survival (PFS). Univariate and multivariate pooled analyses (after correcting for 12 covariates) were employed with Kaplan-Meier and Cox regression to model the impact of bone metastasis at presentation per treatment line stratified by study. For the subset of patients with bone metastasis, the impact of another metastatic organ site was explored with multivariate Cox regression models.

Results: 565 out of 1034 (54.6%) patients received first-line systemic treatment for locally advanced or metastatic disease. Bone metastases were present in 140 patients (77 first-line, 63 second line or higher). The unadjusted difference in OS/PFS with or without bone metastasis was statistically significant only for first-line. For OS, the adjusted hazard ratios for bone metastasis presence were 1.33 (95%-CI:0.99-1.78) and 1.11 (95%-CI:0.81-1.52) for first-line/second line or later treated patients, respectively. Likewise, the adjusted hazard ratios for PFS were 1.31 (95%-CI:1.00-1.73) and 1.07 (95%-CI:0.80-1.43). Effects were not statistically significant, despite a trend in first-line patients for both endpoints. Subgroup analyses indicated bone and lymph node metastasis as the most detrimental combination for OS and bone and lung metastasis for PFS.

Conclusions: Adult STS patients receiving palliative systemic therapy with bone metastasis carried an overall

worse prognosis compared to STS patients without bone metastases. Skeletal metastasis was detrimental for both OS and PFS, independent of treatment line. Findings may have implications for the management of these patients.

4.1 Introduction

Soft-tissue sarcomas (STS) constitute a rare group of very heterogeneous mesenchymal tumours that include more than 100 histologic subtypes developed in supportive or connective tissue such as muscle, nerves, blood vessels, and fatty and fibrous tissues [1, 2]. They account for 1-2% of all newly diagnosed malignancies and commonly affect arms, legs, or trunk. Patients with advanced STS have a poor prognosis with a median progression-free survival (PFS) of about 6 months, i.e., for first-line systemic therapy with doxorubicin plus ifosfamide, and median overall survival (OS) of around 12 months.

Some of the most common adult STS histologies are leiomyosarcoma, undifferentiated pleomorphic sarcoma, and liposarcoma [3]. Chemotherapy is the most frequent systemic therapy for unresectable and advanced disease with mostly a palliative intent due to the high percentages of disease progression and mortality. Available treatment options include mostly chemotherapy, for instance, doxorubicin alone or in combination with ifosfamide for the first line and docetaxel plus gemcitabine for second-line or higher treated patients [4, 5]. Pazopanib, a targeted agent, is a treatment option for second-line or higher nonadipogenic STS [5]. The selection of treatment is based on clinical performance, age, histology, disease biology, patient preferences, and availability of novel treatments and studies [6, 7]. Gastrointestinal stromal tumours (GIST) are considered a separate entity since effective targeted treatment is available [8].

Depending on the histology, the majority of STS metastasise primarily to the lungs [9] and sometimes to the lymph nodes, bones, liver, and brain [10, 11]. 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) [12]. However, staging for bone metastases is not routine, at least for most STS, as their occurrence at presentation is generally rather low [6, 13]. A higher incidence rate is associated with some STS subtypes such as alveolar soft part sarcoma, myxoid liposarcoma, angiosarcoma, and rhabdomyosarcoma [12, 14, 15].

In this article, our aim was to investigate (i) whether bone metastasis at presentation is prognostic for OS or PFS of advanced/metastatic STS patients and (ii) which metastatic organ site has the largest impact for patient's OS/PFS combined with bone metastasis at diagnosis in this database.

4.2 Materials and methods

4.2.1 Patients

Patients with advanced or metastatic non-GIST STS from five prospective clinical trials of the EORTC - STBSG database were included (enrolment period from April 2003 to June 2015) [16–20]. These studies assessed the following drugs/drug combinations: eribulin [17], pazopanib [16, 18], doxorubicin plus ifosfamide versus doxorubicin [19], or trabectedin versus doxorubicin [20] – for either first-line or second-line or later treated locally advanced or metastatic population. The intended treatment arm was used instead of the administered treatment arm as the latter included four missing values and variables were almost identical. Bone lesions were typically detected as part of computed tomography (CT) scans to measure tumour lesions at baseline. Details on eligibility criteria and outcomes have been published [16–20].

Individuals were included if they were eligible in their respective trial, started their allocated treatment with an active drug component, and had locally advanced or metastatic STS at study/observational entry. On the other hand, patients who had a performance status of 2 or worse or were diagnosed with GIST were excluded from all analyses. Data on three patients were not available for both OS and PFS. The PRISMA flow diagram is provided in Figure 4.1 [21].

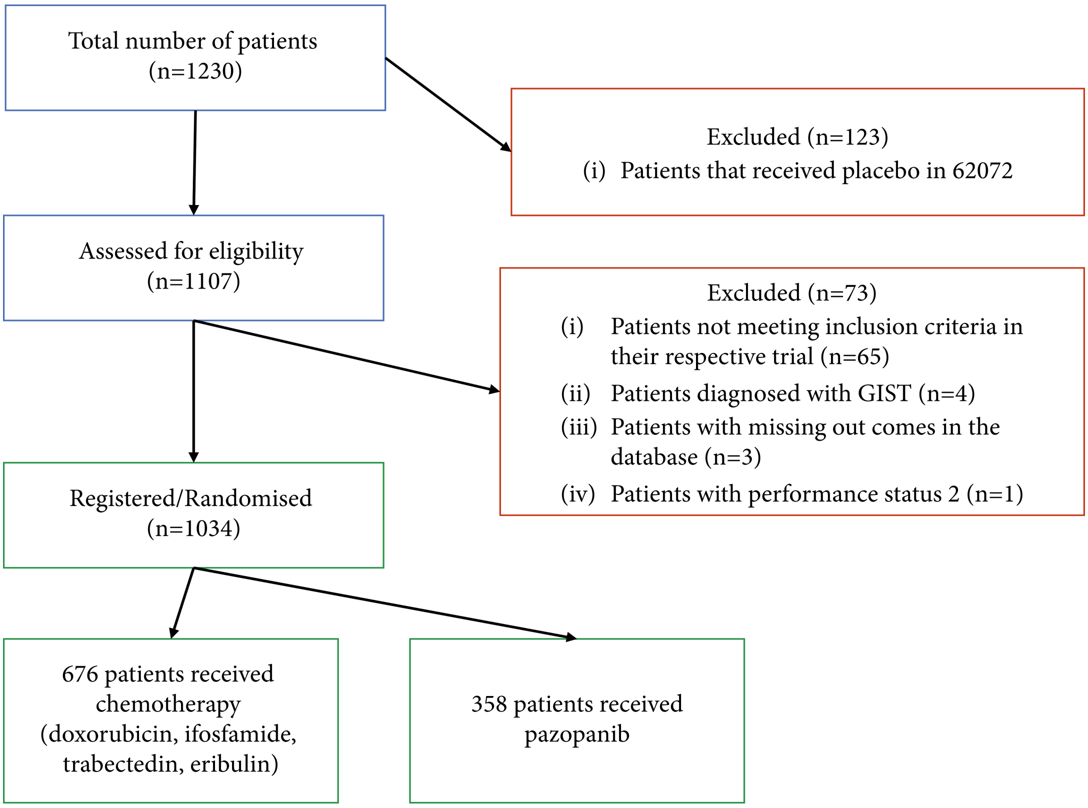

Total number of patients
(n=1230)
Excluded (n=123)
(i) Patients that received placebo in 62072
Assessed for eligibility
(n=1107)
Excluded (n=73)
(i) Patients not meeting inclusion criteria in their respective trial (n=65)
(ii) Patients diagnosed with GIST (n=4)
(iii) Patients with missing out comes in the database (n=3)
(iv) Patients with performance status 2 (n=1)
Registered/Randomised
(n=1034)
676 patients received chemotherapy (doxorubicin, ifosfamide, trabectedin, eribulin)
358 patients received pazopanib

Figure 4.1: **PRISMA flow diagram [21].**

In total, 1034 patients were used to investigate the prognostic significance of bone metastasis at study initiation (presence versus absence). A subgroup of 140 patients was analysed to explore the most prognostic metastatic organ site in the concurrent event of bone metastasis.

4.2.2 Endpoints

The endpoints for this analysis were OS and PFS. OS time was estimated from the date of registration/randomisation (according to the study-specific protocol) until the date of death from any cause. Patients still alive were censored on their last follow-up date. PFS time was estimated from the date of registration/randomisation until the date of disease progression or death from any cause. If neither progression nor death was observed, patients were censored on their last date known to be alive.

4.2.3 Adjusted covariates

Demographic factors at registration/randomisation, line of treatment, histology type, tumour grade, site of the primary tumour, time between histological diagnosis and registration/randomisation, and location of metastasis (bone, liver, lymph nodes, lung, soft tissue (primary or other soft-tissue invasive), or other sites) were considered for the analysis. Soft tissue includes fat, muscles, blood vessels, nerves, tendons, and tissues that surround the bones and joints. Metastasis in other sites referred to ascites, pleural effusion, skin, or other invasive diseases.

Demographic factors included gender, age (less than 40, 40–50, 50–70, or more than 70), and performance status

(0 or 1). The line of treatment was 1 (first-line) or 2+ (second-line or higher). The histological entities of STS were aggregated in five commonly occurring groups: angiosarcoma, leiomyosarcoma, liposarcoma (all subtypes), synovial sarcoma, and an additional group for the remaining STS types (other STS). Diagnosis by local pathologists was used as the central review was incomplete for several patients, which could lead to a substantial loss of data.

The tumour grade was dichotomised as low/intermediate or high. Patients whose tumours were initially diagnosed as low grade were only entered in their specific studies in case of rapid progression before first-line systemic therapy as such clinical behaviour is consistent with a higher-grade tumour rather than a low grade. The site of primary tumour was classified into five locations: extremities, abdomen, thorax, visceral, or other (e.g., primary lung is under the thorax; the thigh is under extremities). The time between histological diagnosis and registration/randomisation was categorized as less than a year, 1–2 years, or more than 2 years.

The six covariates for metastasis in bone, liver, lymph nodes, lung, soft tissue (primary or other soft-tissue invasive), or other sites were defined as absent versus present.

4.2.4 Statistical methods

Covariates were summarised by frequencies and percentages. Univariate analyses were performed for the effect of bone metastasis on OS and PFS per treatment line stratified by study (to account for the variation between the clinical trials) with the Kaplan–Meier method. The log-rank test was used to assess the difference in survival [22]. Moreover, the unadjusted/adjusted effect of bone metastasis on OS and PFS was estimated per treatment line (first-line versus second-line or later) stratified by study with univariate/multivariate Cox regression models, including additionally the baseline variables described in the previous section [23]. For the subset of patients with bone metastasis at presentation, the impact of another metastatic organ site was explored per treatment line with multivariate Cox regression models for PFS and OS stratified by study and adjusted for prognostic variables. The most detrimental combinations are presented.

Outcomes were reported as hazard ratios with 95% confidence intervals (95% CIs). Statistical analyses were performed in SAS software version 9.4 (SAS Institute, Cary NC). All reported values were 2-sided at a 5% significance level.

4.3 Results

4.3.1 Median follow-Up times

The median overall follow-up for all patients was 3 years (IQR: 2.2–4.6) estimated with the reverse Kaplan–Meier method [22]. Patients with bone metastasis were followed for up to 5.5 years, whereas those without bone metastasis for a maximum period of 8.0 years. The median overall follow-up time was 2.8 years (IQR: 2.0–5.6) and 3.0 years (IQR: 2.2–4.6) for patients with and without bone metastasis, respectively. Looking at the survivors' group only, the median follow-up time was 2.0 years (IQR: 1.5–2.4) for bone metastasis presence and 2.2 years (IQR: 1.6–3.2) for bone metastasis absence.

4.3.2 Patient characteristics

In Table 4.1, baseline characteristics for patients with and without metastasis in the bone are shown. Percentages are similar between the two groups. Systemic therapy was given in first-line to 565 patients (54.6%) with metastatic or locally advanced STS. The majority of the patients (676, 65.4%) received chemotherapy (doxorubicin, doxorubicin plus ifosfamide, eribulin, or trabectedin). Pazopanib (n = 358, 34.6%) was the most frequent treatment arm for patients being treated in second-line or higher for metastatic or locally advanced disease. The intended treatment arm and line of treatment in the five EORTC studies are provided in Supplementary materials Table S1.

The metastatic profile of the patients versus metastasis in bone is shown in Table 2. Bone metastases were present in 140 patients (13.5%); 226 patients (21.9%) had liver metastases, 250 patients (24.2%) had lymph node metastases, and 290 patients (28.0%) had metastases in any other site. On the other hand, pulmonary and soft-tissue metastases (locoregional or other soft-tissue invasive) were present in 719 (69.5%) and 556 patients (53.8%), respectively. Soft tissue metastasis per histology type is provided in Supplementary materials Table S2.

4.3.3 Prognostic significance of bone metastasis for OS

From the 894 patients without bone metastasis, 488 (54.6%) were in first-line therapy and 406 (45.4%) were treated in second-line or higher systemic treatment for locally advanced or metastatic disease. Amongst the 140 patients with bone metastasis presence at study entry, 77 (55.0%) were treated in the first line and 63 (45.0%) were treated in the second line or later.

The median first-line OS for patients with or without bone metastases was 0.9 years (95%-CI: 0.7–1.1) and 1.3 years (95%-CI: 1.1–1.4), respectively. For patients treated with second-line or higher systemic treatment, median OS was 0.9 (95%-CI: 0.7–1.2) and 1.0 (95%-CI: 0.9–1.1), respectively. The unadjusted difference in OS for patients with or without metastasis in the bone was statistically significant for first-line ($p < 0.01$) but not for second-line or later systemic treatment ($p = 0.53$). Kaplan–Meier curves are presented in Figure 4.2, including estimates at 1, 2, and 3 years along with their 95%-CIs. The estimated hazard ratio for the presence of bone metastasis was 1.55 (95%-CI: 1.19–2.01) for first-line therapy and 1.10 (95%-CI: 0.81–1.49) for second or further lines of therapy. This means that the presence of bone metastasis increased the hazard of dying by 55.0% for first line. There was no evidence of interaction between bone metastasis and treatment line ($p = 0.13$, Figure S1). The adjusted effect of bone metastasis on OS based on multivariate analysis is provided in Table 4.3. The effect was not statistically significant for any line of treatment ($p > 0.05$)—despite a trend for first-line treatment. The adjusted hazard ratio for bone metastasis presence in first-line systemic treatment was reduced to 1.33 (95% CI: 0.99–1.78). For the population of second line or higher, the adjusted hazard ratio was 1.11 (95% CI: 0.81–1.52).

4.3.4 Prognostic significance of bone metastasis for PFS

The median PFS for patients treated in first line was 4.2 (95% CI: 2.3–5.5) and 6.1 months (95% CI: 5.2–6.7) for bone metastasis presence or absence, respectively. For patients treated in second line or higher, the median PFS was 3.0 (95% CI: 2.7–4.6) and 3.3 (95% CI: 2.8–4.0), respectively. The unadjusted difference in PFS for patients with or without metastasis in the bone was statistically significant for first line ($p < 0.01$) but not for second or further lines ($p = 0.69$). The corresponding estimated survival curves are presented in Figure 4.3, including estimates at 3, 6, and 12 months. The unadjusted estimated hazard ratio for the presence of bone metastasis was 1.43 (95% CI: 1.12–1.84) for first line and 1.06 (95% CI: 0.80–1.40) for the population treated in second line or higher for locally advanced or metastatic disease. This means that the presence of bone metastasis increased the hazard of progression or death by 43.0% for the first line. However, there was no evidence of interaction between bone metastasis and treatment line ($p = 0.09$, Figure S2). Table 4.4 provides the adjusted effect of bone metastasis on PFS based on the multivariate analysis. The effect was not statistically significant for first or further lines ($p > 0.05$)—despite a trend for first line. The adjusted hazard ratio for bone metastasis presence in first line was 1.31 (95%-CI: 1.00–1.73). For population in the second line or higher, the adjusted hazard ratio was 1.07 (95%-CI: 0.80–1.43).

	Metastasis in bone		
	Absent (N=894)	Present (N=140)	Total (N=1034)
	N (%)	N (%)	N (%)
Intended treatment			
Doxorubicin 75 mg/m^2	229 (25.6)	32 (22.9)	261 (25.2)
Doxorubicin 75 mg/m^2/ifosfamide 10g/m^2	187 (20.9)	33 (23.6)	220 (21.3)
Pazopanib 800 mg once daily	309 (34.6)	49 (35.0)	358 (34.6)
Trabectedin 1.3 mg/m^2, 3hrs IV	38 (4.3)	6 (4.3)	44 (4.3)
Trabectedin 1.5 mg/m^2, 24 hrs IV	33 (3.7)	7 (5.0)	40 (3.9)
Eribulin 1.4 mg/m^2 every 3 weeks	98 (11.0)	13 (9.3)	111 (10.7)
Age category			
Less than 40	194 (21.7)	31 (22.1)	225 (21.8)
40 to 50	201 (22.5)	42 (30.0)	243 (23.5)
50 to 70	428 (47.9)	60 (42.9)	488 (47.2)
Older than 70	71 (7.9)	7 (5.0)	78 (7.5)
Gender			
Male	411 (46.0)	61 (43.6)	472 (45.6)
Female	483 (54.0)	79 (56.4)	562 (54.4)
Tumour grade			
Low/intermediate	411 (46.0)	55 (39.3)	466 (45.1)
High	433 (48.4)	79 (56.4)	512 (49.5)
Missing	50 (5.6)	6 (4.3)	56 (5.4)
Line of treatment			
1^{st}	488 (54.6)	77 (55.0)	565 (54.6)
2+	406 (45.4)	63 (45.0)	469 (45.4)
Site of primary tumour			
Thorax	104 (11.6)	16 (11.4)	120 (11.6)
Abdomen	171 (19.1)	20 (14.3)	191 (18.5)
Extremities	330 (36.9)	54 (38.6)	384 (37.1)
Visceral	194 (21.7)	29 (20.7)	223 (21.6)
Others	94 (10.5)	20 (14.3)	114 (11.0)
Missing	1 (0.1)	1 (0.7)	2 (0.2)
Histology type			
Liposarcoma (all subtypes)	92 (10.3)	8 (5.7)	100 (9.7)
Leiomyosarcoma	278 (31.1)	46 (32.9)	324 (31.3)
Angiosarcoma	19 (2.1)	10 (7.1)	29 (2.8)
Synovial sarcoma	129 (14.4)	13 (9.3)	142 (13.7)
Others	376 (42.1)	63 (45.0)	439 (42.5)
Time between histological diagnosis and registration/randomisation			
Less than a year	377 (42.2)	60 (42.9)	437 (42.3)
1-2 years	174 (19.5)	29 (20.7)	203 (19.6)
More than 2 years	343 (38.4)	51 (36.4)	394 (38.1)
Performance status			
0	494 (55.3)	63 (45.0)	557 (53.9)
1	400 (44.7)	77 (55.0)	477 (46.1)

Table 4.1: **Patient baseline characteristics versus metastatic profile in the bone**. Treatment regimens in first-line setting: doxorubicin, doxorubicin+ifosfamide, and trabectedin. Treatments regimens for 2^{nd}-line or later treated patients: pazopanib and eribulin.

	Metastasis in bone		
	Absent (N=894)	Present (N=140)	Total (N=1034)
	N (%)	N (%)	N (%)
Metastasis in soft-tissue			
(primary or other soft-tissue invasive)			
Absent	411 (46.0)	67 (47.9)	478 (46.2)
Present	483 (54.0)	73 (52.1)	556 (53.8)
Metastasis in liver			
Absent	713 (79.8)	95 (67.9)	808 (78.1)
Present	181 (20.2)	45 (32.1)	226 (21.9)
Metastasis in lymph nodes			
Absent	693 (77.5)	91 (65.0)	784 (75.8)
Present	201 (22.5)	49 (35.0)	250 (24.2)
Metastasis in lung			
Absent	292 (32.7)	23 (16.4)	315 (30.5)
Present	602 (67.3)	117 (83.6)	719 (69.5)
Metastasis in other sites			
(ascites, pleural effusion, skin, or other invasive)			
Absent	649 (72.6)	95 (67.9)	744 (72.0)
Present	245 (27.4)	45 (32.1)	290 (28.0)

Table 4.2: **Patient metastatic profile versus metastasis in the bone**.

4.3.5 Prognosis for each metastatic organ site combined with bone metastasis for OS

In our database, 140 patients (13.5%) had bone metastasis and 6 patients had exclusively bone metastasis at presentation. 77 (55.0%) were first-line and 63 (45.0%) second-line or later treated patients. Kaplan–Meier curves for the number of other metastatic organ sites involved together with bone metastasis for OS are depicted in Figure S3 per treatment line. When bone metastasis was present, the number of metastatic organ sites did not seem to affect OS in a proportional manner.

The hazard ratios for the effect of bone metastasis combined with other metastases are presented in Table 4.5. Bone and lymph node metastasis presence were the most adverse for first-line with a hazard ratio equal to 2.97 (95%-CI: 1.53–5.78). For second-line and higher patients, the combination of bone and lymph nodes seemed to be the most detrimental increasing the risk of death by 59%, although not statistically significant (p = 0.39).

4.3.6 Prognosis for each metastatic organ site combined with bone metastasis for PFS

Kaplan–Meier curves for the number of other metastatic organ sites involved with bone metastasis for PFS are provided in Figure S4 per treatment line of systemic treatment. Findings were similar to OS.

The hazard ratios for the combined metastatic profile in the bone and other sites are shown in Table 4.6. The most detrimental combination was bone and lung metastasis, which increased the hazard of progression or death by 180% in first-line treatment (p = 0.03) and 145% in second-line or further lines treatment (p = 0.21).

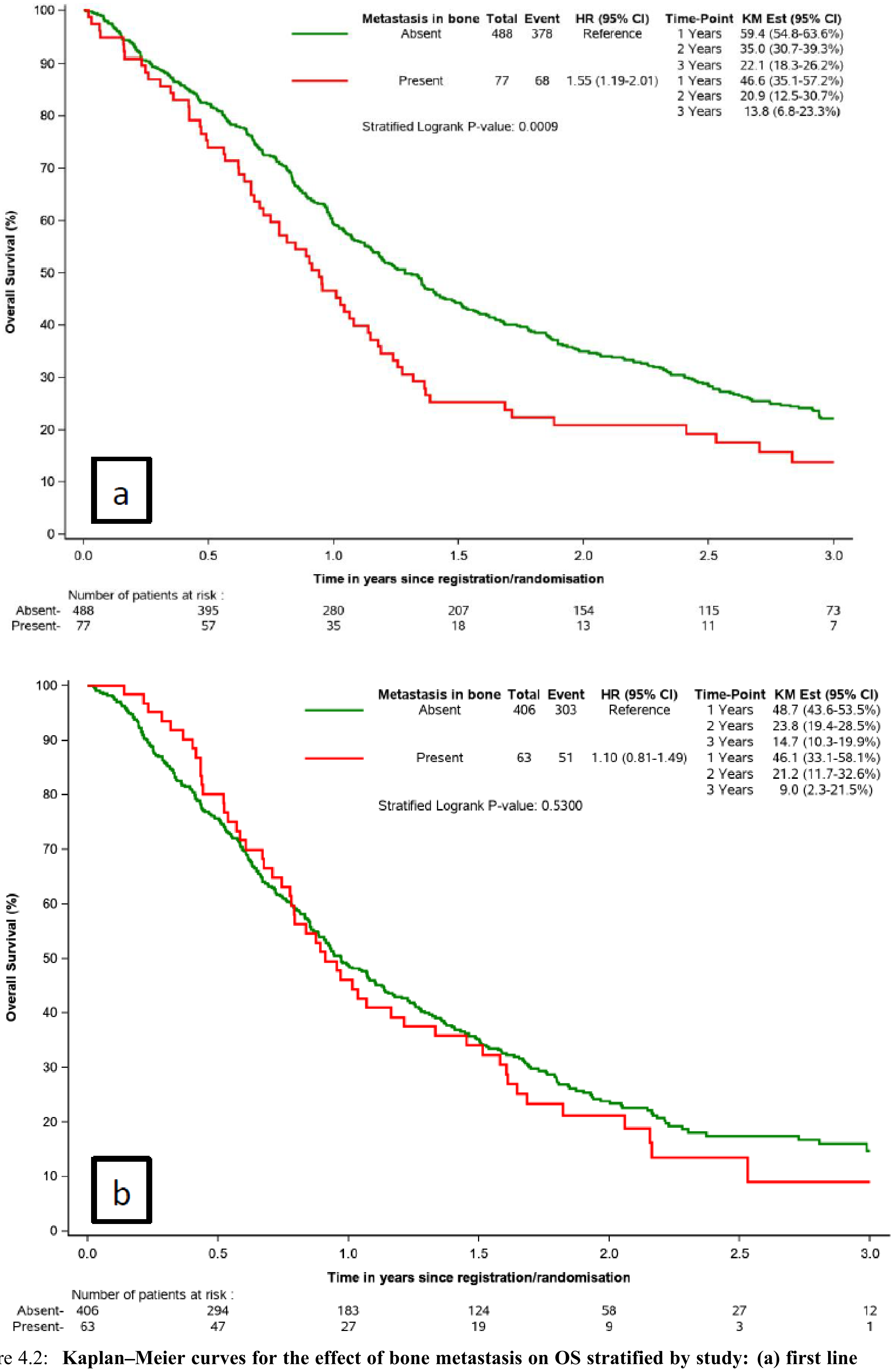
Parameter	**Levels**	**Hazard ratio for first-line treatment** (95% CI)	**p value (first-line)**	**Hazard ratio for second-line or higher treatment** (95% CI)	**p value (second line or higher)**
Metastasis in bone	Absent	1.00	0.061	1.00	0.533
	Present	1.33 (0.99, 1.78)		1.11 (0.81, 1.52)	
Histology type (local review)	Angiosarcoma	1.00		1.00	
	Leiomyosarcoma	0.68 (0.39, 1.18)		1.03 (0.40, 2.65)	
	Liposarcoma (all subtypes)	0.65 (0.35, 1.20)		3.05 (1.10, 8.48)	
	Others	0.89 (0.52, 1.54)		2.00 (0.79, 5.04)	
	Synovial sarcoma	0.95 (0.52, 1.73)		2.35 (0.91, 6.04)	
Age category	Less than 40	1.00		1.00	
	40 to 50	1.05 (0.79, 1.41)		1.25 (0.86, 1.81)	
	50 to 70	1.24 (0.93, 1.66)		1.42 (1.05, 1.91)	
	Older than 70	1.25 (0.61, 2.55)		1.88 (1.21, 2.93)	
Gender	Male	1.00		1.00	
	Female	0.95 (0.77, 1.17)		0.89 (0.70, 1.13)	
Tumour grade	Low/intermediate	1.00		1.00	
	High	1.38 (1.11, 1.71)		1.39 (1.11, 1.75)	
Metastasis in liver	Absent	1.00		1.00	
	Present	1.47 (1.11, 1.95)		1.04 (0.79, 1.36)	
Metastasis in lymph nodes	Absent	1.00		1.00	
	Present	1.25 (0.99, 1.56)		1.34 (1.02, 1.77)	
Metastasis in lung	Absent	1.00		1.00	
	Present	1.25 (0.98, 1.59)		1.14 (0.88, 1.49)	
Metastasis in other sites	Absent	1.00		1.00	
	Present	1.53 (1.20, 1.95)		1.36 (1.07, 1.73)	
Metastasis in soft-tissue (primary or other soft-tissue invasive)	Absent	1.00		1.00	
	Present	1.09 (0.86, 1.38)		1.25 (0.98, 1.60)	
Site of primary tumour	Extremities	1.00		1.00	
	Abdomen	1.27 (0.90, 1.78)		1.10 (0.78, 1.54)	
	Others	1.01 (0.73, 1.39)		1.08 (0.72, 1.62)	
	Thorax	1.00 (0.68, 1.47)		1.29 (0.90, 1.85)	
	Visceral	1.21 (0.88, 1.65)		1.25 (0.89, 1.77)	
Performance status	0	1.00		1.00	
	1	1.67 (1.36, 2.05)		1.70 (1.34, 2.17)	
Time between histological diagnosis and registration/randomisation	Less than a year	1.00		1.00	
	1-2 years	0.78 (0.58, 1.06)		0.94 (0.68, 1.30)	
	More than 2 years	0.61 (0.47, 0.79)		0.60 (0.44, 0.80)	

Table 4.3: **Cox model for the adjusted effect of bone metastasis on OS per line of treatment stratified by study**. Note that 977 (94.5%) of the patients were analysed due to some missing values.

4.4 Discussion

In this research project, we analysed the prognostic impact of bone metastasis at study inclusion for OS and PFS of locally advanced/metastatic STS, separately for first-line and second-line or higher treated patients, with a pooled analysis of five clinical trials from the EORTC - STBSG database. For the subgroup of patients with bone metastasis (n = 140, 13.5%), the most impactful metastatic combination was identified between bone and another site (liver, lymph node, lung, other) for OS and PFS.

There is an increased prevalence of bone metastasis in advanced-stage cancers [24]. The highest incidence can be found in breast, prostate, and lung malignancies [25]. Although bone metastasis is an independent negative prognostic factor with clinical implications for survival and quality of life of patients, a longer survival duration has been observed for breast and prostate cancers with bone metastases, which are hormone-sensitive (median OS 15–27 months) [24–26]. On the other hand, patients with gastrointestinal (GI), lung, and gynaecological cancers



Metastasis in bone	Total	Event	HR (95% CI)	Time-Point	KM Est (95% CI)
Absent	488	378	Reference	1 Years	59.4 (54.8-63.6%)
				2 Years	35.0 (30.7-39.3%)
				3 Years	22.1 (18.3-26.2%)
Present	77	68	1.55 (1.19-2.01)	1 Years	46.6 (35.1-57.2%)
				2 Years	20.9 (12.5-30.7%)
				3 Years	13.8 (6.8-23.3%)

Stratified Logrank P-value: 0.0009

Number of patients at risk :

	0.0	0.5	1.0	1.5	2.0	2.5	3.0
Absent-	488	395	280	207	154	115	73
Present-	77	57	35	18	13	11	7

Metastasis in bone	Total	Event	HR (95% CI)	Time-Point	KM Est (95% CI)
Absent	406	303	Reference	1 Years	48.7 (43.6-53.5%)
				2 Years	23.8 (19.4-28.5%)
				3 Years	14.7 (10.3-19.9%)
Present	63	51	1.10 (0.81-1.49)	1 Years	46.1 (33.1-58.1%)
				2 Years	21.2 (11.7-32.6%)
				3 Years	9.0 (2.3-21.5%)

Stratified Logrank P-value: 0.5300

Number of patients at risk :

	0.0	0.5	1.0	1.5	2.0	2.5	3.0
Absent-	406	294	183	124	58	27	12
Present-	63	47	27	19	9	3	1

Figure 4.2: **Kaplan–Meier curves for the effect of bone metastasis on OS stratified by study: (a) first line and (b) second line or higher population**.

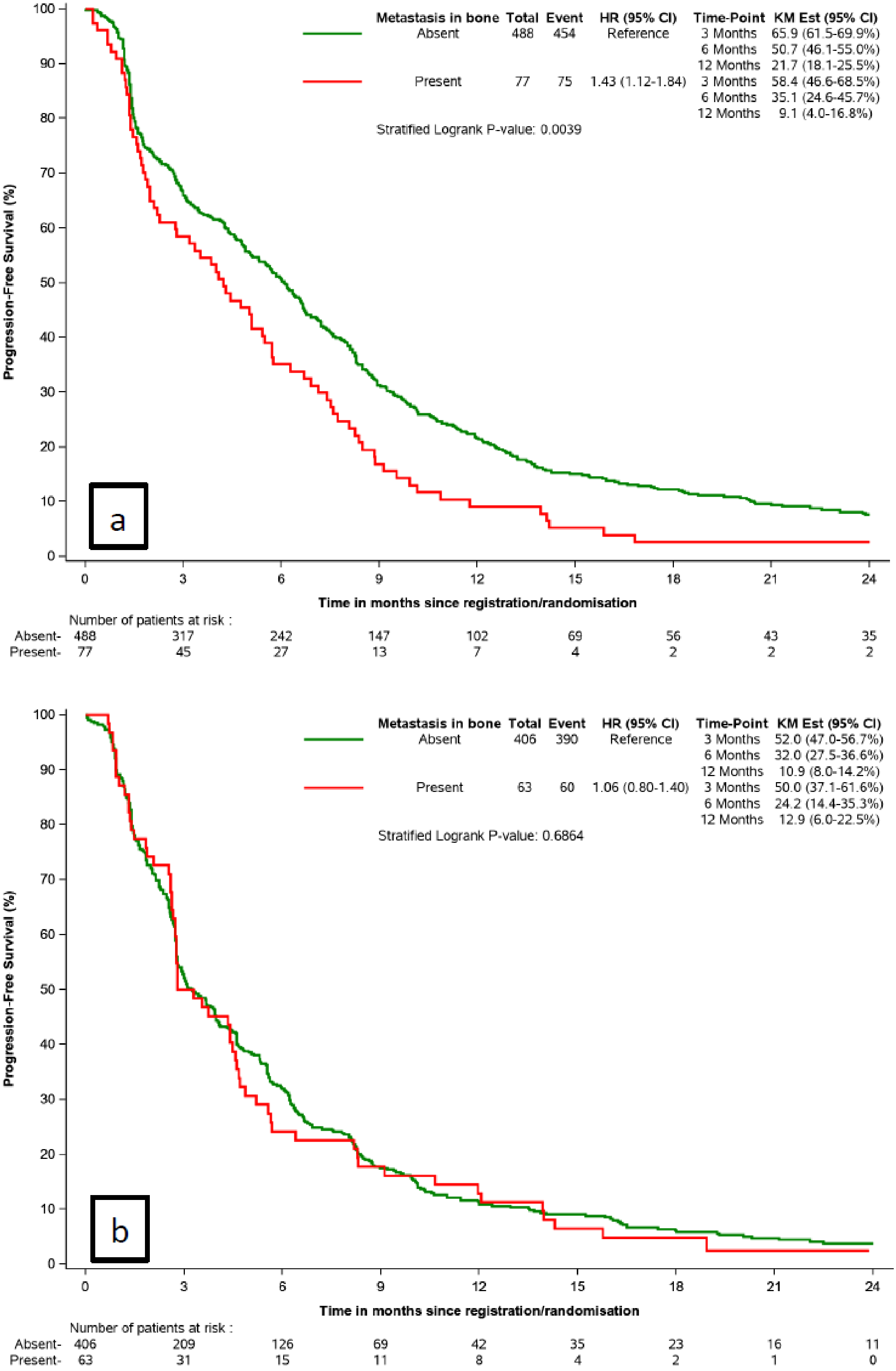
Parameter	**Levels**	**Hazard ratio for first-line treatment** (95% CI)	**p value (first-line)**	**Hazard ratio for second-line or higher treatment** (95% CI)	**p value (second line or higher)**
Metastasis in bone	Absent	1.00	0.061	1.00	0.533
	Present	1.31 (1.00, 1.73)		1.07 (0.80, 1.43)	
Histology type (local review)	Angiosarcoma	1.00		1.00	
	Leiomyosarcoma	1.21 (0.72, 2.04)		0.75 (0.34, 1.67)	
	Liposarcoma (all subtypes)	1.16 (0.66, 2.05)		2.89 (1.15, 7.23)	
	Others	1.33 (0.79, 2.22)		0.81 (0.37, 1.80)	
	Synovial sarcoma	1.39 (0.80, 2.43)		0.85 (0.37, 1.91)	
Age category	Less than 40	1.00		1.00	
	40 to 50	1.10 (0.84, 1.45)		1.00 (0.72, 1.40)	
	50 to 70	0.97 (0.74, 1.27)		1.24 (0.96, 1.61)	
	Older than 70	1.00 (0.53, 1.91)		1.18 (0.80, 1.75)	
Gender	Male	1.00		1.00	
	Female	1.05 (0.86, 1.27)		0.83 (0.67, 1.03)	
Tumour grade	Low/intermediate	1.00		1.00	
	High	1.26 (1.04, 1.53)		1.33 (1.08, 1.63)	
Metastasis in liver	Absent	1.00		1.00	
	Present	1.32 (1.02, 1.72)		1.09 (0.85, 1.39)	
Metastasis in lymph nodes	Absent	1.00		1.00	
	Present	1.13 (0.92, 1.39)		1.18 (0.91, 1.53)	
Metastasis in lung	Absent	1.00		1.00	
	Present	1.32 (1.06, 1.66)		0.93 (0.74, 1.18)	
Metastasis in other sites	Absent	1.00		1.00	
	Present	1.42 (1.13, 1.78)		1.36 (1.09, 1.69)	
Metastasis in soft-tissue (primary or other soft-tissue invasive)	Absent	1.00		1.00	
	Present	1.06 (0.85, 1.33)		1.09 (0.87, 1.36)	
Site of primary tumour	Extremities	1.00		1.00	
	Abdomen	1.17 (0.86, 1.60)		0.85 (0.62, 1.15)	
	Others	1.01 (0.75, 1.36)		0.99 (0.69, 1.44)	
	Thorax	1.33 (0.95, 1.87)		1.11 (0.80, 1.55)	
	Visceral	1.21 (0.90, 1.63)		1.24 (0.92, 1.68)	
Performance status	0	1.00		1.00	
	1	1.27 (1.05, 1.54)		1.10 (0.89, 1.36)	
Time between histological diagnosis and registration/randomisation	Less than a year	1.00		1.00	
	1-2 years	0.93 (0.70, 1.23)		1.09 (0.81, 1.47)	
	More than 2 years	0.77 (0.61, 0.97)		0.74 (0.56, 0.96)	

Table 4.4: **Cox model for the adjusted effect of bone metastasis on PFS per line of treatment stratified by study**. Note that 977 (94.5%) of the patients were analysed due to some missing values.

usually have the lowest survival in case of bone metastasis (median OS < 12 months). A larger tumour burden is associated with a worse OS.

In general, patients who present with metastatic STS have a poor prognosis regardless of the systemic treatment they receive [27, 28]. A pooled analysis of metastatic STS patients (lesions in lung, liver, bone, or other site)—who received first-line chemotherapy in fifteen EORTC trials—suggested that lung involvement only was an independent prognostic factor in favour of OS in contrast with other metastatic sites [29]. An improved median survival time has also been observed in other studies with isolated lung versus bone metastasis [10, 30].

Ferguson et al. (2006) investigated histologic bone invasion in extremity STS at a reference sarcoma center between 1986 and 2001 based on magnetic resonance imaging (MRI) [31]. In total, 48/874 patients had evidence of bone invasion at presentation. Interestingly, these patients presented with a significantly higher proportion of larger and deeper tumours. They found that bone invasion was a precursor of a poor OS and was associated with a more aggressive clinical course. Younis et al. (2020) used the Surveillance, Epidemiology and End Results (SEER)



(a)

Metastasis in bone	Total	Event	HR (95% CI)	Time-Point	KM Est (95% CI)
🟩 Absent	488	454	Reference	3 Months	65.9 (61.5-69.9%)
				6 Months	50.7 (46.1-55.0%)
				12 Months	21.7 (18.1-25.5%)
🟥 Present	77	75	1.43 (1.12-1.84)	3 Months	58.4 (46.6-68.5%)
				6 Months	35.1 (24.6-45.7%)
				12 Months	9.1 (4.0-16.8%)

Stratified Logrank P-value: 0.0039

Number of patients at risk :

	0	3	6	9	12	15	18	21	24
Absent-	488	317	242	147	102	69	56	43	35
Present-	77	45	27	13	7	4	2	2	2

(b)

Metastasis in bone	Total	Event	HR (95% CI)	Time-Point	KM Est (95% CI)
🟩 Absent	406	390	Reference	3 Months	52.0 (47.0-56.7%)
				6 Months	32.0 (27.5-36.6%)
				12 Months	10.9 (8.0-14.2%)
🟥 Present	63	60	1.06 (0.80-1.40)	3 Months	50.0 (37.1-61.6%)
				6 Months	24.2 (14.4-35.3%)
				12 Months	12.9 (6.0-22.5%)

Stratified Logrank P-value: 0.6864

Number of patients at risk :

	0	3	6	9	12	15	18	21	24
Absent-	406	209	126	69	42	35	23	16	11
Present-	63	31	15	11	8	4	2	1	0

Figure 4.3: **Kaplan–Meier curves for the effect of bone metastasis on PFS stratified by study: (a) first line and (b) second line or higher population**.

Parameter	**Levels**	**Hazard ratio for first-line treatment** (95% CI)	**p value (first-line)**	**Hazard ratio for second-line or higher treatment** (95% CI)	**p value (second line or higher)**
Bone and liver metastases	Bone present—liver absent	1.00	0.114	1.00	0.436
	Bone present—liver present	1.83 (0.86, 3.90)		1.56 (0.51, 4.79)	
Bone and lymph node metastases	Bone present—lymph nodes absent	1.00	0.001	1.00	0.389
	Bone present—lymph nodes present	2.97 (1.53, 5.78)		1.59 (0.55, 4.54)	
Bone and lung metastases	Bone present—lung absent	1.00	0.881	1.00	0.915
	Bone present—lung present	0.93 (0.37, 2.36)		0.92 (0.19, 4.47)	
Bone and and soft-tissue metastases	Bone present—soft-tissue absent	1.00	0.278	1.00	0.679
	Bone present—soft-tissue present	0.65 (0.30, 1.42)		0.83 (0.34, 2.02)	
Bone and other metastases	Bone present—other absent	1.00	0.044	1.00	0.497
	Bone present—other present	0.45 (0.21, 0.98)		0.68 (0.23, 2.04)	

Table 4.5: **Cox model for the effect of bone metastasis combined with another metastatic organ site on OS stratified by study**. Hazard ratios were adjusted for demographic characteristics, histological entity, tumour grade, site of primary tumour, and time between histological diagnosis and registration/randomisation.

Parameter	**Levels**	**Hazard ratio for first-line treatment** (95% CI)	**p value (first-line)**	**Hazard ratio for second-line or higher treatment** (95% CI)	**p value (second line or higher)**
Bone and liver metastases	Bone present—liver absent	1.00	0.185	1.00	0.149
	Bone present—liver present	1.68 (0.78, 3.63)		2.16 (0.76, 6.15)	
Bone and lymph node metastases	Bone present—lymph nodes absent	1.00	0.040	1.00	0.742
	Bone present—lymph nodes present	1.99 (1.03, 3.85)		0.84 (0.30, 2.34)	
Bone and lung metastases	Bone present—lung absent	1.00	0.030	1.00	0.205
	Bone present—lung present	2.80 (1.10, 7.09)		2.45 (0.61, 9.84)	
Bone and and soft-tissue metastases	Bone present—soft-tissue absent	1.00	0.299	1.00	0.175
	Bone present—soft-tissue present	0.68 (0.33, 1.40)		1.94 (0.75, 5.04)	
Bone and other metastases	Bone present—other absent	1.00	0.112	1.00	0.082
	Bone present—other present	0.56 (0.27, 1.15)		0.41 (0.15, 1.12)	

Table 4.6: **Cox model for the effect of bone metastasis combined with another metastatic organ site on PFS stratified by study**. Hazard ratios were adjusted for demographic characteristics, histological entity, tumour grade, site of primary tumour, and time between histological diagnosis and registration/randomisation.

registry to identify risk factors for early bone metastasis and prognostic factors of survival in 180 extremities of STS patients with skeletal metastasis from 2010 to 2015 [30]. The authors concluded that high tumour grade, deep location to fascia, and regional lymph node metastasis were significant risk factors at diagnosis. Resection of the primary sarcoma was the only significant predictor of survival in the presence of bone metastasis.

A metastatic bone profile may be part of STS patients' natural history, which negatively affects their prognosis. Here, patients with STS of the extremities, abdomen, thorax, visceral or other sites of primary tumour were included. A higher incidence rate of bone metastasis, amongst the four main sarcoma subtypes, was detected for angiosarcoma (10/29, 34.5%) and leiomyosarcoma (46/324, 14.2%), which matches previously reported findings [12, 14]. According to our pooled analysis, the unadjusted difference in OS/PFS for patients with or without bone metastasis was statistically significant for first-line treatment. However, this difference was not significant when adjusting for other prognostic factors. Nevertheless, an overall worse status is suggested for patients suffering from bone metastasis.

A strength of this work is the large patient cohort combined with a minimal amount of missing data. From 1034 patients included here, the tumour grade was missing for 56 patients (5.4%) and the site of primary tumour for two patients (0.2%). The 12 remaining variables were complete, which demonstrates a high-quality data collection in the five EORTC studies. All multivariate Cox models were built adjusting the effects (hazard ratios) for these covariates. In our dataset, 6/1034 patients had exclusively bone metastasis at diagnosis, and therefore a separate analysis of this small subgroup could not be performed. Tentative explanations of this small number could be that (i) bone lesions alone are typically challenging to measure and most trials require a measurable disease to assess response/progression per RECIST 1.1 criteria [32], (ii) bone metastasis at diagnosis is a sign of extensive disease. A limitation of this work is the retrospective exploratory nature. For this analysis, we pooled both randomised and nonrandomised studies from the EORTC - STBSG database to increase the statistical power, which is likely to have introduced some selection bias in the population. Moreover, the interval of follow-up procedures for tumour reevaluation differed between the five trials analysed (e.g., every six or twelve weeks during treatment), which might have had an impact on PFS duration. A subgroup analysis was performed for 140 patients with bone metastasis at presentation to identify the metastatic organ site combination that is the most detrimental for OS and PFS. Due to the limited number of patients per treatment line (77 first-line, 63 second-line or later), results should be interpreted with caution.

Historically, there is heterogeneity in diagnostic tools for bone metastases. Routine use of imaging to detect bone lesions at diagnosis is not standard of care, nor has it ever been, at least for the majority of STS. Most likely, these lesions are detected in a routine computed tomography (CT), which can only detect more advanced bone metastases—e.g., rib or spine metastasis or pelvic disease or when investigating persistent bone pain. The use of more sensitive imaging techniques for screening, such as whole-body MRI and $[^{18}F]$ 2-fluoro-2-deoxy-D-glucose positron emission tomography/CT (FDG PET/CT), more routinely could increase the detection of bone metastases at diagnosis. However, as FDG PET could also miss bone metastases, whole-body MRI might be a preferable choice (e.g., for myxoid liposarcoma). As patients with metastatic STS survive these days somewhat longer than 20–25 years ago due to advances in supportive and multidisciplinary care, the prevalence of bone invasion is difficult to be ascertained and an increase will inevitably be observed.

According to the latest clinical practice guidelines for diagnosis, treatment, and follow-up of soft-tissue and visceral sarcomas, MRI is the main imaging modality if the primary STS is in the extremities, pelvis, and trunk [5]. Standard X-rays might also help to rule out a bone tumour to detect bone erosion and to show calcifications. When managing patients with advanced/metastatic STS and surgery of lung metastases is selected, it is mandatory to perform an abdominal CT scan and a bone scan or FDG PET to confirm that bone or other lesions are not present. In the case of skeletal metastases, radiotherapy should be considered for palliation of bone lesions at risk of fracture. Orthopaedic intervention is sometimes justified to improve the quality of life of these patients.

4.5 Conclusions

Adult STS patients receiving palliative systemic therapy with bone metastasis demonstrated an overall poor prognosis. A metastatic profile in the bone was detrimental for both OS and PFS in any treatment line, although not statistically significant. The hazard ratios–unadjusted and adjusted—were larger for patients treated in a first-line advanced or metastatic setting. A combined bone/lymph nodes metastatic presentation had the worst OS prognosis. For PFS, bone plus lung metastasis was the most detrimental combination. Of note, such combinations were statistically significant for first-line treatment. These findings may have implications for managing advanced/metastatic STS patients with bone metastasis at diagnosis.

Declarations

Data availability statement

The data that support the findings of this study are available from 5 clinical trials (62012, 62043, 62052, 62072, 62091) of the European Organisation for Research and Treatment of Cancer - Soft Tissue and Bone Sarcoma Group (EORTC - STBSG) database. Data can be requested via https://www.eortc.org/data-sharing/.

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Acknowledgements

The authors would like to thank Nathan Touati and Nicolas Sauvé for their support in realizing this research project. 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 as well as the EORTC Cancer Research Fund (ECRF). This work was supported by the European Organisation for Research and Treatment of Cancer - Soft Tissue and Bone Sarcoma Group (EORTC - STBSG).

Online supplementary materials

The Supplementary materials of this Chapter are available online at https://github.com/GKantidakis/Thesis_supplementary_materials/blob/main/Chapter4/Supplementary%20materials.docx.

References

[1] 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.

[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, jan 2021. ISSN 15320979. doi: 10.1097/PAS.0000000000001552. URL https://pubmed.ncbi.nlm.nih.gov/32796172/.

[3] E. K. Singhi, D. C. Moore, and A. Muslimani. Metastatic soft tissue sarcomas: A review of treatment and new pharmacotherapies. *Pharmacy and Therapeutics*, 43(7):410, jul 2018. ISSN 10521372. URL /pmc/articles/PMC6027857//pmc/articles/PMC6027857/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6027857/.

[4] G. K. In, J. S. Hu, and W. W. Tseng. Treatment of advanced, metastatic soft tissue sarcoma: Latest evidence and clinical considerations. *Therapeutic Advances in Medical Oncology*, 9(8):533–550, aug 2017. ISSN 17588359. doi: 10.1177/1758834017712963.

[5] A. Gronchi, A. B. Miah, A. P. Dei Tos, N. Abecassis, J. Bajpai, S. Bauer, R. Biagini, S. Bielack, J. Y. Blay, S. Bolle, S. Bonvalot, I. Boukovinas, J. V.M.G. Bovee, K. Boye, B. Brennan, T. Brodowicz, A. Buonadonna, E. De Álava, X. G. Del Muro, A. Dufresne, M. Eriksson, F. Fagioli, A. Fedenko, V. Ferraresi, A. Ferrari, A. M. Frezza, S. Gasperoni, H. Gelderblom, F. Gouin, G. Grignani, R. Haas, A. B. Hassan, S. Hecker-Nolting, N. Hindi, P. Hohenberger, H. Joensuu, R. L. Jones, C. Jungels, P. Jutte, L. Kager, B. Kasper, A. Kawai, K. Kopeckova, D. A. Krákorová, A. Le Cesne, F. Le Grange, E. Legius, A. Leithner, A. Lopez-Pousa, J. Martin-Broto, O. Merimsky, C. Messiou, O. Mir, M. Montemurro, B. Morland, C. Morosi, E. Palmerini, M. A. Pantaleo, R. Piana, S. Piperno-Neumann, P. Reichardt, P. Rutkowski, A. A. Safwat, C. Sangalli, M. Sbaraglia, S. Scheipl, P. Schöffski, S. Sleijfer, D. Strauss, S. Strauss, K. Sundby Hall, A. Trama, M. Unk, M. A.J. van de Sande, W. T. A. van der Graaf, W. J. van Houdt, T. Frebourg, P. G. Casali, and S. Stacchiotti. Soft tissue and visceral sarcomas: ESMO–EURACAN–GENTURIS Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 32(11):1348–1365, 2021. ISSN 15698041. doi: 10.1016/j.annonc.2021.07.006. URL https://doi.org/10.1016/j.annonc.2021.07.006.

[6] 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/.

[7] 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.

[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, jul 2018. ISSN 22965262. doi: 10.1159/000487511. URL https://www.karger.com/Article/Abstract/487511.

[9] K. G. Billingsley, M. E. Burt, E. Jara, R. J. Ginsberg, J. M. Woodruff, D. H. Y. Leung, and M. F. Brennan. Pulmonary metastases from soft tissue sarcoma: Analysis of patterns of disease and postmetastasis survival. *Annals of Surgery*, 229(5):602–612, 1999. ISSN 00034932. doi: 10.1097/00000658-199905000-00002. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1420804/.

[10] 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/.

[11] B. H. Hansen, J. Keller, M. Laitinen, P. Berg, S. Skjeldal, C. Trovik, J. Nilsson, A. Walloe, A. Kalén, and R. Wedin. The Scandinavian Sarcoma Group skeletal metastasis register: Survival after surgery for bone metastases in the pelvis and extremities. *Acta Orthopaedica Scandinavica, Supplement*, 75(311):11–15, apr 2004. ISSN 03008827. doi: 10.1080/00016470410001708270. URL https://pubmed.ncbi.nlm.nih.gov/15188660/.

[12] 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.

[13] P. L. Jager, H. J. Hoekstra, J. Albertus Leeuw, W. T. A. van der Graaf, E. G.E. De Vries, and D. Albertus Piers. Routine bone scintigraphy in primary staging of soft tissue sarcoma: Is it worthwhile? *Cancer*, 89(8):1726–1731, 2000. ISSN 0008543X. doi: 10.1002/1097-0142(20001015)89:8<1726::AID-CNCR12>3.0.CO;2-V.

[14] H. Yoshikawa, T. Ueda, S. Mori, N. Araki, S. Kuratsu, A. Uchida, and T. Ochi. Skeletal metastases from soft-tissue sarcomas. Incidence, patterns, and radiological features. *The Journal of Bone and Joint Surgery*, 79(4):548–552, 1997. ISSN 0301620X. doi: 10.1302/0301-620x.79b4.7372.

[15] H. Yoshikawa, A. Myoui, T. Ochi, N. Araki, T. Ueda, I. Kudawara, K. Nakanishi, H. Tanaka, and H. Nakamura. Bone metastases from soft tissue sarcomas. *Seminars in Musculoskeletal Radiology*, 3(2):183–189, 1999. ISSN 1098898X. doi: 10.1055/s-2008-1080061.

[16] S. Sleijfer, I. Ray-Coquard, Z. Papai, A. Le Cesne, M. Scurr, P. Schöffski, F. Collin, L. Pandite, S. Marréaud, A. De Brauwer, M. Van Glabbeke, J. Verweij, and J. Y. Blay. Pazopanib, a multikinase angiogenesis inhibitor, in patients with relapsed or refractory advanced soft tissue sarcoma: A phase II study from the European organisation for research and treatment of cancer-soft tissue and bone sarcoma group (EORTC Study 620. *Journal of Clinical Oncology*, 27(19):3126–3132, 2009. ISSN 0732183X. doi: 10.1200/JCO.2008.21.3223.

[17] P. Schöffski, I. L. Ray-Coquard, A. Cioffi, Bui-Nguyen B., S. Bauer, J. T. Hartmann, A. Krarup-Hansen, V. Grünwald, R. Sciot, H. Dumez, J. Y. Blay, A. Le Cesne, J. Wanders, C. Hayward, S. Marréaud, M. Ouali, and P. Hohenberger. Activity of eribulin mesylate in patients with soft-tissue sarcoma: A phase 2 study in four independent histological subtypes. *The Lancet Oncology*, 12(11):1045–1052, 2011. ISSN 14702045. doi: 10.1016/S1470-2045(11)70230-3.

[18] W. T. A. van der Graaf, J. Y. Blay, S. P. Chawla, D. W. Kim, B. Bui-Nguyen, P. G. Casali, P. Schöffski, M. Aglietta, A. P. Staddon, Y. Beppu, A. Le Cesne, H. Gelderblom, I. R. Judson, N. Araki, M. Ouali, S. Marréaud, R. Hodge, M. R. Dewji, C. Coens, G. D. Demetri, C. D. Fletcher, A. P. Dei Tos, and P. Hohenberger. Pazopanib for metastatic soft-tissue sarcoma (PALETTE): A randomised, double-blind, placebo-controlled phase 3 trial. *The Lancet*, 379(9829):1879–1886, 2012. ISSN 1474547X. doi: 10.1016/S0140-6736(12)60651-5.

[19] I. Judson, J. Verweij, H. Gelderblom, J. T. Hartmann, P. Schöffski, J. Y. Blay, J. M. Kerst, J. Sufliarsky, J. Whelan, P. Hohenberger, A. Krarup-Hansen, T. Alcindor, S. Marréaud, S. Litière, C. Hermans, C. Fisher, P. C. W. Hogendoorn, A. P. Dei Tos, and W. T. A. van der Graaf. Doxorubicin alone versus intensified doxorubicin plus ifosfamide for first-line treatment of advanced or metastatic soft-tissue sarcoma: A randomised controlled phase 3 trial. *The Lancet Oncology*, 15(4):415–423, 2014. ISSN 14745488. doi: 10.1016/S1470-2045(14)70063-4.

[20] B. Bui-Nguyen, J. E. Butrynski, N. Penel, J. Y. Blay, N. Isambert, M. Milhem, J. M. Kerst, A. K. L. Reyners, S. Litière, S. Marréaud, F. Collin, and W. T. A. van der Graaf. A phase IIb multicentre study comparing the efficacy of trabectedin to doxorubicin in patients with advanced or metastatic untreated soft tissue sarcoma: The TRUSTS trial. *European Journal of Cancer*, 51(10):1312–1320, 2015. ISSN 18790852. doi: 10.1016/j.ejca.2015.03.023.

[21] A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. A. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen, and D. Moher. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of clinical epidemiology*, 62(10):e1—-e34, 2009. ISSN 18785921. doi: 10.1016/j.jclinepi.2009.06.006.

[22] 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.

[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] B. Yücel, M. G. Celasun, B. Öztoprak, Z. Hasbek, S. Bahar, T. Kaçan, A. Bahçeci, and M. M. Şeker. The negative prognostic impact of bone metastasis with a tumor mass. *Clinics*, 70(8):535–540, 2015. ISSN 18075932. doi: 10.6061/clinics/2015(08)01.

[25] E. Svensson, C. F. Christiansen, S. P. Ulrichsen, M. R. Rørth, and H. T. Sørensen. Survival after bone metastasis by primary cancer type: A Danish population-based cohort study. *BMJ Open*, 7(9):1–7, 2017. ISSN 20446055. doi: 10.1136/bmjopen-2017-016022.

[26] J. F. Huang, J. Shen, X. Li, R. Rengan, N. Silvestris, M. Wang, L. Derosa, X. Zheng, A. Belli, X. L. Zhang, Yan M. Li, and A. Wu. Incidence of patients with bone metastases at diagnosis of solid tumors in adults: a large population-based study. *Annals of Translational Medicine*, 8(7):482, 2020. ISSN 23055839. doi: 10.21037/atm.2020.03.55.

[27] P. C. Ferguson, B. M. Deheshi, P. Chung, C. N. Catton, B. O'Sullivan, A. Gupta, A. M. Griffin, and J. S. Wunder. Soft tissue sarcoma presenting with metastatic disease. *Cancer*, 117(2):372–379, 2011. ISSN 0008543X. doi: 10.1002/cncr.25418.

[28] E. Younger, O. Husson, B. Asare, C. Benson, I. Judson, A. Miah, S. Zaidi, A. Dunlop, O. Al-Muderis, W. J. Van Houdt, R. L. Jones, and W. T. A. van der Graaf. Metastatic Soft Tissue Sarcomas in Adolescents and Young Adults: A Specialist Center Experience. *Journal of Adolescent and Young Adult Oncology*, 9(6): 628–638, 2020. ISSN 2156535X. doi: 10.1089/jayao.2020.0010.

[29] 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. Marréaud, 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.

[30] M. H. Younis, S. Summers, and J. Pretell-Mazzini. Bone metastasis in extremity soft tissue sarcomas: risk factors and survival analysis using the SEER registry. *Musculoskeletal Surgery*, pages 1–10, 2020. ISSN 20355114. doi: 10.1007/s12306-020-00673-9. URL https://doi.org/10.1007/s12306-020-00673-9.

[31] P. C. Ferguson, A. M. Griffin, B. O'Sullivan, C. N. Catton, A. M. Davis, A. Murji, R. S. Bell, and J. S. Wunder. Bone invasion in extremity soft-tissue sarcoma: Impact on disease outcomes. *Cancer*, 106(12): 2692–2700, 2006. ISSN 0008543X. doi: 10.1002/cncr.21949.

[32] E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer*, 45(2):228–247, 2009. ISSN 09598049. doi: 10.1016/j.ejca.2008.10.026. URL http://dx.doi.org/10.1016/j.ejca.2008.10.026.

Part II

Statistical models versus machine learning to predict survival for sarcoma and non-sarcoma clinical data

5

Neural networks for survival prediction in medicine using prognostic factors: a review and critical appraisal

This chapter is based on: G. Kantidakis, A. D. Hazewinkel, and M. Fiocco. Neural networks for survival prediction in medicine using prognostic factors: a review and critical appraisal. *Computational and Mathematical Methods in Medicine*, 2022:1176060, 2022.

Abstract

Survival analysis deals with the expected duration of time until one or more events of interest occur. Time to the event of interest may be unobserved, a phenomenon commonly known as right censoring, which renders the analysis of these data challenging. Over the years, machine learning algorithms have been developed and adapted to right-censored data. Neural networks have been repeatedly employed to build clinical prediction models in healthcare with a focus on cancer and cardiology.

We present the first ever attempt at a large-scale review of survival neural networks (SNNs) with prognostic factors for clinical prediction in medicine. This work provides a comprehensive understanding of the literature (24 studies from 1990 - August 2021, global search in PubMed). Relevant manuscripts are classified as methodological/technical (novel methodology or new theoretical model; 13 studies) or applications (11 studies). We investigate how researchers have used neural networks to fit survival data for prediction. There are two methodological trends: either time is added as part of the input features and a single output node is specified, or multiple output nodes are defined for each time interval.

A critical appraisal of model aspects that should be designed and reported more carefully is performed. We identify key characteristics of prediction models (i.e., number of patients/predictors, evaluation measures, calibration), and compare ANN's predictive performance to the Cox proportional hazards model. The median sample size is 920 patients, and the median number of predictors is 7. Major findings include poor reporting (e.g., regarding missing data, hyperparameters), as well as inaccurate model development/validation. Calibration is neglected in more than half of the studies. Cox models are not developed to their full potential, and claims for the performance of SNNs are exaggerated.

Light is shed on the current state of art of SNNs in medicine with prognostic factors. Recommendations are made for the reporting of clinical prediction models. Limitations are discussed, and future directions are proposed for researchers who seek to develop existing methodology.

5.1 Introduction

There is a growing interest by the medical community in applying machine learning (ML) to predict clinical outcomes [1]. This interest springs from the collection of large-volume patient information in electronic health records, and the growing availability of mixed data - for instance clinical and molecular. ML techniques are assumption-free and data-adaptive, making them attractive for modelling complex data. Artificial Neural Networks (ANNs) and other ML techniques have been used in healthcare for clinical diagnosis, prediction and to support decision making, e.g., in the domains of cancer and cardiology [2, 3].

Survival analysis (also called time-to-event analysis) is used to estimate the lifespan of a particular population under study. Survival data are omnipresent in medicine where the focus is on modelling a particular event of interest (for example disease-progression or death). This 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. The presence of censored observations makes the analysis of these data and the direct application of ML algorithms challenging, requiring modifications to the traditional approaches. As such, prediction of survival outcomes with ANNs - one of the most popular machine learning techniques in healthcare - poses unique hurdles with respect to the development and use of effective algorithms that can deal with right censoring (main focus here).

The most popular statistical model to analyse time-to-event data in medical research is the Cox proportional hazards defined as $\lambda(t|X) = \lambda_0(t)exp(X^T\beta)$, where X is the vector of covariates and $\lambda_0(t)$ is the baseline hazard function which is left unspecified. The effect of the covariates on the hazard is modeled by the parametric part $exp(X^T\beta)$ leading to the proportional hazard regression model [4]. Possible alternatives include parametric regression methods which make strong assumptions about the time distribution (e.g., exponential, Weibull or log-normal), and flexible non-parametric methods that do not make any prior assumptions regarding the time or the predictors (e.g Random Survival Forest, ANNs) [5–7]. A well-known non-parametric method to estimate the survival function was proposed by Kaplan-Meier [8]. It is used to estimate the fraction of patients alive after a specific starting point (for example, start of treatment).

ANNs have been widely used for survival data. Two decades ago, Ripley B. and Ripley R. published an overview that identifies the most appropriate survival neural networks (SNNs) for medical applications [9]. In their paper, they show different ways of adapting classification networks to survival data, and describe the disadvantages of these methods. An example of a work outside the medical field is discussed by Baesens *et al.* (2005) [10]; in this work various SNNs in context of personal loan data are used where the performance is compared to the Cox proportional hazards model [4]. In a recent comprehensive survey, Wang *et al.* (2019) [11] 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. Nevertheless, despite their non-negligible development in medicine for time-to-event data, a comprehensive review on SNNs using prognostic factors is missing. Prognostic factors are patient / disease characteristics (such as age, sex, or disease stage), which can be used to estimate the impact on survival, disease recurrence, or on others clinical outcomes. Typically prognostic factors do not include images (pathology images, tumor slices, whole slide images, etc.) or genetic marker sequences of DNA (variables from the area of bioinformatics).

In this article, we fill this gap with a structured overview of SNNs in clinical prediction with prognostic factors which can be used as a guideline for future research. Our aim is to provide a broad understanding of the literature (1st January 1990 - 31st August 2021), as part of a growing trend towards personalised medicine [1]. 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. During the 1990s, there were several modelling attempts, followed by a stagnation in scientific publications. In the past years, however, the advancement of machine learning has led to an increased interest from the medical community, where neural networks are now viewed as a promising modern approach to modelling medical data. In this review, we distinguish, following a chronological order, between methodological manuscripts (novel method or a new theoretical model) or applications that may build on existing methods to improve or adapt them based on the data at hand. The major distinctions between SNNs are 3-fold: a) data structure; some authors rely on a long format transformation of the dataset, whereas others use the original dataset, b) incorporating time information in the SNN; time is either added as part of the input features of the SNN, while specifying a single output node, or this step is omitted and multiple output nodes are specified - one for each time interval, c) estimation of outcome (output layer of the networks): some SNNs predict survival probabilities directly, while others estimate (conditional) death probabilities (hazard), from which the former can be calculated.

This work is supplemented with a critical appraisal on model aspects to be designed and reported more carefully in future studies. Key characteristics of prediction models (i.e., number of patients/predictors, evaluation measures, validation, calibration) are listed for methodological papers and applications, and the predictive performance of SNNs is compared to the Cox proportional hazards model (if reported in the papers). We conclude with recommendations on the correct application of SNNs in context of clinical prediction models, and discuss limitations and potential directions of future research. Particular interest is on SNNs applied to cancer prediction in contrast to other medical fields.

This manuscript is organised as follows. In Section ”Conducting the review” we describe our search and review strategy. Section ”Methodologies” focuses on the various SNN approaches identified. We present in a chronological order ”Early methodological approaches”, ”Approaches at the beginning of new millennium”, and ”Modern methodological approaches”. Section ”Applications” summarizes 11 applications to real or simulated data. In Section ”A critical perspective” we perform a critical appraisal of relevant studies, considering their ”General study characteristics”, ”Model development” aspects, ”Model validation”, and ”Comparison with Cox model’s performance”. Section ”Discussion” provides a discussion of current limitations and future directions.

5.2 Conducting the review

We searched the Medline biomedical database from 1st January 1990 to 31st August 2021 and identified 261 relevant studies where survival prediction was estimated using ML techniques. An additional 15 studies were identified by looking at references of selected papers and a previous literature overview by Ripley B. and Ripley R. in 2001 [9]. After removing duplicates and performing a screening of title and abstract, a total of 62 articles were considered.

Our search strategy is summarized in Figure 5.1 as a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram [12]. We identified 24 relevant studies, 13 methodological and 11 applications. Studies were considered eligible if they described the development of an SNN prediction model using prognostic factors, or its application (may build on an existing method to improve it) to real-word medical data or simulation studies. We define an SNN prediction model as an ANN adapted to survival data and capable of making individual patient predictions with prognostic factors. We excluded studies that focused on other ML approaches, performed standard ANN classification/regression, used an ANN as an extension of Cox regression, or were solely concerned with feature selection/reduction. Applications involving non-human subjects, images (pathology images, magnetic resonance imaging, tumor slices etc.) and computational biology analysis (e.g., predictions of gene expression) were disregarded. All non-original articles (e.g., reviews, tutorials) were excluded. The reader can find the search string in PubMed and the detailed list of inclusion/exclusion criteria in the Supplementary Material.

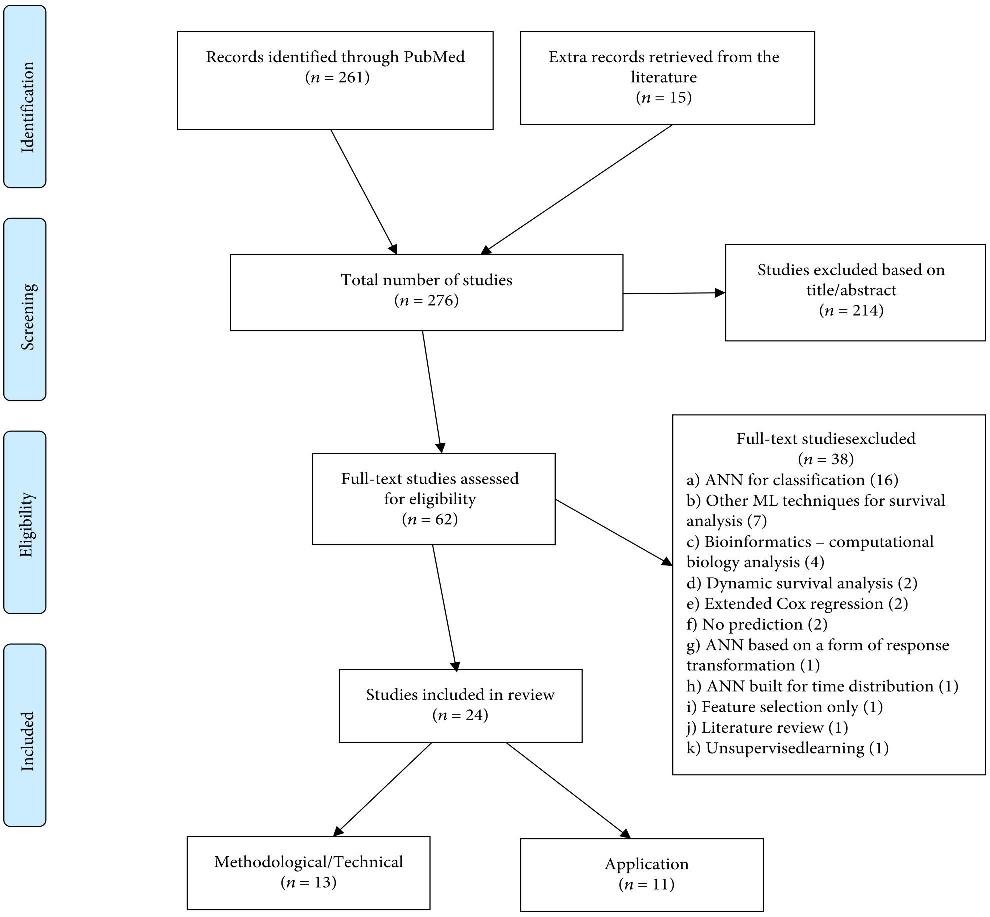

Identification
Screening
Eligibility
Included
Records identified through PubMed (n = 261)
Extra records retrieved from the literature (n = 15)
Total number of studies (n = 276)
Studies excluded based on title/abstract (n = 214)
Full-text studies assessed for eligibility (n = 62)
Full-text studiesexcluded (n = 38)
a) ANN for classification (16)
b) Other ML techniques for survival analysis (7)
c) Bioinformatics – computational biology analysis (4)
d) Dynamic survival analysis (2)
e) Extended Cox regression (2)
f) No prediction (2)
g) ANN based on a form of response transformation (1)
h) ANN built for time distribution (1)
i) Feature selection only (1)
j) Literature review (1)
k) Unsupervisedlearning (1)
Studies included in review (n = 24)
Methodological/Technical (n = 13)
Application (n = 11)

Figure 5.1: Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flowchart. Reasons for exclusion of the 214 studies in screening step 1 were: ML techniques for classification (n = 98), predictions based on individual's images (n = 25), models with focus on feature selection (n = 18), bioinformatics/computational biology analysis only (n = 15), other ML techniques for survival analysis (n = 15), unsupervised learning (n = 10), and other reasons (n = 33) including ML techniques for risk group stratification (n = 6), systematic/literature review (n = 6), new prediction tool (n = 5), ML techniques for regression (n = 4), ensemble of different ML techniques (n = 3), no prediction (n = 3), letter to the editor (n = 2), model for non-humans (n = 2), models with focus on feature reduction (n = 1), and tutorial/case study (n = 1).

5.3 Methodologies

In this Section, we present the methodological approaches of neural networks for survival analysis in chronological order. The majority of the techniques were developed in the 1990s, or early 2000s, followed by a long period with hardly any contributions in the field. Recently, the interest in the development of new methods has been rekindled, and modern approaches have been developed in specialized state-of-the-art software such as `keras` [13] in Python or R programming languages, which offer tremendous capabilities in modelling architecture and optimisers. Available options move beyond typical Feed Forward ANNs (FFANNs) and include deep learning and recurrent neural networks (RNNs), which were originally used only in non-medical context, for example for speech recognition and natural language processing. Table 5.1 provides notations used in the manuscript.

5.3.1 Basic components of survival neural networks

Neural networks have a layered structure which is based on a collection of units (also called nodes or neurons) for each layer. The input layer picks up the signals and passes them to the next layer which is called "hidden" after the application of a (usually non-linear) activation function. SNNs can have one or multiple hidden layers next to each other that connect with the previous layer. Signals are transmitted towards the output layer which is the last layer of

Notation	Description
T	Survival time
T_{max}	Maximum follow-up time (in years)
q_k	Conditional survival probability in (output) unit k
p_k	Conditional event probability in (output) unit k, with $p_k = 1 - q_k$
O_k	Output unit k
$\boldsymbol{w}$	Connection weight matrix
$\boldsymbol{\beta}$	Vector of regression coefficients
$\boldsymbol{x}$	Covariate matrix
$\boldsymbol{x_i}$	Vector of p covariates for individual i
Y_{ki}	Observed outcome of individual i in unit k
ϕ_h	Activation function for the hidden layer
ϕ_o	Activation function for the output layer
α	Bias unit (node)
E	Error (loss) function for the ANN
δ_{ik}	Event indicator of individual i for time interval $k = 1, \cdots, K$
p_{ki}	Probability that patient i relapses in time period (interval) k
γ_k	Cumulative event probability in (output) unit k

Table 5.1: Notations used in this review.

units where desired predictions are obtained. For SNNs, the output layer predicts (conditional) event probabilities or survival probabilities. A bias unit is an extra node added to each pre-output layer that stores the value 1 (it allows the activation function to be shifted to left or right to better fit the data). Bias units are not connected to any previous layer. Connections between the artificial units of different layers are called edges. These have a weight which adjusts through training increasing or decreasing the strength of each connection's signal. The simplest type of a neural network is a FFANN where the information moves in only one direction - forward - from the input units to the hidden units (if any) and to the output units. Recently, more and more researchers build deep neural networks which are ANNs with multiple hidden layers between the input and the output layer. Recurrent neural networks are also a class of FFANNs where connections between units form a directed or an undirected graph along a temporal sequence (of time intervals).

There are two basic data formulations for right-censored survival data which is the main focus here. For some methodologies, the wide data format is sufficient (standard data format with a single line per patient). However, several methods require data transformation into a long format where each patient is replicated multiple times with the survival times being divided into a set of k non-overlapping time intervals indicating months or years. Different terminologies such as prognostic variables, survival covariates, covariate vector, prognostic / clinical features, or predictors are used to denote the input units (features) of SNNs for the purpose of text enrichment. Note that some of the networks can include time-varying covariates (variables that change values over time during the follow-up period) as part of the input units if a methodology necessitates data transformation into a long format.

An example of two basic architectures for SNNs is illustrated in Figure 5.2. These are FFANNs with one hidden layer. The network's architecture depends on whether the time (interval) is coded as part of the prognostic variables or not.

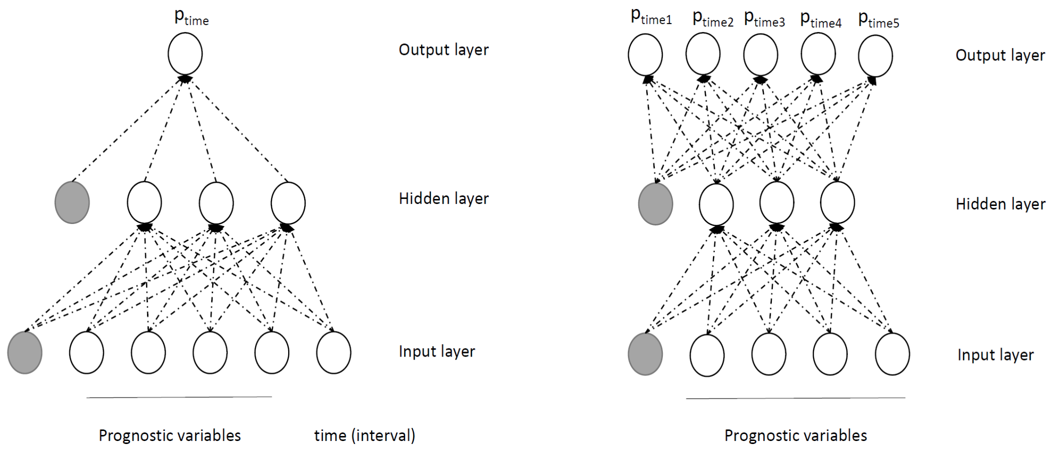


Figure 5.2: **Two basic architectures of survival neural networks.** Left panel: A network where time (interval) is coded as a prognostic variable (input feature). Data transformation into a long format is required for each patient. The output layer makes predictions in a given time interval. Right panel: A network where time (interval) is not coded as part of the prognostic variables. The wide data format is adequate for each patient. The output layer makes predictions at multiple sequential (non-overlapping) time intervals.

5.3.2 Early methodological approaches

The first attempt of modelling neural networks for censored data was made by Ravdin and Clark in 1992 [14]. The authors use a simple 3-layer FFANN and code time as an additional prognostic variable. Input features are replicated for several time intervals $[1, \cdots, T_{max}]$ with equal event rates, where T_{max} is the maximum follow-up time (in years). A patient who experienced the event of interest, is replicated exactly T_{max} times, while a censored patient is replicated only until the time of censoring. The output layer contains a single output unit representing the survival status and is set to 0 for all time intervals where the subject is alive, and to 1 for the time interval where the event of interest occurred (and the following intervals up to T_{max}). The `hyperbolic tangent` activation function is used for the units in the hidden and output layers. To correct for the bias introduced by the data transformation in a long format (as the number of deaths is over-represented in the late intervals), a selective sampling approach is performed, such that the proportion of deaths matches the information of the Kaplan-Meier [8] estimate. Selective sampling, however, is not an exact procedure and weighting cases would be a preferable approach [9]. The output layer provides death probabilities and can be seen as a prognostic index. An advantage of the methodology proposed by Ravdin and Clark is that time-varying covariates can be included, as subject entries are duplicated across multiple time periods.

In 1994, De Laurentiis and Ravdin proposed two alternative FFANNs [15]. The first is very similar to Ravdin and Clark's approach, and also specifies the time interval as an additional input variable. In this model, the distinct time intervals are selected such that each interval reflects a constant increase in event probability. Again, no data is present for censored cases after the last interval on study. Bias is controlled in a similar fashion, by obtaining the same frequency of censoring and events. The second FFANN proposed by De Laurentiis and Ravdin is a multiple time point model. This network does not require any modification of the wide data format and can accommodate only baseline characteristics and no time-varying covariates. The output layer is a vector with multiple output units (nodes) of I_k non-overlapping ordered intervals, and estimates event (death) probabilities. In the training data censored cases can be imputed at given times of follow-up (e.g., by Cox regression), or, alternatively, these output units can be deactivated. This approach mimics a k-class classification problem.

In the same year, Liestøl *et al.* proposed ANN generalizations of standard regression models used for survival analysis [16]. They constructed ANNs comparable to the 2^{nd} network proposed by De Laurentijs and Ravdin, with and without the hidden layer. These networks have k output units estimating hazard scores and are denoted as chain-binomial models. In principle, these networks can be viewed as a modification of Cox regression models, where the time axis has been partitioned into a number of disjoint intervals (grouped survival data). Such a model for the observed data may be specified via the conditional survival probability $q_k = P(T \geq t_k | T \geq t_{k-1})$, with

$k = 1, \cdots, K$. To implement it in a shallow network (no hidden layers) with K output nodes, the following parametrisation $w_{1j} = w_{2j} = \cdots = w_{Kj} = \beta_j$, $j = 1, \cdots, p$ can be applied, where w_{kj} is the weight assigned to the connection between input node j and output node K. This implies that all connections arising from the same input node j have the same weight. Then, for the output nodes the following function will be computed:

$$O_k(\boldsymbol{x}; \boldsymbol{\beta}, w_{k0}) = g(\boldsymbol{\beta}^T \boldsymbol{x} + w_{k0}), \tag{5.1}$$

with $\boldsymbol{x}$ the input variables, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$, and w_{k0} the weight from the bias node of the input layer. By applying a `sigmoidal` (logistic) activation function $g(x) = \frac{\exp(x)}{1+\exp(x)}$, we obtain an output, O_k, which corresponds to the event (death) probabilities p_k of the grouped version of the Cox model [4]. Applying the activation function $g(x) = 1 - \exp(-\exp(x))$, ensures estimation of event probabilities as in the grouped version of the Prentice and Gloeckler model [17]. The log-likelihood function of such a model corresponds to the negative error (loss) function:

$$E = -\sum_i \sum_k \{Y_{ki} \log(O_k(\boldsymbol{x_i}; \boldsymbol{w})) + (1 - Y_{ki}) \log(1 - O_k(\boldsymbol{x_i}; \boldsymbol{w}))\}, \tag{5.2}$$

for an individual $i = 1, \cdots, n$ of output unit $k = 1, \cdots, K$ having covariate vector $\boldsymbol{x_i}$, and observed responses (target values) Y_{ki}. This loss function is minimized with respect to $\boldsymbol{w}$, the connection weight matrix, using a back-propagation algorithm. Liestøl *et al.* (1994) suggested extensions to non-linear and non-proportional ANNs, which would require dropping the weight constraint, and adding a hidden layer to the previous shallow network. This would lead to an increase in the number of parameters. A non-linear and non-proportional ANN introduced in this way could be more appropriate in dealing with prognostic factors of non-linear and time-dependent effects.

Another attempt at adapting ANNs to survival data was made by Lapuerta *et al.* in 1995 [18]. Here, the output variable of the FFANN represents the time of occurrence of clinical coronary events. Time is divided into three 40-month periods plus an additional period in which no event occurred during the 120 months. The initial values for the output vectors denote event (1), no event (0) or censorship (as an unknown outcome with the symbol ?). To improve predictive ability, two separate networks are used to impute missing outcomes of early censored cases in each training set for the second period (40-79 months) and third period (80 - 120 months). Imputations are not performed in the test data. The authors create a predictor network where the output neuron with the highest value indicates the most likely outcome between four different classes. This approach might become cumbersome in terms of computational cost as it requires the use of multiple ANNs.

In 1998, Street used a standard FFANN with the `hyperbolic tangent` activation function for the units in the hidden and output layers [19]. The output layer consists of 11 ordered categories, $(0, 1], (1, 2], \cdots (9, 10]$ years, plus a final category denoting time of more than 10 years (in which the event did not occur). The network estimates the probability of disease-free survival up to a particular year, learning multiple classes in parallel. The output node is +1 as long as an individual is recurrence-free and -1 thereafter. Censored cases are incorporated directly in the training set using the probability that a patient will have disease recurrence before a certain time. The probability is obtained by employing a variation of the standard Kaplan-Meier method. Hereto, each censored individual may relapse at time t, given that no relapse has occurred at $t - 1$, and the disease-free survival time is used as the starting time (instead of time 0). Street uses the probabilities generated by the ANN to separate cases into those with ”good” and ”bad” prognosis and to estimate survival curves for individual patients. The author scales the probabilities to the range of the activation function by using *activation = 2*probabilities* - 1 and specifying the relative entropy error function. Street's approach cannot be considered as a classification problem because of the many incomplete data cases (it is unknown whether an individual is recurrence-free for these instances).

Biganzoli *et al.* (1998) introduced the partial logistic ANN (PLANN) [20]. This is a variation of the network proposed by Ravdin and Clark in 1992. It has a single hidden layer, one unit (node) in the output layer, and uses the time indicator as an additional input variable. Each prognostic variable is replicated for the number of intervals until death or end of follow-up. A major difference from Ravdin and Clark's approach is that here patients are not included after the time interval of death. Figure 5.3 shows a visual illustration of Biganzoli's PLANN. Nodes are represented by circles and the connections between them by dashed lines. The weights for the connection of the bias node with the hidden layer and the output layer are denoted by α_h and α_k, respectively. The weights for the connections between input and hidden nodes and hidden and output nodes are denoted w_{jh} and w_{hK}, respectively. The input layer consists of J nodes, given by the covariates, the time indicator, and a single bias node (0). The hidden layer consists of H nodes and one bias node (0). There is a single output unit (node) ($K = 1$) which computes conditional failure probabilities.

The output $\widehat{y_k}$ of a PLANN with a single hidden layer for an individual i can be defined as:

$$\widehat{y_k}(\boldsymbol{x_i}, \boldsymbol{w}) = \phi_o(\alpha_k + \sum_{h=1}^{H} w_{hk}\phi_h(\alpha_h + \sum_{j=1}^{J} w_{jh}x_{ij})), \tag{5.3}$$

for $j = 1, \cdots, J$ input nodes; $k = 1$ unique output node; ϕ_o and ϕ_h are the activation functions of the output and the hidden layer, respectively; x_{ij} represent the input value for an individual i and covariate j; α_h and α_k are the constant bias nodes for the input and the hidden layers, respectively. In general, ϕ_o will depend on the specified regression problem. For this SNN, Biganzoli *et al.* used the logistic activation function for both the hidden and output layer. FFANNs with logistic outputs (such as PLANN) can be regarded as flexible regression models for conditional probability estimation [21, 22].

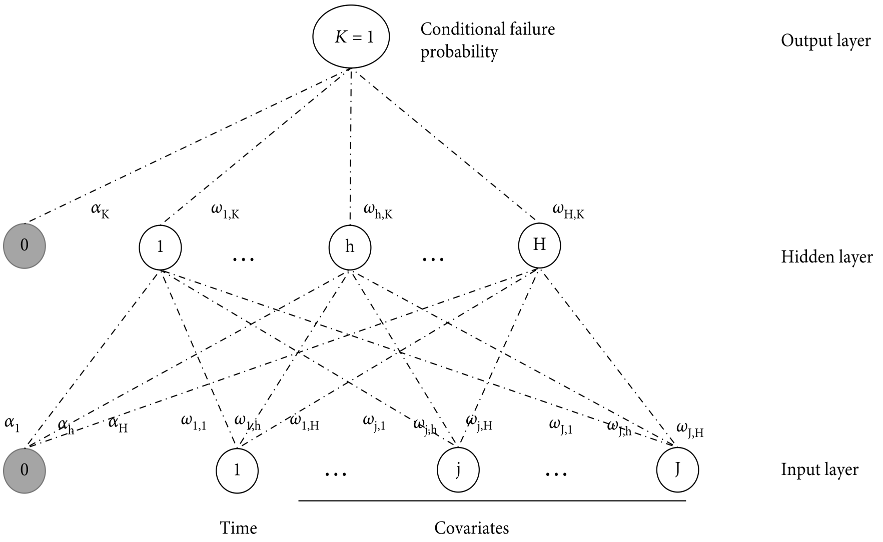


Figure 5.3: Visualization of the PLANN by Biganzoli *et al.* (1998) [20]

To enable inclusion of covariates, Cox proposed the proportional odds model [4] for grouped survival times. The formula below shows that discrete hazard rates can be modelled using a logistic regression model:

$$h_l(\boldsymbol{x_i}) = \frac{\exp(\theta_l + \boldsymbol{\beta}^T \boldsymbol{x_i})}{1 + \exp(\theta_l + \boldsymbol{\beta}^T \boldsymbol{x_i})}, \tag{5.4}$$

where $\theta_l = \log(\frac{h_l(0)}{1-h_l(0)})$ of $l = 1, 2, \cdots, L$ disjoint intervals $A_l = (t_{l-1}, t_l]$ with $t_0 = 0$ and l_i the interval of observation for the i^{th} subject.

PLANN is a generalization of partial logistic regression. The output values provide smoothed estimates of discrete hazards $h_l(\boldsymbol{x_i}, a_l)$ for the midpoint a_l of the time interval A_l. The survival is estimated as $S(t_l) = \prod_{l=1}^{L}(1 - h_l(\boldsymbol{x_i}, a_l))$. The error function of the model, for a given individual, i, is defined as

$$E(\boldsymbol{x_i}, a_l) = -\sum_{i=1}^{n}\sum_{l=1}^{l_i}\{\delta_{il}\log(h_l(\boldsymbol{x_i}, a_l)) + (1 - \delta_{il})\log(1 - h_l(\boldsymbol{x_i}, a_l))\}, \tag{5.5}$$

with δ_{il} the event indicator (1 at the interval of the event of interest, and 0 otherwise). This error function is equivalent to the cross-entropy error function and to Equation (5.2). A weight decay penalty term is added to the weights in Equation (5.5) to avoid overfitting ($E^* = E + \lambda \sum w^2$, regularisation L_2).

Biganzoli *et al.* used PLANN for flexible modeling of the hazard function of different cancer datasets, in an explanatory analysis. This approach has several favourable characteristics, including the presence of an analytical mathematical formulation, monotonicity of the survival curves and the option to include time-varying covariates, as the neural network is fitted to data that has been transformed to long format.

5.3.3 Approaches at the beginning of new millennium

Lisboa *et al.* extended the PLANN approach in 2003 by introducing a Bayesian framework with automatic relevance determination (ARD) [23]. This approach, called the PLANN-ARD, was inspired by David Mackay's 1995 review of Bayesian supervised ANNs [24]. PLANN-ARD is robust in estimating weight parameters, and carries out model selection, via regularization included within a Bayesian framework which consists of a sequential 3-step approach:

1. A penalty term, $L(\boldsymbol{w}, \boldsymbol{k})$, is added to the objective function (5.5) (similar to weight decay) where $\boldsymbol{k}$ is a set of Bayesian regularization parameters. The penalized objective function is $S(\boldsymbol{w}, \boldsymbol{k}) = E + L(\boldsymbol{w}, \boldsymbol{k})$.

2. Regularization parameters are estimated to control the penalty term.

3. Model selection is performed by interpreting the evidence in favor of candidate networks (hyperparameter selection).

For tuning the hyperparameters, the empirical Bayes approach is preferable to cross-validation, as the latter is frequently very computationally intensive. PLANN-ARD soft-prunes irrelevant variables to carry out model selection (as part of the Bayesian framework). The authors suggest that this methodology can be more efficient in the allocation of patients into prognostic groups compared to the Cox model.

Given that enough hidden units are specified, ANNs can approximate any functional relationships (i.e, interactions between covariates) [25, 26]. In 2004, Ripley R. *et al.* proposed two more discrete-time FFANNs [27]. Here, time is split into five non-overlapping time periods ($I_1; (0, 1]$, $I_2 : [1, 2)$, $I_3 : [2, 3)$, $I_4 : [3, 5)$ and $I_5 : [5, \infty)$). No multiple records (repeated entries of the same individual in the data) are needed for these approaches.

For the first network, the likelihood is calculated by $\prod_{i=1}^{N} \sum_{k=m_i+1}^{l_i} p_{ki}$, where m_i is the last time period the i^{th} patient is known to have survived without relapse, l_i is the final time period during which the patient may have relapsed, and p_{ki} is the probability that the i^{th} patient relapses in time period k. Ignoring the ordering of time periods the model can be estimated as

$$\log(p_k) - \log(p_1) = \eta_k(x) \quad (k = 2, 3, 4, 5), \tag{5.6}$$

with $\eta_k(x) = y_k - y_1$ using an ANN with the `softmax` activation function for the units of the output layer. The probabilities are computed as $p_k = \frac{\exp(y_k)}{\sum_l \exp(y_l)}$ (`softmax` formula) where y_k are outputs of the network.

The second network relies on more complex methodology which incorporates ordinal outcomes. This ANN has a single output unit to model the function η, which is now independent of the output class k. The cumulative event probabilities, $\gamma_k = F(t_k|\boldsymbol{x})$, are modelled as

$$\log(\frac{\gamma_k}{1 - \gamma_k}) = t_k - \eta(\boldsymbol{x}) \quad (k = 1, 2, 3, 4), \tag{5.7}$$

where t_k indicates the end of the k^{th} time period. Constraints on $t_k : t_1 \leq t_2 \leq t_3 \leq t_4$ are set to ensure that γ_k are increasing (ordinality of outcomes).

5.3.4 Modern methodological approaches

Deep learning ANNs are frequently used for prediction of output features - especially in the context of image classification [28, 29]. Applying deep learning methodology to medical survival data, however, poses the risk of overfitting, as the available sample sizes are typically small. In 2019, Matsuo *et al.* predicted survival by using a deep neural network (DNN) with a hierarchical structure and FFANNs in the first layers of the model [30]. The DNN contains 2 sub-networks with fully connected layers to jointly optimize the C-index and Mean Absolute Error (MAE). For each sub-network, the optimization is performed separately. The C-index quantifies the probability that the predicted event times of two randomly selected individuals have the same order as their true event times. Due to the presence of censored data not all pairs can be compared; this implies that a pair of subjects are comparable if the earliest time is an event, or both are events. The C-index is a measure of probability of concordance between the observed and the predicted survival. The MAE is defined as the absolute difference between the observed survival time and the survival time predicted by the sub-network. The authors found that the DNN performance improved on inclusion of more clinical features (input variables). A drawback of DNNs is that they are frequently computationally intensive and can be too complex for clinical insights.

In 2020, Bora Lee *et al.* developed time-binned neural networks to predict recurrence-free survival of non-small-cell lung cancer after surgery, using 30 clinico-pathological features [31]. The authors present one supervised learning binned-time survival analysis model (called s-DeepBTS) and one semi-unsupervised learning model (called su-DeepBTS). Here, we focus only on the supervised learning model s-DeepBTS. This is a shallow network where the output layer provides the survival probability in each pre-defined time interval (recurrence-free survival in months). The output value, y_j, is 1 when a patient is alive without relapse at the beginning of the j^{th} time interval I_j, and 0 after relapse. For censored patients, y_j is 1 until a patient is lost to follow-up and $\prod_{i=t_i \leq I_j} \left(\frac{1-d_i}{n_i}\right)$ after censoring occurs (Kaplan-Meier survival probability), where n_i is the total number of samples without recurrence at the beginning of the j^{th} time interval, and d_i is the number of events. The activation function of the output layer is the `sigmoid` (logistic). The root mean squared error (RMSE) between the true y_j and the predicted $\hat{y_j}$ is used as loss function.

Survival recurrent networks

Oh *et al.* (2018) use a survival recurrent network to train time-sequential outcome data for gastric cancer patients [32]. Their model is a DNN containing four recurrent neural network (RNN) layers in a total of seven layers, with the number of nodes gradually reduced across hidden layers. This network takes as inputs patient prognostic features and the survival probability of the previous year. In the following year, a comparison is performed between the predicted and observed survival probabilities. Survival for each time interval is denoted as either 1 (alive), 0 (dead) at the time of observation, or, for censored cases, as a ranking score in between 0 and 1. The predicted survival probability is updated every year with a weight, which is a tuneable parameter. The input layer consists of 25 prognostic features plus two survival features. The output layer consists of two nodes and is activated with the `softmax` function. As part of the procedure, variables of an individual are embedded (categorical variables are mapped to a vector of continuous numbers) for purposes of dimensionality reduction. This ANN approach to modelling survival data is complex which means it could lead to a poor generalizability on new data (overfitting training data), and/or a less intuitive interpretability of results.

A comparable learning algorithm was developed by Han *et al.* in 2018 [33]. Han *et al* describes a deep learning based survival model that can analyze patients lost to follow-up in a sequential manner (Figure 5.4). The network contains an input layer, 3 hidden layers with the number of nodes reduced across the layers, and an output layer. Information is updated every year. It is composed of three learning systems: nine clinical features $\boldsymbol{x}$, the survival probability p_{t-1} for the previous time of follow-up sequentially updated (10^{th} input feature), and non-parametric ranking scores $0 < r < 1$ for censored cases. Each time the ANN predicts the survival probability for the following year p_t. The recurrent loop reinforces training of the network sequentially, updating the residuals λ between the real outcome Y (1 = alive and 0 = dead) and the probability $\hat{Y}$, predicted by the SNN. A modulating parameter connects the residuals with the survival probabilities. As in Oh *et al.*'s network, the output layer contains two

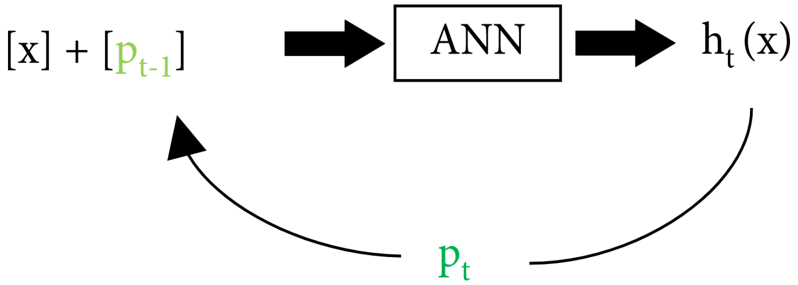


Figure 5.4: A schematic representation of the SNN by Han *et al.* in 2018 adapted from [33] built for 242 patients with synovial sarcoma. Here ANN means artificial neural network, $\boldsymbol{x}$ is the set of 9 clinical features, p_{t-1} is the survival probability of the previous year $t-1$ sequentially updated (10^{th} input feature), $h_t(\boldsymbol{x})$ is the predicted survival risk (alive/death probability), and p_t the predicted survival probability for the following year t.

nodes that are activated with the `softmax` function, which represent the predicted alive/death probability. This DNN might be biased because individuals who survive longer will be used more times for re-training, resulting in connection weight matrix, $\boldsymbol{w}$, optimized for longer survivors.

In 2019, Sung *et al.* developed RNNs with long short-term memory (RNN-LSTM), with the purpose of performing a risk classification for the prevention of cardiovascular disease, using national time-series health examination data [34]. This model includes a large number of patients (361239), randomly sampled in South Korea. The authors transform the binary output variable into multiple time-point output vectors for specific time-point analysis. The output layer includes yearly intervals from 2-10 years. The RNN-LSTM estimates the probability of survival for each interval. To take into account censored individuals, the probability of disease is estimated using Kaplan-Meier methodology. This network can incorporate time-varying covariates, and, in this application, provides more accurate predictions than the Cox model, suggesting such an approach may be well-suited to time-series data in particular.

5.4 Applications

In our search, we identified eleven applications, of which eight used real data and three used simulated data to investigate model behaviour in different scenarios. In some of these studies, the original methods were modified to improve prediction. Furthermore, as interpretability of results is crucial for clinical decision making, some studies focused on extracting interpretations from the ANNs (often called ”black boxes” as they do not provide insights on the structure of the function they approximate). The applications make use of different performance measures, which is likely due to the dynamic evolution of the field over the last decades.

In 2000, Xiang *et al.* [35] compared three different approaches in a simulation study. Nine data designs were simulated with 2 or 4 covariates, various censoring patterns, interaction between covariates, as well as proportional or non-proportional hazards. Survival times were generated using inverse probability transformations (details in the paper). For the purposes of this review, we only consider the SNN developed by Liestøl and colleagues [16] as the other two networks do not meet our search inclusion criteria. Time was divided into three distinct intervals, in which the hazard was assumed to be constant. The authors chose the general form of the method (no proportional hazards - dropping the weight constraint - see Section ”Early methodological approaches”). A simple FFANN with one input, one hidden and one output layer was developed. The quasi-Newton algorithm was used to minimize the negative log-likelihood. The performance of the SNN varied according to 9 underlying data designs, but none outperformed the Cox regression model. In 2003, Kattan [36] applied the same methodology to 3 large urological datasets. For this study, the author preserved proportional hazards for the network (by applying weight constraints). The author claims that - although theoretically attractive - ML techniques often do not result in an improved prediction accuracy.

Chi *et al.* (2007) [37] applied the SNN developed by Street [19] to two breast cancer datasets. The FFANN had three layers with `sigmoid` activation functions. It predicted the disease-free survival probability for each time unit. A slight modification was made to the labelling of the output vectors, using +1 up to recurrence time and 0 thereafter. The authors concluded that ANNs can successfully predict the probability of disease recurrence.

The PLANN-ARD Bayesian framework has been used several times for prediction in medical studies. In 2006, Jones *et al.* applied PLANN-ARD to data on patients with laryngeal squamous carcinoma [38], in which 97.9% (855 out of 873) died from the disease. When comparing the SNN to a Cox model, the authors found that the SNN performed better in separating patients' survival based on dichotomous variables. In 2007, Taktak *et al.* performed a double-blind multi-centre study for uveal melanomas [39]. They applied a PLANN-ARD, using 5-fold cross-validation to tune the hyperparameters instead of an empirical Bayes approach. A Bayesian mechanism was used to compensate for skewness in the data vector, resulting from the necessary data replication when transforming the data to long format [23]. The authors found a better performance of the SNN when compared to the semi-parametric Cox model and other models (the log-normal model, the partial spline model, and the partial logistic radial basis function network).

Five years after the development of PLANN-ARD, Lisboa *et al.* [40] applied the approach to breast cancer data. They extended the existing methodology to a competing risks model, where the two competing events are disease-free survival and breast cancer related mortality. This SNN provided a smoothed estimate for the hazards over time (assumptions about proportionality not required). The Bayesian framework for variable selection was extended to allow for continuous variables. To evaluate performance, a time-dependent C-index was used, which is an extension of the Area Under the Receiver Operating Characteristics (AUROC) measure [41]. The authors concluded that PLANN-ARD was a useful tool for risk assessment,as it distinguished high and low risk patients better than the Cox model.

In 2008, Amiri *et al.* applied a hierarchical ANN for risk assessment of gastric cancer patients [42]. Input features consisted exclusively of binary covariates. The network was a simple feed-forward with three nodes in the hidden layer, which computed the probability of survival in different periods. The authors observed that the SNN had a smaller mean standard error for the survival probabilities than the Cox proportional hazards model. They noted, however, that the baseline survival of the SNN may be unreliable as a consequence of the small sample size (N = 330) of the study.

In 2013, Biglarian *et al.* compared the PLANN method with Cox models in a simulation study [43]. Percentage of censoring was chosen between 20.0 and 80.0% and the data were simulated with linear and non-linear effects for the hazards. Model hyperparameters were tuned (a set of parameters was identified that leads to best performing model in the training data), using the Bayesian Information Criterion (BIC). Model fit was assessed in the test set, using the Mean Squared Error (MSE). This study concluded that prediction accuracy in more complex datasets depends on the level of censorship. Use of PLANN was suggested for data with a high percentage of censoring and for modelling complex interactions.

Spelt *et al.* applied PLANN to predict long-term survival after liver resection of metastases, for patients with metastatic colorectal cancer [44]. The model was an extension of the network by Biganzoli [20] and used an ensemble of SNNs. Training and validation were performed using 5-fold cross-validation, applied to 20 slightly different datasets, which were created by performing multiple imputations of missing values on the original data. The networks were combined within a single prediction model. The output of the ensemble was the mean output of all individual SNNs. Harrell's C-index was used as an performance measure [45]. Building on the work of Lippmann and Shahian (1997) for odds ratios [46], time-dependent hazard ratios for each variable in the SNNs were provided. Prognostic variables were ranked and minimized for the trained SNN. Order of variable relevance was obtained by measuring the change in baseline C-index (model with all variables) after removal of each of the risk factors, one at a time.

In 2018, Gong *et al.* investigated the PLANN approach in a simulation with a view to the field of pharmacometrics [47]. As in the study by Biglarian *et al.*, Gong *et al.* investigated both different proportional hazard functions (linear, non-linear) and different censoring percentages. To interpret the results, the authors employed the connection weights algorithm proposed by Garson (1991) [48] to calculate the relative importance of each input variable, and evaluated this method in a high-dimensional setting. Performance was assessed using the C-index, and the authors found that PLANN outperformed Cox regression. PLANN was less sensitive to changes in sample size and censoring percentage than Cox regression, and achieved the best performance when predictor variables assumed non-linear relationships in the hazard function. Additionally, for high-dimensional simulated data, PLANN was able to identify all pre-defined influential variables.

Kantidakis *et al.* (2020) compared PLANN with Cox models for large liver transplantation data (n = 62294 patients, 97 predictors). The authors described novel extensions to existing PLANN architecture (i.e, hyperparameters, activation functions, time interval specification) [49]. The extended PLANNs were tuned with the Integrated Brier Score (IBS) as the main criterion, which is a global summary of Brier score over the whole range up to the time horizon of the study (10 years) [50, 51]. The SNNs showed better performance than the Cox models based on IBS at 10 years, and the extended PLANN with 1 hidden layer was as calibrated as the Cox model with all variables (the predicted survival probabilities were similar to the observed survival probabilities estimated by using Kaplan-Meier's methodology). Emphasis was given on the advantages and pitfalls of each method and on the interpretability of the ML techniques. As in Gong *et al.*, the connection weights algorithm (Garson 1991) was used to identify the strongest prognostic factors.

5.5 A critical perspective

In this Section, we critically appraise relevant characteristics of the 13 methodological and the 11 application studies selected for this review (details on Figure 5.1). Excel sheets were constructed (available in the online version) that list the relevant prediction model characteristics. Additional information is provided in the Supplementary Material (overview of the extracted items in each study, 9 tables regarding the study characteristics).

5.5.1 General study characteristics

Of the 24 studies, 21 (87.5%) made use of existing data, while three (12.5%) applied the methods to simulated datasets. Descriptive statistics are shown in Table 5.2. The median total sample size was 920 patients, the median number of predictors was 7 (low-dimensional data), and the median percentage of censoring was 70.8% (10 of 24 studies considered multiple outcomes). Medical applications were mainly in the field of oncology (73.5%, 25 datasets). The majority of these studies conducted research on breast cancer (10 datasets), and cervical cancer, gastric cancer or prostate cancer (2 datasets each). Other fields of application comprised cardiovascular disease, coronary artery disease, liver transplantation and post-partum amenorrhea (2, 1, 1, and 1 datasets, respectively).

Clinical endpoints of interest included overall survival (analysed 16 times, 47.1%) and disease-free (or progression free, recurrence-free, relapse-free) survival (analysed 12 times, 35.3%). Remaining endpoints were breast cancer specific mortality (5.9%), death or hospitalization due to cardiovascular events (5.9%), menstruation-free survival (2.9%) and time to clinical artery events (2.9%).

	Min	1^{st} Qu.	Median	3^{rd} Qu.	Max	Excel lines
Total sample size	96	242	920	1616	361239	33
# of predictors	1	5	7	25.75	97	32
% of events	6.60	21.32	29.25	47.58	97.90	20

Table 5.2: General characteristics for the 24 studies. If multiple outcomes were predicted, multiple lines were used in the extraction sheet. Maximum number of lines was 34 (10 studies used multiple outcomes). For simulation studies, the number of predictors and percentage of events were not considered, unless they were fixed (e.g., not varied across simulations).

The strategy used to address the missing data (if any) was unclear for 9/21 (42.9%) studies (disregarding the 3 simulation studies that did not contain any missing data). Single or multiple stochastic imputation was used for 6 studies (28.6%) and ad-hoc approaches (separate attribute or mean / median imputation) were used in 5 studies (23.8%). One dataset had no missing data. Ad-hoc approaches to missing data can be problematic, as they can alter the distribution of a variable (if there is a substantial number of missing values). Multiple stochastic imputation, which replaces each missing value with multiple plausible values, is the preferable option [52], as the variability in multiple predictions reflects the uncertainty of the imputation process. It is understandable that multiple imputations may not be considered due to computational cost. Nevertheless, a single stochastic imputation is still superior to an ad-hoc fix, since imputation algorithms are more likely to preserve the original data structure. Examples of such algorithms are k-nearest neighbor and random forest (missForest) [53, 54].

5.5.2 Model development

Different aspects of model development for SNNs were considered: 1) whether the hyperparameters were tuned and which was the performance criterion for model development. 2) how the prognostic variables were scaled, 3) which programming language was used.

Hyperparameters are fundamental to the architecture of an ANN. They fine-tune the performance of a prediction model, preventing overfitting and providing generalizability of the model to new ”unseen” data. Choice of hyperparameters can be a challenge in the modern era of building SNNs with state-of-the-art software that allows for numerous choices. Commonly tuned parameters were penalty terms in the likelihood (e.g., weight decay) and the number of units (nodes) in the hidden layer(s). In the majority of studies (15, 62.5%), the approach to training hyperparameters was unclear, with 6 of these studies (25.0%) failing to report whether parameters were tuned or default values were chosen. In 4 studies (16.7%) parameters were tuned, in 3 studies (12.5%) some parameters were tuned and some were assigned default values, while in 2 studies (8.3%) default values only were chosen for the hyperparameters. The performance criterion for model development (hyperparameter tuning) was examined across the 24 studies. The training criterion was unclear for 6 studies (25.0%). For 5 studies (20.8%), neural network hyperparameters were trained based on the log-likelihood, for 3 studies based on the C-index (12.5%), and for 2 studies (8.3%) based on the Area Under the Curve (AUC). Other criteria used for model development are provided in the Supplementary Material. Better reporting of the choice of hyperparameters (which parameters were selected) and of the training procedure (how they were tuned) is needed. This will help researchers to better understand how the model was developed and will facilitate reproducibility.

In ANNs, input features are typically scaled to ensure that all features have a comparable scale, which allows an update of the same rate, resulting in faster algorithm convergence. The procedure was unclear in 10 of the 24 studies (41.7%), scaling was unnecessary in 7 studies (29.2%), and normalization (minimum and maximum values of features are used for scaling) was applied in 5 studies (20.8%). Standardization (mean and standard deviation of features are used for scaling) was applied in only 2 studies (8.3%). A precise description of the scaling approach (normalization or standardization) should be provided by researchers.

The programming language used for the development of the ANN was unclear in 7 studies (29.2%). Python was employed in 4 (16.7%) and R in 2 (8.3%) of the more recent studies. In the previous decades, Matlab was used 3 times (12.5%), NeuralWare 3 (12.5%), S-plus 3 (12.5%), while Epilog Plus and PlaNet were used 1 time each (4.2%). There is a trend towards employing Python, utilizing the `keras` and `Theano` libraries, which can build state-of-the-art ANNs with multiple options for layers, optimisers and error (loss) functions. These two libraries also have an interface available to the R programming language. It is strongly encouraged to share code developed for new methodologies or applications of existing methodologies in publicly available repositories (e.g., GitHub) to support reproducability and good clinical practice.

5.5.3 Model validation

We examined the validation approach for each of the 34 outcomes (clinical endpoints of the studies). Single random split was used 17 times (50.0%), with the data split into single train-test or train-validation-test parts. When the data are split into train-test parts the best model for training data is chosen based on model's performance on test data, whereas when the data are split into train-validation-test sets the best model for training data is selected based on the performance of the model on validation data. Then the test data are used to internally validate the performance of the model on new patients. Resampling (cross-validation or nested cross-validation) was used 9 times (26.5%). External validation (testing the original prediction model in a set of new patients from a different year, location, country etc.) was used 4 times (11.8%). External validation involved the chronological split of data into training and test parts 3 times (temporal validation), and validation of a new dataset 1 time. Multiple random split was used 2 times (5.9%), with the data split into train-test or train-validation-test data multiple times. Validation was not performed for 2 datasets (5.9%). We recommend reporting the steps of the validation approach in detail, to avoid misconceptions. In case of complex procedures, a comprehensive representation of the validation procedures can be insightful. Researchers should aim at performing both internal and external validations, if possible, to maximize the reliability of the prediction models.

Table 5.3 shows the performance measures used for model validation in the 24 studies. A popular measure in the survival field, the C-index, was employed in 8 studies (33.3%, as C-index or time-dependent C-index) and AUC in 5 studies (20.8%). Notably, during the screening process, several manuscripts were identified where AUC and C-statistic were used interchangeably. While there is a link between the dynamic time-dependent AUC and the C-index (the AUC can be interpreted as a concordance index employed to assess model discrimination) [55], the two are not identical and some caution is required. Apart from the C-index, there was no other established measure in the 24 studies (large variability). This issue is of paramount importance as validation (and development) of the SNNs depends on a suitable performance measure. Any candidate measure should take into account the censoring mechanism. By employing performance measures that are commonly used in traditional classification ANNs, such as accuracy, some SNNs were suboptimally validated. Consistency in the use of performance measures should also be considered. In the simulation study of Biglarian *et al.* in 2013 [43], hyperparameter values for PLANN were based on the Bayesian Information Criterion (BIC), while validation of the SNN performance on the test data was performed using the Mean Squared Error (MSE), and the comparison with Cox model was based on the C-index. Proper measures should be employed for model development and validation of time-to-event data (see the book of van Houwelingen and Putter [5]).

Reporting of confidence intervals for the predictive measures was examined; 13 studies (54.2%) did not provide confidence intervals. Repeated data resampling was practiced in 6 studies (25.0%). The following remaining approaches were observed: repeating the simulations 500 times; rerunning the SNN 10 times for each covariate; and using a non-parametric confidence interval based on Gaussian approximation (4.2% each). The method of choice was unclear in 2 studies (8.3%). There is a strong need for the development of methods which reflect the amount of uncertainty of an evaluation criterion. This would provide additional insights into the predictive accuracy of the model.

Another important aspect of a prediction model is calibration. It refers to the agreement between observed survival probabilities estimated with Kaplan-Meier's methodology and the predicted outcomes. Typically, a plot is produced where the subjects are divided into 10 groups based on the deciles of predicted probabilities. Observed survival probabilities are plotted against predicted. In this review, calibration plots were available for only 11 studies (45.9%). Calibration of the SNNs was not assessed in most studies, and as such a neutral comparison with the Cox proportional hazards model could not be established. This is in accordance with the findings of Christodoulou *et al.* (2019) [56], which pinpoint an urgent need for more attention in calibration of modern ML techniques versus traditional regression methods to achieve a fair model comparison in the classification setting.

Performance criterion	N (%)
C-index	7 (29.2%)
AUC	5 (20.8%)
log-likelihood	3 (12.5%)
Accuracy	2 (8.3%)
Global Chi-squared statistic of Cox regression	2 (8.3%)
Brier Score	1 (4.2%)
Comparison of predicted probabilities with Kaplan-Meier	1 (4.2%)
Integrated Brier Score (IBS)	1 (4.2%)
Mean Absolute Error (MAE)	1 (4.2%)
McNemar's test	1 (4.2%)
Mean Squared Error (MSE)	1 (4.2%)
Prognostic risk group discrimination	1 (4.2%)
Sensitivity	1 (4.2%)
Separation of cases into good and bad prognosis	1 (4.2%)
Specificity	1 (4.2%)
Survival curves comparison with log-rank test	1 (4.2%)
Time-dependent C-index (C^{td})	1 (4.2%)
Wilcoxon test (separation of cases into good and bad prognosis)	1 (4.2%)

Table 5.3: The performance measures used for model validation across the 24 studies.

5.5.4 Comparison with Cox model's performance

The Cox proportional hazards regression model assumes proportionality of hazards across different prognostic groups over time. Any interaction between predictors and/or time needs to be manually specified by the user (e.g., fractional polynomials, splines). This may be difficult when a large set of prognostic factors is available. ML techniques such as ANNs, which are flexible and data-adaptive, relax this assumption and can naturally incorporate multi-way interactions between the input features. This characteristic together with the rise of computational power and the collection of large-volumes of data (with electronic healthcare records) has contributed to the popularity of ANNs. However, the Cox model remains the most common choice for survival data. Therefore, any new prediction model including SNNs should be compared to the traditional Cox model to be considered in clinical practice.

Of 24 studies, 19 reported comparisons between Cox models and SNNs. We assessed whether interaction terms were specified in the models to obtain optimal predictive performance in Cox regression. Fifteen studies (78.9%) did not consider interaction terms between the predictors, information was unclear for 2 studies (10.5%), and 2 simulation studies considered interaction terms when applicable (10.5%). This result suggests suboptimal attention to the development of Cox models, which in turn undermines inferences made regarding comparative SNN and Cox model performance. For datasets with a large number of prognostic factors ($p > 10$), a number of interaction terms can be selected based on external knowledge and clinical expertise (see [6]).

Secondly, the author's claim for the performance of SNN was investigated. Among the 19 studies comparing SNN and Cox model's performance, 9 (47.4%) claimed better predictive performance of the SNN, while 5 reported a similar or better performance (26.3%) of the SNN compared to the Cox model. The performance was similar to Cox's model in 5 studies (26.3%). These result may be influenced by publication bias, as articles with favorable results are more likely to be published than articles with poor results.

A fair comparison between SNN and Cox model approaches to modelling survival data should include model validation with proper evaluation measures, a comparison of calibration curves and the inclusion of non-linear terms and interactions for Cox models, where applicable and possible. On the preface of his textbook on clinical prediction models (2019), Steyerberg reflects on exaggerated claims of modern method performance, which are lacking in convincing presentation of evidence and frequently involve suboptimal strategy choices for the regression model competitor [57].

5.6 Discussion

To the best of our knowledge, this is the first ever attempt at a large-scale review of SNNs in medicine using prognostic factors (1st January 1990 - 31st August 2021). It included 24 studies (13 methodological and 11 applications) where ANNs were employed for time-to-event prediction with right-censored data, mainly in the field of oncology (73.5%, 25 datasets) with a particular focus on breast cancer research (10 datasets). This might be due to the fact that survival analysis is well-suited to long-term outcome prediction (e.g overall survival), which is of primary interest in the field of oncology. Several methodologies were developed in the 1990s and were in later years applied to more complex datasets for clinical prediction. The majority of the SNNs were simple FFANNs, with the exception of some recent publications, which made use of deep ANNs and Survival Recurrent Networks. Amongst the methods used, two general trends can be distinguished: networks with a unique output unit and a time indicator variable added as an extra input feature, and networks with multiple outputs representing k non-overlapping time intervals. The former approach requires that the data are replicated multiple times, for each of the time intervals considered, and allows for the incorporation of time-varying covariates.

We excluded studies where SNNs were built for bioinformatics - computational biology analysis, dynamic survival analysis, focused on ANN extensions of the Cox model, studies that did not evaluate model performance, or where predictions were based on individual's images (pathology images, magnetic resonance imaging, tumor slices etc.) (see Section "Conducting the review"). We addressed this review in a pragmatic way using the biomedical database PubMed and focusing on SNNs for prediction using prognostic factors. We acknowledge that we may have missed some articles during the process. Below, we briefly summarize some other important methodological developments of the last three decades.

In 1995, Faraggi and Simon [58] extended the Cox model by replacing the linear function $\beta^T \boldsymbol{x_i}$ with the output $\phi_o(\boldsymbol{x_i}, \boldsymbol{w})$ of an ANN with `logistic` hidden and `linear` output layers. No bias unit is specified for the output layer, and the model is subject to the proportional hazards assumption. A modern deep survival analysis approach related to Faraggi and Simon's work was described by Katzman *et al.* in 2018 [59]. Here, the authors construct a deep FFANN where the output of the network is a single unit which predicts the log-risk function and can be used to extend Cox regression (`DeepSurv`; an open source Python module). `DeepSurv` provides personalised treatment recommendations and is capable of predicting the effect of a specific patient's characteristics on the risk of failure. A practical extension of such work could involve the use of convolutional neural networks on medical imaging data for risk prediction (out of scope here). Very recently, a multilayer deep learning Cox-based prediction model (another extension of the linear function $\beta^T \boldsymbol{x_i}$) was proposed by Sun *et al.* [60] for high dimensional survival data in a genome wide association study, and was also applied by Hao *et al.* [61] in ultra-high-dimensional genomic data (number of predictors $> 10^5$). It is shown that it can not only outperform several existing survival prediction models (Random Survival Forest, Cox LASSO, Cox Ridge) in terms of accuracy, but also detect clinically meaningful risk subgroups by effectively learning the complex structures among genetic variants.

In 2006, Biganzoli *et al.* extended the PLANN methodology to competing risks (PLANNCR), in a study of primary invasive breast cancer [62]. PLANNCR is an ANN for the joint modelling of discrete cause-specific hazards and can be used for both discrete and grouped survival data. The output layer contains multiple nodes (competing risks) that estimate discrete conditional event probabilities. PLANNCR uses `logistic` and `softmax` functions for the hidden and output layer, respectively. The error function that is minimized corresponds to the multinomial likelihood. The degree of smoothing for output nodes is modulated by the number of hidden nodes and the penalization of the error function (weight decay in the loss function). PLANNCR can be implemented using standard ANN software that is able to accommodate multiple classification. In 2009, Lisboa *et al.* published an ARD extension of PLANNCR (PLANNCR-ARD) [63]. The authors apply the methodology for local and distal recurrence of breast cancer, in an approach that requires no prior domain knowledge, and performs model selection within a Bayesian framework. Kantidakis *et al.* performed a simulation study in 2021 [64] to compare the predictive performance of PLANN original [20] and PLANN extended (1 hidden layer) [49] with Cox models for non-complex clinical data (small /medium sample size, low dimensional). Methods were compared for scenarios where different percentages (20, 40, 61, 80%) of censored data were present. ML and Cox models showed similar predictive performance on simulated data for most scenarios. C-index, Brier score, or Integrated Brier Score were used for the comparison. Results of this study show that the statistical models were often better calibrated.

In 2013, Fornili *et al.* presented a simple FFANN for the purpose of analyzing disease dynamics in a survival analysis context [65]. This SNN - applied to breast cancer data - specifies, for the output unit, the smoothed hazard as a function of time interval and prognostic factors. This approach is known as Partial Exponential ANN (PEANN), and is a non-linear extension of generalized linear models for right-censored survival data [66] and a direct extension of the PLANN method for piece-wise data. The network uses the `logistic` and the `exponential` functions for the hidden and output layers, respectively. Such method is best-suited to modelling the hazard shape of diseases with a long follow-up, and allows for the exploration of non-linear and non-additive effects.

Ching *et al.* developed *Cox-nnet* in 2018 [67] - a new ANN framework for patient prognosis using transcriptomics data. This FFANN has an input layer, one fully connected hidden layer with 143 nodes (set as the square root of more than 20000 input features) and one output ”Cox regression” layer. To avoid overfitting, different regularization methods are employed, such as ridge (weight decay), dropout, and a combination of ridge and dropout [68]. The author compared an ANN with no hidden layer (shallow), a single hidden layer, and two hidden layers, and found that a single layer neural network had the best performance based on C-index.

Very recently, two novel deep learning approaches have been published for dynamic survival analysis. Changhee Lee *et al.* proposed *Dynamic-DeepHit* for longitudinal and time-to-event data with competing risks to issue dynamically updated survival predictions for cystic fibrosis patients [69]. This network is trained by leveraging a combination of loss functions that capture the right-censoring and the associations of longitudinal measurements with disease progression. It provides a remarkable improvement in discriminating individual risks of different causes of failure. This model can also provide useful clinical insights by identifying covariates which are influential for different competing risks (risk predictions interpretation). In the same year, Jarrett *et al.* developed temporal convolutional networks for Alzheimer's disease (called *MATCH-Net*) [70]. This CNN is designed to capture temporal dependencies and heterogeneous interactions in covariates and patterns of missingness for personalised risk prognosis. Its performance is compared with statistical and deep learning benchmarks showing incremental sources of gain from various design choices.

Unclear addressing of missing data (42.9%) or ad-hoc methods (23.8%)
Unclear reporting of hyperparameters (62.5%)
Unclear reporting of the performance criterion for model development (25.0%)
Unclear scaling of prognostic factors (41.7%)
Unclear programming language for SNNs (29.2%)
Large variability and improper performance measures for survival data
External validation for only 4 outcomes (11.8%)
No confidence intervals for the predictive measures (54.2%)
No calibration plots (54.2%)
No interactions in Cox regression or unclear reporting (89.5%)

Table 5.4: Summary of the findings from the critical appraisal across the 24 manuscripts.

A critical appraisal was carried out to pinpoint current limitations and identify future research directions. Our findings are summarized in Table 5.4. Based on these findings, we make the following recommendations. Complete and transparent reporting of modelling steps and analysis is necessary (e.g., more details on training and test data), to enable reproducibility, and to allow critical appraisal of the results by a wider audience [71, 72]. In the event of missing data, a single or multiple imputation approach should be used, prior to SNN development (see also Section ”General study characteristics”), to avoid discarding patients from nearly complete records. Hyperparameter selection and training should be more extensive with the performance criterion for model development clearly reported. Careful tuning of parameters can prevent overfitting and improves the generalizability of the prediction model. When developing an SNN, the following elements must be considered: the number of hidden nodes, the penalty terms, the activation functions and the optimizers. Of particular importance is the choice of performance measure for model validation, which we observed to be sometimes poorly chosen (see Section ”Model validation”). A suitable performance measure should take into account the censoring mechanism (see the book of van Houwelingen and Putter [5]). Additionally, model calibration should be assessed, preferably through calibration plots. In the studies of our review, the median sample size was 920 patients and the median number of predictors was 7 (low-dimensional data). Larger datasets and/or more predictors are needed for better model development/validation and improved generalizability. These aspects are of great value as suboptimal clinical prediction models are responsible for research waste [57, 73]. Comparisons of SNNs with conventional regression models should be made in a fair manner, with the conventional models fully developed and interactions and/or non-linear terms included when appropriate.

When comparing SNN methods to traditional approaches in simulation, scenarios with different sample sizes, censoring percentages, and numbers of covariates (fixed and/or time-varying) can be considered. Comparing SNNs in low and high dimensional settings is relevant to areas of study like bioinformatics. ANNs are often referred to as ”black boxes”, due to the lack of interpretability (ANNs do not provide coefficients/hazard ratios as a Cox model does). The more complicated (deep) an ANN is, the more challenging interpretation of results becomes. As interpretability is necessary for clinical decision making, more emphasis should be placed on the development of methods which can facilitate SNN model interpretation. In Section ”Applications”, we discussed several applications that attempt to address this aspect. Olden’s 2004 article provides a comparison of different techniques for ANN interpretability (e.g., variable importance) [74].

In the studies considered in this review, variability of performance (e.g., through the use of confidence intervals) was not well documented. The studies that did employ confidence intervals, typically used a resampling approach. Multiple resampling of all empirical data using bootstrapping can be an advantageous approach when sample size is limited, as it avoids the need to split the data for model development. While confidence intervals are necessary for model assessment, obtaining them can be computationally expensive. Further methods and guidelines for obtaining confidence intervals are needed. Another aspect which is under-reported in studies concerns the stability of SNN. ML techniques are algorithmic approaches that inherently rely on random processes to obtain generalisable models (e.g., for ANNs, values of weights are randomly initialized). Consequently, when rerunning the same model on the same data, there will be variations in output. In the event of a well-tuned model, these variations will be small and the model can be described as stable. In contrast, an incorrect approach to hyperparameter tuning may result in an unstable model with large variations. When validating an SNN, we recommend rerunning the model several times under the same parameterisation, to evaluate the stability of network’s performance.

In Section ”Methodologies”, 13 methodologies were presented for survival prediction with SNNs. Some studies predict survival probabilities in the units of the output layer, which allows the estimated survival curves to be non-monotonic (such networks cannot be forced to generate monotonically decreasing output units that predict survival probabilities) [10]. This can be avoided by predicting conditional hazard probabilities instead (from which survival probabilities can be readily calculated), as it is done, for example, in the PLANN and PLANN-ARD methods. We recommend that future ANN methodologies either estimate the (smoothed) hazard function in the output unit(s), or alternatively add constraints to ensure monotonicity of the survival curve. Furthermore, in this review, all neural networks were developed for right-censored data. Future work should focus on building SNNs for other types of censoring such as left or interval censoring, which are less common in practice compared to right censoring.

SNNs developed in recent years usually have more complicated structures and make use of multiple hidden layers (deep learners). It should be noted, however, that increasing the complexity of an ML prediction model does not necessarily translate to improved performance on new clinical data. An increase in the complexity, and by extent flexibility, of a network, may produce a model that is too attuned to the training data with poorer generalization to new data (overfitting), resulting in less accurate survival probabilities than a simpler network. Additionally, increasing complexity will pose additional challenges regarding interpretation. For clinical survival data using prognostic factors, sample size and number of predictors is likely to be insufficient for employing such advanced ML techniques. This may explain the frequent use of PLANNs in applications, as a PLANN guarantees survival curve monotonicity, relaxes proportional hazard assumption, and employs a relatively simple network structure.

5.7 Conclusions

Nowadays, prediction models are ubiquitous in a wide range of research fields (e.g., medicine, engineering, finance) and are becoming increasingly relevant in the medical field, as a result of the large-scale data collection and the increase in biological knowledge. In this paper, we discussed clinical prediction models with SNNs in the healthcare domain using prognostic factors, which can be used as guidance for future works. Light was shed on SNN approaches developed and applied from 1990 to August 2021. We assessed various methodological and practical aspects, including study characteristics, model development/validation, and comparison with Cox models. It is our opinion that, in the future, artificial intelligence and related algorithms (e.g., ANNs and SNNs) might become an integral part of personalised and evidence-based medicine. This review and critical appraisal hopely provides enough stimuli to researchers to be inspired from these methods, and seek for new developments.

List of abbreviations

ANN, artificial neural network; AUC, Area Under the Curve; DNN, deep neural network; FFANN, feed forward artificial neural network; ML, machine learning; 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; RNN, recurrent neural network; SNN, survival neural network.

Declarations

Data availability statement

The excel sheets (SNNs review - short, SNNs review - long) developed for the critical appraisal of the 24 studies are provided online.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgements

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 did not have any involvement in the designing of this study, the collection, analysis and interpretation of data, the writing of this report, or the decision to submit the article for publication. The authors would like to thank Saskia Litière for providing her feedback during the preparation of this manuscript.

Online supplementary materials

The Supplementary material and the excel sheets of this Chapter are available online at https://github.com/GKantidakis/Thesis_supplementary_materials/tree/main/Chapter5.

References

[1] 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.

[2] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis. Machine learning applications in cancer prognosis and prediction, 2015. ISSN 20010370.

[3] N. Shahid, T. Rappon, and W. Berta. Applications of artificial neural networks in health care organizational decision-making: A scoping review, 2019. ISSN 19326203.

[4] 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.

[5] 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.

[6] F. E. Harrell Jr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer, 2nd edition, 2015. ISBN 978-3-319-19425-7. doi: https://doi.org/10.1007/978-3-319-19425-7. URL http://www.springer.com/series/692.

[7] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, sep 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.

[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] B. D. Ripley and R. M. Ripley. Neural networks as statistical methods in survival analysis. *Clinical Applications of Artificial Neural Networks*, pages 237–255, 2001.

[10] B. Baesens, T. Van Gestel, M. Stepanova, D. Van Den Poel, and J. Vanthienen. Neural network survival analysis for personal loan data. *Journal of the Operational Research Society*, 56(9):1089–1098, 2005. doi: 10.1057/palgrave.jors.2601990.

[11] 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.

[12] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, 151(4):264–269, 2009.

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

[14] P. M. Ravdin and G. M. Clark. A practical application of neural network analysis for predicting outcome of individual breast cancer patients. *Breast Cancer Research and Treatment*, 22(3):285–293, 1992.

[15] M. De Laurentiis and P. M. Ravdin. Survival analysis of censored data: Neural network analysis detection of complex interactions between variables. *Breast Cancer Research and Treatment*, 32:113–118, 1994.

[16] K. Liestol, P. K. Andersen, and U. Andersen. Survival analysis and neural nets. *Statistics in Medicine*, 13 (12):1189–1200, 1994. doi: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780131202.

[17] R. L. Prentice and L. A. Gloeckler. Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data. *Biometrics*, 34:57–67, 1978. doi: 10.2307/2529588.

[18] P. Lapuerta, Azen S. P., and LaBree L. Use of neural networks in predicting the risk of coronary artery disease. *Computers and Biomedical Research*, 28(1):38–52, 1995. doi: 10.1006/cbmr.1995.1004.

[19] W. N. Street. A Neural Network Model for Prognostic Prediction. *ICML*, pages 540–546, 1998.

[20] 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.

[21] C. M. Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995. ISBN 978-0-19-853864-6.

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

[23] 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.

[24] D. J. C. Mackay. Probable networks and plausible predictions - a review of practical bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505, 1995. doi: 10.1088/0954-898X_6_3_011.

[25] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. doi: 10.1016/0893-6080(89)90020-8.

[26] 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.

[27] R. M. Ripley, A. L. Harris, and L. Tarassenko. Non-linear survival analysis using neural networks. *Statistics in Medicine*, 23(5):825–842, 2004. doi: 10.1002/sim.1655.

[28] C. Affonso, A. L. D. Rossi, F. H. A. Vieira, and A. C. P. F. de Carvalho. Deep learning for biological image classification. *Expert Systems with Applications*, 85:114–122, nov 2017. ISSN 09574174. doi: 10.1016/j.eswa.2017.05.039.

[29] M. Xin and Y. Wang. Research on image classification model based on deep convolution neural network. *Eurasip Journal on Image and Video Processing*, 2019(1):1–11, 2019. doi: 10.1186/s13640-019-0417-8. URL https://jivp-eurasipjournals.springeropen.com/articles/10.1186/s13640-019-0417-8.

[30] K. Matsuo, S. Purushotham, B. Jiang, R. S. Mandelbaum, T. Takiuchi, Y. Liu, and L. D. Roman. Survival outcome prediction in cervical cancer: Cox models vs deep-learning model. *American Journal of Obstetrics and Gynecology*, 220(4):381.e1—-381.e14, 2019. doi: 10.1016/j.ajog.2018.12.030.

[31] B. Lee, S. H. Chun, J. H. Hong, I. S. Woo, S. Kim, J. W. Jeong, J. J. Kim, H. W. Lee, S. J. Na, K. S. Beck, B. Gil, S. Park, H. J. An, and Y. H. Ko. DeepBTS: Prediction of Recurrence-free Survival of Non-small Cell Lung Cancer Using a Time-binned Deep Neural Network. *Scientific Reports*, 10(1):1–10, 2020. ISSN 20452322. doi: 10.1038/s41598-020-58722-z. URL http://dx.doi.org/10.1038/s41598-020-58722-z.

[32] S. E. Oh, S. W. Seo, M. G. Choi, T. S. Sohn, J. M. Bae, and S. Kim. Prediction of Overall Survival and Novel Classification of Patients with Gastric Cancer Using the Survival Recurrent Network. *Annals of Surgical Oncology*, 25(5):1153–1159, 2018. doi: 10.1245/s10434-018-6343-7.

[33] I. Han, J. H. Kim, H. Park, H. S. Kim, and S. W. Seo. Deep learning approach for survival prediction for patients with synovial sarcoma. *Tumor Biology*, 40(9), 2018. doi: 10.1177/1010428318799264.

[34] J. M. Sung, I. J. Cho, D. Sung, S. Kim, H. C. Kim, M. H. Chae, M. Kavousi, O. L. Rueda-Ochoa, M. Arfan Ikram, O. H Franco, and H. J. Chang. Development and verification of prediction models for preventing cardiovascular diseases. *PLoS ONE*, 14(9), 2019. doi: 10.1371/journal.pone.0222809.

[35] A. Xiang, P. Lapuerta, A. Ryutov, J. Buckley, and S. Azen. Comparison of the performance of neural network methods and Cox regression for censored survival data. *Computational Statistics & Data Analysis*, 34(2): 243–257, 2000. doi: https://doi.org/10.1016/S0167-9473(99)00098-5. URL www.elsevier.com/locate/csda.

[36] M. W. Kattan. Comparison of Cox regression with other methods for determining prediction models and nomograms. *Journal of Urology*, 170(6):S6—-S10, 2003. doi: 10.1097/01.ju.0000094764.56269.2d.

[37] C. L. Chi, W. N. Street, and W. H. Wolberg. Application of Artificial Neural Network-Based Survival Analysis on Two Breast Cancer Datasets. *AMIA Annual Symposium Proceedings*, pages 130—-134, 2007.

[38] A. S. Jones, A. G. F. Taktak, T. R. Helliwell, J. E. Fenton, M. A. Birchall, D. J. Husband, and A. C. Fisher. An artificial neural network improves prediction of observed survival in patients with laryngeal squamous carcinoma. *European Archives of Oto-Rhino-Laryngology*, 263(6):541–547, jun 2006. doi: 10.1007/s00405-006-0021-2.

[39] A. Taktak, L. Antolini, M. Aung, P. Boracchi, I. Campbell, B. Damato, E. Ifeachor, N. Lama, P. Lisboa, C. Setzkorn, V. Stalbovskaya, and E. Biganzoli. Double-blind evaluation and benchmarking of survival models in a multi-centre study. *Computers in Biology and Medicine*, 37(8):1108–1120, 2007. doi: 10.1016/j.compbiomed.2006.10.001.

[40] P. J. G. Lisboa, T. A. Etchells, I. H. Jarman, M. S. Hane Aung, S. Chabaud, T. Bachelot, D. Perol, T. Gargi, V. B., S. Bonnevay, and S. Négrier. Time-to-event analysis with artificial neural networks: An integrated analytical and rule-based study for breast cancer. *Neural Networks*, 21(2-3):414–426, 2008. doi: 10.1016/j.neunet.2007.12.034.

[41] L. Antolini, P. Boracchi, and E. Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, 2005. doi: 10.1002/sim.2427.

[42] Z. Amiri, K. Mohammad, M. Mahmoudi, H. Zeraati, and A. Fotouhi. Assessment of gastric cancer survival: using an artificial hierarchical neural network. *Pakistan Journal of Biological Sciences*, 11(8):1076–1084, 2008. doi: 10.3923/pjbs.2008.1076.1084.

[43] A. Biglarian, E. Bakhshi, A. R. Baghestani, M. R. Gohari, M. Rahgozar, and M. Karimloo. Nonlinear survival regression using artificial neural network. *Journal of Probability and Statistics*, 2013, 2013. doi: https://doi.org/10.1155/2013/753930.

[44] L. Spelt, J. Nilsson, R. Andersson, and B. Andersson. Artificial neural networks-A method for prediction of survival following liver resection for colorectal cancer metastases. *European Journal of Surgical Oncology*, 39(6):648–654, 2013. doi: 10.1016/j.ejso.2013.02.024.

[45] 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.

[46] R. P. Lippmann and D. M. Shahian. Coronary Artery Bypass Risk Prediction Using Neural Networks. *The Annals of thoracic surgery*, 63(6):1635–1643, 1997.

[47] X. Gong, M. Hu, and L. Zhao. Big Data Toolsets to Pharmacometrics: Application of Machine Learning for Time-to-Event Analysis. *Clinical and Translational Science*, 11(3):305–311, 2018. doi: 10.1111/cts.12541.

[48] G. D. Garson. Interpreting Neural Network Connection Weights. *AI Expert*, 6(4):46–51, 1991.

[49] 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(277), 2020. ISSN 14712288. doi: 10.1186/s12874-020-01153-1.

[50] 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.

[51] U. B. Mogensen, H. Ishwaran, and T. A. Gerds. Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. *Journal of Statistical Software*, 50(11):1–23, 2012. ISSN 1548-7660. doi: 10.18637/jss.v050.i11.

[52] S. van Buuren. *Flexible imputation of missing data*. CRC press, 2nd edition, 2018. ISBN 9781138588318.

[53] L. Beretta and A. Santaniello. Nearest neighbor imputation algorithms: A critical evaluation. *BMC Medical Informatics and Decision Making*, 16(Suppl 3), jul 2016. ISSN 14726947. doi: 10.1186/s12911-016-0318-z.

[54] 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.

[55] P. Blanche, J. F. Dartigues, and doi = 10.1002/sim.5958 file = :C\:/Users/kanti/Dropbox/George PhD projects/Competing risks/Literature/About prediction/OK - (2013) Estimating and comparing time-dependent AUROC for CRs.pdf:pdf journal = Statistics in Medicine keywords = AUC,Competing risks,Discrimination,Inverse probability of censoring weighting,Prognosis,Survival analysis number = 30 pages = 5381–5397 title = Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks volume = 32 year = 2013 Jacqmin-Gadda, H.

[56] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models, 2019. ISSN 18785921.

[57] E. W. Steyerberg. *Clinical prediction models: A Practical Approach to Development, Validation, and Updating*. Springer, 2nd edition, 2019. doi: https://doi.org/10.1007/978-3-030-16399-0. URL https://www.springer.com/gp/book/9783030163983.

[58] D. Faraggi and R. Simon. A neural network model for survival data. *Statistics in Medicine*, 14(1):73–82, 1995. doi: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780140108.

[59] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1), 2018. doi: 10.1186/s12874-018-0482-1.

[60] T. Sun, Y. Wei, W. Chen, and Y. Ding. Genome-wide association study-based deep learning for survival prediction. *Statistics in Medicine*, 39(30):4605–4620, 2020. ISSN 10970258. doi: 10.1002/sim.8743.

[61] L. Hao, J. Kim, S. Kwon, and I. D. Ha. Deep learning-based survival analysis for high-dimensional survival data. *Mathematics*, 9(11):1–18, 2021. ISSN 22277390. doi: 10.3390/math9111244.

[62] 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.

[63] 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.

[64] 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.

[65] M. Fornili, F. Ambrogi, P. Boracchi, and E. Biganzoli. Piecewise exponential artificial neural networks (PEANN) for modeling hazard function with right censored data. *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*, pages 125–136, 2013. doi: 10.1007/978-3-319-09042-9_9.

[66] E. Biganzoli, P. Boracchi, and E. Marubini. A general framework for neural network models on censored survival data. *Neural Networks*, 15(2):209–218, 2002. doi: 10.1016/s0893-6080(01)00131-9. URL `www.elsevier.com/locate/neunet`.

[67] T. Ching, X. Zhu, and L. X. Garmire. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Computational Biology*, 14(4), apr 2018. ISSN 15537358. doi: 10.1371/journal.pcbi.1006076.

[68] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. URL `http://jmlr.org/papers/v15/srivastava14a.html`.

[69] C. Lee, J. Yoon, and M. Van Der Schaar. Dynamic-DeepHit: A Deep Learning Approach for Dynamic Survival Analysis with Competing Risks Based on Longitudinal Data. *IEEE Transactions on Biomedical Engineering*, 67(1):122–133, 2020. ISSN 15582531. doi: 10.1109/TBME.2019.2909027.

[70] D. Jarrett, J. Yoon, and M. Van Der Schaar. Dynamic Prediction in Clinical Survival Analysis Using Temporal Convolutional Networks. *IEEE Journal of Biomedical and Health Informatics*, 24(2):424–436, 2020. ISSN 21682208. doi: 10.1109/JBHI.2019.2929264.

[71] 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), 2015. ISSN 17417015. doi: 10.1186/s12916-014-0241-z. URL `http://www.biomedcentral.com/1741-7015/13/1`.

[72] 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`.

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

[74] J. D. Olden, M. K. Joy, and R. G. Death. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, 178(3-4):389–397, 2004. doi: 10.1016/j.ecolmodel.2004.03.013.

6

Survival prediction models since liver transplantation - comparisons between Cox models and machine learning techniques

This chapter is based on: G. Kantidakis, H. Putter, C. Lancia, J. 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, 2020.

Abstract

Background: Predicting survival of recipients after liver transplantation is regarded as one of the most important challenges in contemporary medicine. Hence, improving on current prediction models is of great interest. Nowadays, there is a strong discussion in the medical field about machine learning (ML) and whether it has greater potential than traditional regression models when dealing with complex data. Criticism to ML is related to unsuitable performance measures and lack of interpretability which is important for clinicians.

Methods: In this paper, ML techniques such as random forests and neural networks are applied to large data of 62294 patients from the United States with 97 predictors selected on clinical/statistical grounds, over more than 600, to predict survival from transplantation. Of particular interest is also the identification of potential risk factors. A comparison is performed between 3 different Cox models (with all variables, backward selection and LASSO) and 3 machine learning techniques: a random survival forest and 2 partial logistic artificial neural networks (PLANNs). For PLANNs, novel extensions to their original specification are tested. Emphasis is given on the advantages and pitfalls of each method and on the interpretability of the ML techniques.

Results: Well-established predictive measures are employed from the survival field (C-index, Brier score and Integrated Brier Score) and the strongest prognostic factors are identified for each model. Clinical endpoint is overall graft-survival defined as the time between transplantation and the date of graft-failure or death. The random survival forest shows slightly better predictive performance than Cox models based on the C-index. Neural networks show better performance than both Cox models and random survival forest based on the Integrated Brier Score at 10 years.

Conclusions: In this work, it is shown that machine learning techniques can be a useful tool for both prediction and interpretation in the survival context. From the ML techniques examined here, PLANN with 1 hidden layer predicts survival probabilities the most accurately, being as calibrated as the Cox model with all variables.

6.1 Introduction

Liver transplantation (LT) is the second most common type of transplant surgery in the United States after kidney [1]. Over the last decades, the success of liver transplants has improved survival outcome for a large number of patients suffering from chronic liver disease everywhere on earth [2]. Availability of donor organs is a major limitation especially when compared with the growing demand of liver candidates due to the enlargement of age limits. Therefore, improvement on current prediction models for survival since LT is important.

There is an open discussion about the value of machine learning (ML) versus statistical models (SM) within clinical and healthcare practice [3–7]. For survival data, the most commonly applied statistical model is the Cox proportional hazards regression model [8]. This model allows a straightforward interpretation, but is at the same time restricted to the proportional hazards assumption. On the other hand, ML techniques are assumption-free and data adaptive which means that they can be effectively employed for modelling complex data. In this article, the results between SM and ML techniques are assessed based on a 3-stage comparison: predictive performance for large sample size/large number of covariates, calibration (absolute accuracy) which is often neglected, and interpretability in terms of the most prognostic factors identified. Advantages and disadvantages for each method are detailed.

ML techniques need a precise set of operating conditions to perform well. It is important that a) the data have been adequately processed so that the inputs allow for good learning, b) modern method is applied using state-of-the-art programming software and c) proper tuning of the parameters is performed to avoid sub-optimal or default choices for parameters which downgrade the algorithm's performance. Danger of overfitting is associated with ML approaches (as they employ complex algorithms). A note of caution is required during model training to prevent from overfitting, e.g. the selection of suitable hyper-parameters. Needless to say, overfitting might also occur with a traditional model if it is too complex (estimation of too many parameters) thus limiting generalizability outside training instances.

Neural networks have been commonly applied in healthcare. Consequently, different approaches for time-to-event endpoints are present in the literature. Biganzoli *et al.* proposed a partial logistic regression approach of feed forward neural networks (PLANN) for flexible modelling of survival data [9]. By using the time interval as an input in a longitudinally transformed feed forward network with logistic activation and entropy error function, they estimated smoothed discrete hazards at each time interval in the output layer. This is a well known approach for modelling survival neural networks [10]. In 2000, Xiang *et al.* [11] compared the performance of 3 existing neural network methods for right censored data (the Faraggi-Simon [12], the Liestol-Andersen-Andersen [13] and a modification of the Buckley-James method [14]) with Cox models in a Monte Carlo simulation study. None of the networks outperformed the Cox models and they only performed as good as Cox for some scenarios. Lisboa *et al.* extended the PLANN approach introducing a Bayesian framework which can perform Automatic Relevance Determination for survival data (PLANN-ARD) [15]. Several applications of the PLANN and the PLANN-ARD methods can be found in the literature [16–19]. They show potential for neural networks in systems with non-linearity and complex interactions between factors. Here extensions of the PLANN approach for big LT data are examined.

The clinical endpoint of interest for this study is overall graft-survival defined as the time between LT and graft-failure or death. Predicting survival after LT is hard as it depends on many factors and is associated with donor, transplant and recipient characteristics whose importance changes over time and per outcome measure [20]. Models that combine donor and recipient characteristics have usually better performance for predicting overall graft-survival and particularly those that include sufficient donor risk factors have better performance for long-term graft survival [21]. The aims of this manuscript can be summarised as: i) potential role of ML as a competitor of traditional methods when complexity of the data is high (large sample size, high dimensional setting), ii) identification of potential risk factors using 2 ML methods (random survival forest, survival neural networks) complementary to the Cox model, iii) use of variable selection methods to compare predictive ability with the models including the non-reduced set of variables, iv) evaluation of predictions and goodness of fit, and v) clinical relevance of the findings (potential for medical applications).

The paper is organized as follows. Section ”Methods” presents details about data collection and the imputation technique, SMs and ML. Further sections discuss model training, predictive performance assessment on test data, and details about interpretability of the models. Comparisons between models based on global performance measures, prediction error curves, variable importance and calibration plots are discussed in the section ”Results”. The article is concluded by the ”Discussion” about findings, limitations of this work and future perspectives. All analyses were performed in R programming language version 3.5.3 [22]. Preliminary results were presented at 40th Annual Conference of the International Society for Clinical Biostatistics [23].

6.2 Methods

An analysis is presented on survival data after LT based on 62294 patients from the United States. Information was collected from the United Network of Organ Sharing (UNOS) [1]. After extensive pre-processing from a set of more than 600 covariates, 97 variables were included in the final dataset based on clinical and statistical considerations (see Additional file 1); 52 donor and 45 liver recipient characteristics (missing values were imputed). As the UNOS data is large in both number of observations and covariates, it is of interest to see how ML algorithms - which are able to capture naturally multi-way interactions between variables and can deal with big datasets - will perform compared to Cox models. The clinical endpoint is overall graft-survival (OGS) the time between LT and graft-failure or death). The choice for this endpoint was made for two reasons 1) it is of primary interest for clinicians and 2) it is the most appropriate outcome measure to evaluate the efficacy of LT, because it incorporates both patient mortality and survival of the graft [21].

This section is divided into different subsections including the necessary components of analyses for OGS (provided in ”Results” section). We discuss in detail both Cox models and ML techniques (Random Survival Forest, Survival Neural Networks). Elements of how the models were trained and how the predictive performance was assessed on the test data are presented. More technical details are provided in the supplementary material. We conclude this extensive section with a focus on methods to extract interpretation for the ML approaches.

6.2.1 Data collection and imputation technique

UNOS manages the Organ Procurement and Transplantation Network (OPTN) and together they collect, organise and maintain data of statistical information regarding organ transplants in the Scientific Registry of Transplant Recipients (SRTR) database [2]. SRTR gathers data from local Organ Procurement Organisations (OPO) and from OPTN (primary source). It includes data from transplantations performed in the United States from 1988 onwards. This information is used to set priorities and seek improvements in the organ donation process.

The data provided by UNOS included 62294 patients who underwent LT surgery from 2005 to 2015 (project under DUA number 9477). Standard analysis files contained 657 variables for both donors and patients (candidates and recipients). Among these, 97 candidate risk factors - 52 donor and 45 patient characteristics - were pre-selected before carrying out analysis. This resulted in a final dataset with 76 categorical and 21 continuous variables amounting to 2.2% missing data overall. The percentage of missing values for each covariate varied from 0 to 26.61% (no missing values for 26 covariates, up to 1% missingness for 51 covariates, 1 to 10% for 11 variables, 10 to 25% for 7 variables and 25 to 26.61% for only 2 variables). Analysis on the complete case would reduce

[1] UNOS is a non-profit and scientific organisation in the United States which arranges organ donation and transplantation. For more information visit its website https://unos.org.

[2] Dictionary for variables details is provided at: https://www.srtr.org/requesting-srtr-data/saf-data-dictionary/.

the available sample size from 62294 to 33394 patients leading to a huge waste of data. Furthermore, this could lead to invalid results (underestimation or overestimation of survival) if the excluded group of patients represents a subgroup from the entire sample [24]. To reconstruct the missing values the `missForest` algorithm [25] was applied for both continuous and categorical variables. This is a non-parametric imputation method that does not make explicit assumptions about the functional form of the data and builds a random forest model for each variable (500 trees were used). It specifies the model to predict missing values by using information based on the observed values. It is the most exhaustive and accurate of all random forests algorithms used for missing data imputation, because all possible variable combinations are checked as responses.

6.2.2 Cox proportional hazard regression models

In survival analysis, the focus is on the time till the occurrence of the event of interest (here graft-failure or death). The Cox proportional hazards model is usually employed to estimate the effect of risk factors on the outcome of interest [8].

Data with sample size n consist of the independent observations from the triple (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 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 as follows:

$$h(t|X) = h_0(t)\exp(X^T\boldsymbol{\beta}), \tag{6.1}$$

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 corresponding partial likelihood can be written as:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp\left(\sum_{k=1}^{p} \beta_k X_{ik}\right)}{\sum_{j\in R(t_i)} \exp\left(\sum_{k=1}^{p} \beta_k Z_{jk}\right)}, \tag{6.2}$$

where D is the set of failures, and $R(t_i)$ is the risk set at time t_i of all individuals who are still in the study at the time just before time t_i. This function is then maximised over $\boldsymbol{\beta}$ to estimate the model parameters.

Two other Cox models were employed 1) a Cox model with a backward elimination and 2) a penalised Cox regression with the Least Angle and Selection Operator (LASSO). Both models have been widely used for variable selection. We aim to compare these more parsimonious models versus a Cox model with all variables in terms of predictive performance. For the first, a numerically stable version of the backward elimination on factors was applied using a method based on Lawless and Singhal (1978) [26]. This method estimates the full model and computes approximate Wald statistics by computing conditional maximum likelihood estimates - assuming multivariate normality of estimates. Factors that require multiple degrees of freedom are dropped or retained as a group. The latter approach uses a combination of selection and regularisation [27]. Denote the log-partial likelihood by $\ell(\boldsymbol{\beta}) = logL(\boldsymbol{\beta})$. The vector $\boldsymbol{\beta}$ is estimated via the criterion:

$$\hat{\boldsymbol{\beta}} = \text{argmin}[\ell(\boldsymbol{\beta})], \quad \text{subject to} \sum_{j=1}^{p} |\beta_j| \leq s \tag{6.3}$$

with s a user specified positive parameter.

Equation (6.3) can also be rewritten as

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\text{argmin}} \left(\ell(\beta) + \lambda_{LASSO} \sum_{j=1}^{p} |\beta_j|\right). \tag{6.4}$$

The quantity $\sum_{j=1}^{p}|\beta_j|$ is also known as the L_1-norm and performs regularisation to the log-partial likelihood. The term λ_{LASSO} is a non-negative constant that assigns the amount of penalisation. Larger values for the parameter mean larger penalty to the β_j coefficients and enlarged shrinkage towards zero.

The tuning parameter s in equation (6.3) or equivalently parameter λ_{Lasso} in equation (6.4) is the controlling mechanism for the variance of the model. Higher values reduce further the variance but introduce at the same time more bias (variance-bias trade off). To find a suitable value for this parameter 5-fold cross-validation was performed to minimise the prediction error; here in terms of the cross-validated log-partial likelihood (CVPL) [28]

$$CVPL(s) = \sum_{i=1}^{n}(\ell(\hat{\beta}_{(-i)}(s)) - \ell_{(-i)}(\hat{\beta}_{(-i)}(s))), \tag{6.5}$$

where $\ell_{(-i)}(\beta)$ is the partial log-likelihood of equation (6.2) when individual i is excluded. Therefore, the term $\ell(\hat{\beta}_{(-i)}) - \ell_{(-i)}(\hat{\beta}_{(-i)})$ represents the contribution of observation i. The value that maximizes $\ell_{(-i)}(\beta_{(-i)})$ is denoted by $\hat{\beta}_{(-i)}$.

6.2.3 Random forests for survival analysis

Random Survival Forests (RSFs) are an ensemble tree method for survival analysis of right censored data [29] adapted from random forests [30]. 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) [31].

For RSFs, randomness is introduced in two ways: bootstrapping 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 most commonly used splitting criteria are the log-rank test by Segal [32] and the log-rank score test by Hothorn and Lausen [33]. Each terminal node should have at least a pre-specified number of unique events. Combining information from the $\mathcal{B}$ trees, survival probabilities and ensemble cumulative hazard estimate can be calculated using the Kaplan-Meier and Nelson-Aalen methodology, respectively.

The fundamental principle behind each survival tree is the conservation of events. It is used to define ensemble mortality, a new type of predicted outcome for survival data derived from the ensemble cumulative hazard function (comparable to the prognostic index based on the Cox model). 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 $\mathcal{H}$ [29]. 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. However, combining an ensemble of trees downgrades significantly the intuitive interpretation of a single tree.

6.2.4 Survival neural networks

Artificial neural networks (NNs) are a machine learning method able to model non-linear relationships between prognostic factors with great flexibility. These systems are inspired from biological neural networks that aimed at imitating the human brain activity [34]. A NN has a layered structure and is based on a collection of connected units called nodes or neurons which comprise a layer. The input layer picks up the signals and passes them through transformation functions to the next layer which is called ”hidden”. A network may have more than one hidden layer that connects with the previous and transmit signals towards the output layer. Connections between artificial neurons are called edges. Artificial neurons and edges have a weight (connection strength) which adjusts as learning proceeds. It increases or decreases the strength of the signal of each connection according to its sign. For the purpose of training, a target is defined, which is the observed outcome. The simplest form of a NN is the single layer feed-forward perceptron with the input layer, one hidden layer and the output layer [35].

The application of NNs has been extended to survival analysis over the years [13]. Different approaches have been considered; some model the survival probability $\mathcal{S}(t)$ directly or the unconditional probability of death $\mathcal{F}(t)$ whereas other approaches estimate the conditional hazard $h(t)$ [10]. They can be distinguished according to the method used to deal with the censoring mechanism. Some networks have k output nodes [36] - where k denotes k separate time intervals - while others have a single output node.

In this research, the method of Biganzoli was applied, which specifies a partial logistic feed-forward artificial neural network (PLANN) with a single output node [9]. This method uses as inputs the prognostic factors and the survival times to increase the predictive ability of the model. Data have to be transformed into a longitudinal format with the survival times being divided into a set of k non-overlapping intervals (months or years) $I_k = (\tau_{k-1}, \tau_k]$, with $0 = \tau_o < \tau_1 < \cdots < \tau_k$ a set of pre-defined time points. In this way, the time component of survival data is taken into consideration. On the training data, each individual is repeated for the number of intervals he/she was observed in the study and on the test data for all time intervals. PLANN provides the discrete conditional probability of dying $\mathcal{P}(T \in I_k \mid T > \tau_{k-1})$ using as transformation function of both input and output layers the logistic (sigmoid) function:

$$f(\eta) = \frac{1}{1+e^{-\eta}}, \tag{6.6}$$

where $\eta = \sum_{i=1}^{p} w_i X_i$ is the summed linear combination of the weights w_i of input-hidden layer and the input variables X_i $(i = 1, 2, \cdots, p)$.

The contribution to the log-likelihood for each individual is calculated all over the intervals one is at risk. The output node is one large target vector with 0 if the event did not occur and 1 if the event occurred in a specific time interval. Therefore, such a network first estimates the hazard for each interval $h_k = P(\tau_{k-1} < T \leq \tau_k | T > \tau_{k-1})$ and then $S(t) = \prod_{k:t_k \leq t}(1 - h_k)$.

In this work, novel extensions in the specification of the PLANN are tested. Two new transformation functions were investigated for the input-hidden layer the rectified linear unit (ReLU)

$$f(\eta) = \eta^{+} = \max(0, \eta), \tag{6.7}$$

which is the most used activation function for NNs and the hyperbolic tangent (tanh)

$$f(\eta) = \frac{1-e^{-2\eta}}{1+e^{-2\eta}}. \tag{6.8}$$

These functions can be seen as different modulators of the degree of non-linearity implied by the input and the hidden layer.

The PLANN was expanded in 2 hidden layers with same node size and identical activation functions for input-hidden 1 and hidden 1 - hidden 2 layers. The k non-overlapping intervals of the survival times were treated as k separate variables. In this way, the contribution of each interval to the predictions of the model using the relative importance method by Garson [37] and its extension for 2 hidden layers can be obtained (see subsection Interpretability of the models below and Additional file 1).

6.2.5 Model training

The split sample approach was employed; data was split randomly into two complementary parts, a training set (2/3) and a test set (1/3) under the same event/censoring proportions. To tune a model, 5-fold cross validation was performed in the training set for the machine learning techniques (and for Cox LASSO). Training data was divided into 5 folds. Each time 4 folds were used to train a model and the remaining fold was used to validate its performance and the procedure was repeated for all combination of folds. Tuning of the hyper-parameters was done using grid search and performance of final models was assessed on the test set. Analyses were performed in R programming language version 3.5.3 [22]. Package of implementation for RSFs and NNs as well as technical details regarding the choice of tuning parameters and the cross-validation procedure for each method are provided in Additional file 2.

6.2.6 Assessing predictive performance on test data

To assess the final predictive performance of the models the concordance index, the Brier score, and the Integrated Brier Score (IBS) were applied.

The most popular measure of model performance in a survival context is the concordance index [38] which computes the proportion of pairs of observations for which the survival times and model predictions order are concordant taking into account censoring. It takes values typically in the range 0.5 - 1 with higher values denoting higher ability of the model to discriminate and 0.5 indicating no discrimination. The C-index cannot be defined for neural network models since it relies on the ordering of individuals according to prognosis and there is no unique ordering between the subjects. At one year individual i may have better survival probability than individual j, but this could be reversed for a different time point.

The C-index provides a rank statistic between the observations that is not time-dependent. Following van Houwelingen and le Cessie [39] a time-dependent prediction error is defined as

$$\mathit{Brier}(y, \hat{S}(t_0|x)) = (y - \hat{S}(t_0|x))^2, \tag{6.9}$$

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. The expected value with respect to a new observation Y_{new} under the true model $S(t_0|x)$ can be written as:

$$E[\mathit{Brier}(Y_{new}, \hat{S}(t_0|x))] = S(t_0|x)(1 - S(t_0|x)) + (S(t_0|x) - \hat{S}(t_0|x))^2. \tag{6.10}$$

The Brier Score consists of two components: the ”true variation” $S(t_0|x)(1 - S(t_0|x))$ and the error due to the model $(S(t_0|x) - \hat{S}(t_0|x))^2$. A perfect prediction is only possible if $S(t_0|x) = 0$ or $S(t_0|x) = 1$. In practice the two components cannot be separated since the true $S(t_0|x)$ is unknown.

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 [40]. Then an estimate of the average prediction error of the prediction model $\hat{S}(t|x)$ at time $t = t_0$ is

$$Err_{Score}(\hat{S}, t_0) = \frac{1}{n}\sum_i 1\{d_i = 1 \vee t_i > t_0\}\frac{Score(1\{t_i > t_0\}, \hat{S}(t_0|x_i))}{\hat{C}(\min(t_i-, t_0)|x_i)} \tag{6.11}$$

In (6.11), $\frac{1}{\hat{C}(\min(t_i-, t_0)|x_i)}$ is a weighting scheme known as inverse probability of censoring weighting (IPCW) and $Score$ is the Brier Score for the prediction model. It ranges typically from 0 to 0.25 with a lower value meaning smaller prediction error.

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 $\int_0^{t_{hor}} Err_{Score}(\hat{S}, t_0)dt_0$ (here t_{hor} = 10 years) [41]. IBS provides the cumulative prediction error up to t_{hor} at all available times ($t^* = 1, 2, \cdots,$ 10 years) and takes values in the same range as the Brier score. In this study, we use IBS as the main criterion to evaluate the predictive ability of all models up to 10 years.

6.2.7 Interpretability of the models

Interpretation of models is of great importance for the medical community. It is well known that Cox models offer a straightforward interpretation through hazard ratios.

For neural networks with one hidden layer the connection weights algorithm by Garson [37] – later modified by Goh [42] – can provide information about the mechanism of the weights. The idea behind this algorithm is that inputs with larger connection weights produce greater intensities of signal transfer. As a result, these inputs will be more important for the model. Garson's algorithm can be used to determine relative importance of each input variable, partitioning the weights in the network. Their absolute values are used to specify percentage of importance. Note that the algorithm does not provide the direction of relationships, so it remains uncertain whether the relative importance indicates a positive or a negative effect. For details about the algorithm see [43]. During this work, the algorithm was extended for 2 hidden layers to obtain the relative importance of each variable (for the implementation see algorithm 1 on Additional file 1).

Random survival forest relies on two methods which can provide interpretability: variable importance (VIMP) and minimal depth [44]. The former is associated with the prediction error before and after the permutation of a prognostic factor. Large importance values indicate variables with strong predictive ability. The latter is related to the forest topology as it assesses the predictive value of a variable by computing its depth compared to the root node of a tree. VIMP is more frequently reported than minimal depth in the literature [45]. For both methods interpretation is available only for variable entities and not for each variable level.

6.3 Results

Administrative censoring was applied to the UNOS data at 10 years. Median follow-up is equal to 5.36 years (95% CI: 5.19 - 5.59 years) and it was estimated with reverse Kaplan-Meier [46]. Clinical endpoint is overall graft-survival (OGS). From the total number of patients, 69.1% was alive/censored and 30.9% experienced the event of interest (graft-failure or death). 3 models were used from the Cox family to predict survival outcome: a) a model with all 97 prognostic factors, b) a model with backward selection and c) a model based on the LASSO method for variable selection. Furthermore, 3 machine learning methods were employed: a) a random survival forest, b) a NN with one hidden layer and c) a NN with two hidden layers.

6.3.1 Comparisons between models

In this section a direct comparison of the 6 models is illustrated in terms of variable importance on the training set and predictive performance on the test set. Specification of the variables with dummy coding included 119 variable levels from the 97 potentially prognostic factors. For NNs - to apply and extend the methodology of Biganzoli - follow-up time was divided into 10 time intervals $(0, 1], (1, 2], \cdots, (9, 10]$ denoting years since transplantation. For Cox models and RSF exact time points were used.

Cox model assumes that each covariate has a multiplicative effect in the hazard function (which is constant over time). Estimating a model with 97 prognostic factors leads inevitably to a violation of the proportional hazards assumption for some covariates (17 out of 97 here). This means that hazard ratios for those risk factors are the mean effects on the outcome which is still a valuable information for the clinicians. To consider all possible non-linear effects on interactions leads to a complex model where too many parameters need to be estimated and the interpretability becomes very difficult. On the other hand, ML techniques do not make any assumptions about the data structure and therefore their performance is not affected by the violation of PH. The backward and the LASSO methods selected 28 (out of 97) and 45 predictors (out of 119 dummy coded), respectively. Selection of a smaller set of variables by Cox backward was expected, since it is a greedier (heuristic) method than LASSO penalized regression. The 12 most influential variables for the Cox model with all variables were selected by both methods (see table 6.2). 5 of these variables: *re-transplantation*, *donor type*, *log(Total cold ischemic time)*, *diabetes* and *pre-treatment status* violated the PH assumption.

5-fold cross-validation in the training data resulted in the following optimal hyper-parameters combinations for the machine learning techniques:

- For the Random Survival Forest `nodesize` = 50, `mtry` = 12, `nsplit` = 5 and `ntree` = 300. Stratified bootstrap sub-sampling of half the patients was used per tree (due to the large training time required).
- For the neural network with 1 hidden layer `activation function` = ”sigmoid” (for the input-hidden layer), `node size` = 85, `dropout rate` = 0.2, `learning rate` = 0.2, `momentum` = 0.9 and `weak class weight` = 1.
- For the neural network with 2 hidden layers `activation function` = ”sigmoid” (for the input-hidden 1 and the hidden 1-hidden 2 layers), `node size` = 110, `dropout rate` = 0.1, `learning rate` = 0.2, `momentum` = 0.9 and `weak class weight` = 1.

6.3.2 Global performance measures

The global performance measures on test data are provided in Table 6.1. Examining the Integrated Brier Score (IBS), the NNs with 1 and with 2 hidden layers have the lowest (IBS = 0.180) followed by the RSF (IBS = 0.182). Cox models have a comparable performance (IBS = 0.183). Therefore, the predictive ability of Cox backward and Cox LASSO is the same as the less parsimonious Cox model with all variables in terms of IBS. The best model in terms of C-index is the Random Survival Forest (0.622) while the Cox models with all variables has slightly worse performance. C-index for Cox backward and Cox LASSO are respectively 0.615 and 0.614.

	IBS	C-index
Cox all variables	0.183	0.620
Cox backward	0.183	0.615
Cox LASSO	0.183	0.614
RSF	0.182	**0.622**
Neural Network 1h	**0.180**	-
Neural Network 2h	**0.180**	-

Table 6.1: Integrated Brier Score (IBS) and C-index on the test data. Neural network 1h and 2h refer to a neural network with one and two hidden layers respectively.

Stability of the networks was investigated by rerunning the same models on the test data, and showed that the NN with 1 hidden layer had stable predictive performance and variable importance. In contrast, the NN with 2 hidden layers was quite unstable regarding variable importance. This behavior might be related to the vast amount of weights that had to be trained for this model which can lead to overfitting (in total 26621 connection weights were estimated for a sample size of 41530 patients in long format; whereas for the NN with 1 hidden layer 11136 connection weights). For the RSF, model obtained remarkable stability in terms of performance error after a particular number of trees (`ntree` = 300 was selected).

6.3.3 Prediction error curves

Figure 6.1 shows the average prediction Brier error over time for all models. Small differences can be observed between Cox models and RSF. The NNs with 1 hidden and with 2 hidden layers have almost identical evolution over time achieving better performance than the Cox models and the RSF.

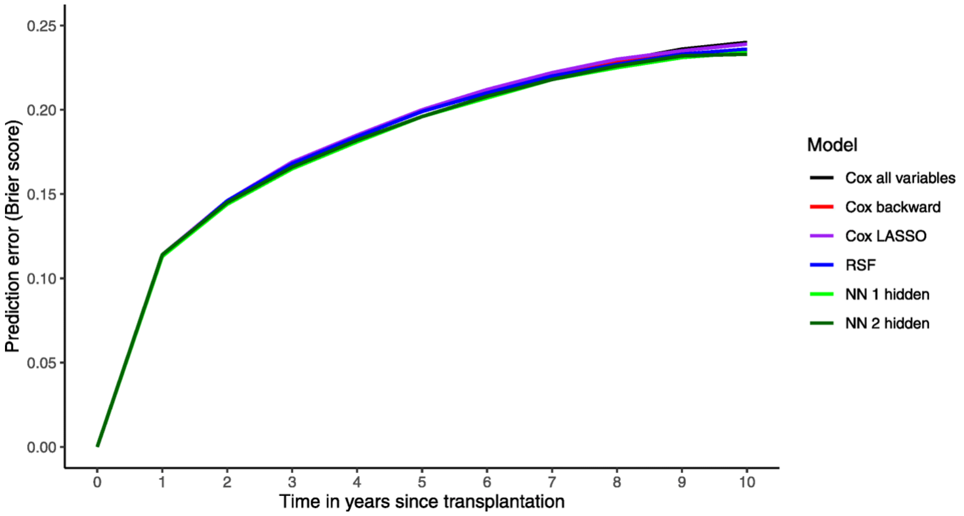


Figure 6.1: Prediction error curves for all models.

6.3.4 Variable importance

	Cox all variables	Cox backward	Cox LASSO
	HR (95% CI)	HR (95% CI)	HR
Re-transplantation	1.602 (1.491-1.721)	1.608 (1.501-1.722)	1.558
Donor age	1.010 (1.008-1.011)	1.011 (1.009-1.012)	1.009
Donor type DCD$^{(a)}$	1.483 (1.362-1.616)	1.443 (1.338-1.556)	1.298
log(Total cold ischemic time)	1.258 (1.192-1.327)	1.285 (1.221-1.353)	1.191
Diabetes	1.173 (1.125-1.225)	1.176 (1.128-1.226)	1.136
Race Black$^{(b)}$	1.240 (1.171, 1.314)	1.261 (1.193-1.332)	1.186
Life support	1.343 (1.240-1.454)	1.375 (1.272-1.487)	1.304
Recipient age	1.007 (1.005-1.009)	1.008 (1.006-1.010)	1.006
Incidental tumour	1.314 (1.202, 1.437)	1.315 (1.203-1.437)	1.203
Hypertensive bleeding	1.296 (1.185, 1.418)	1.301 (1.190-1.423)	1.214
HCV$^{(c)}$ serology status	1.147 (1.091-1.206)	1.148 (1.094-1.205)	1.166
Pre-treatment status ICU$^{(d)}$	1.240 (1.143, 1.346)	1.253 (1.160-1.354)	1.164

(a): Donor type DCD (Donor Circulatory Dead) vs DBD (Donor after Brain-Dead), (b): Race Black vs White,

(c): Chronic hepatitis C virus, (d): Intense Care Unit vs Non-hospitalised/Hospitalised

Table 6.2: Hazard ratios along with their 95% confidence intervals for the 12 most influential variables for the Cox models. Variables are presented in decreasing order according to the absolute z-score values (12.90 to 5.16) for the Cox model with all variables. Predictors shown are the most prognostic as their z-scores values correspond to low and very significant p-values. These variables were also selected by both Cox backward and Cox LASSO model which verifies their prognostic ability for Cox models.

In this section, the models are compared based on the most prognostic variables identified from the set of 97 predictors - 52 donor and 45 recipient characteristics. Hazard ratios of the 12 most prognostic variables for the Cox models are shown in Table 6.2, based on the absolute z-score values for the Cox model with all variables. The strongest predictor is *re-transplantation*. Having been transplanted before increases the hazard of graft-failure or death by more than 55%. The other most detrimental variables are *donor age* and *donor type circulatory dead*. One unit increase for donor age rises the hazard by around 1% while having received the graft from a donor circulatory versus brain-dead increases the hazard by more than 29% for all models. The rest of the factors which have an adverse effect are: *cold ischemic time*, *diabetes*, *race*, *life-support*, *recipient age*, *incidental tumour*, *spontaneous hypertensive bleeding*, *serology status of HCV* and *intense care unit before the operation*.

Neural network 1h	Rel-Imp	Neural network 2h	Rel-Imp	RSF	VIMP
Re-transplantation	0.035	Re-transplantation	0.028	Donor age	0.010
Life-support	0.025	HCV$^{(d)}$ serology status	0.025	Re-transplantation	0.009
Pre-treatment status ICU$^{(a)}$	0.023	Life-support	0.024	Life support	0.007
Donor type DCD$^{(b)}$	0.023	Donor age	0.023	HCV$^{(d)}$ serology status	0.007
Race Black$^{(c)}$	0.022	Diabetes	0.021	Pre-treatment status	0.006
HCV$^{(d)}$ serology status	0.022	Pre-treatment status ICU$^{(a)}$	0.020	Recipient age	0.004
Diabetes	0.020	Working income	0.020	Aetiology	0.003
Donor age	0.020	Race Black$^{(c)}$	0.019	log(Last serum creatinine)	0.003
Working income	0.018	Previous abdominal surgery	0.015	Functional status	0.002
Functional status Total assistance$^{(e)}$	0.017	Donor pre-recovery diuretics	0.015	log(Total cold ischemic time)	0.002
Aetiology HCV	0.017	Aetiology Cholestatic	0.011	Race	0.002
Hypertensive bleeding	0.017	Functional status Total assistance$^{(e)}$	0.015	Diabetes	0.002

(a): Intense Care Unit vs Non-hospitalised/Hospitalised (b): Donor type DCD (Donor Circulatory Dead) vs DBD (Donor after Brain-Dead),

(c): Race Black vs White, (d): Chronic hepatitis C virus, (e): Total assistance vs No assistance

Table 6.3: The 12 most prognostic factors for the neural networks with 1 and 2 hidden layers (Rel-Imp: relative importance) and for the Random Survival Forest (VIMP: variable importance). Note that the NN utilises time intervals as one of the input variables (check the contribution of time intervals in Table 1 of Additional file 1). For RSF importance is measured for each variable without distinction for each level.

In Table 6.3 the most prognostic factors for the machine learning techniques are presented. The top predictors are provided in terms of relative importance (Rel-Imp) for the PLANN models and in terms of variable importance (VIMP) for the RSF. For the NNs, the strongest predictor is *re-transplantation* (Rel-Imp 0.035 for 1 hidden and 0.028 for 2 hidden layers), which is the second strongest for the RSF (VIMP 0.009). According to the tuned RSF, the most prognostic factor for the overall graft-survival of the patient is *donor age* (VIMP 0.010).

Other strong prognostic variables for the NN with 1 hidden layer are *life support* (Rel-Imp 0.025), *intense care unit before the operation* (Rel-Imp 0.023) and *donor type circulatory dead versus brain-dead* (Rel-Imp 0.023). For the NN with 2 hidden layers other very prognostic variables are *serology status for HCV* (Rel-Imp 0.025), *life support* (Rel-Imp 0.024) and donor age (Rel-Imp 0.023).

For the RSF *life support* (VIMP 0.007), *serology status for HCV* (VIMP 0.007) and *intense care unit before the operation* (VIMP 0.006). Note that variable *total cold ischemic time* which was identified as the 4th most prognostic for the Cox model with all variables and the 10th most prognostic for random survival forest is not in the list of the 12 most prognostic for both NNs.

6.3.5 Individual predictions

In this section, the predicted survival probabilities are compared for 3 new hypothetical patients and 3 patients from the test data.

In Figure 6.2a) the patient with reference characteristics shows the best survival. The highest probabilities are predicted by the RSF and the lowest by the Cox model. The same pattern occurs for the patient that suffers from diabetes (orange lines). The patient with diabetes who has been transplanted before has the worst survival predictions. In this case the NN predicts the highest survival probabilities and the Cox model built using all the prognostic factors the lowest.

In Figure 6.2b) the estimated survival probabilities are showed by the Cox model with all variables, the tuned RSF and the tuned PLANN with 1 hidden layer for 3 patients from the test set. The first patient shows the highest survival predictions by the 3 models. The RSF provides the highest survival probabilities and the NN the lowest. The second patient experiences lower survival probabilities (orange lines) whereas the third patient shows the lowest survival probabilities overall. For the second patient the NN predicts the lowest survival probabilities over time and for the third the Cox model.

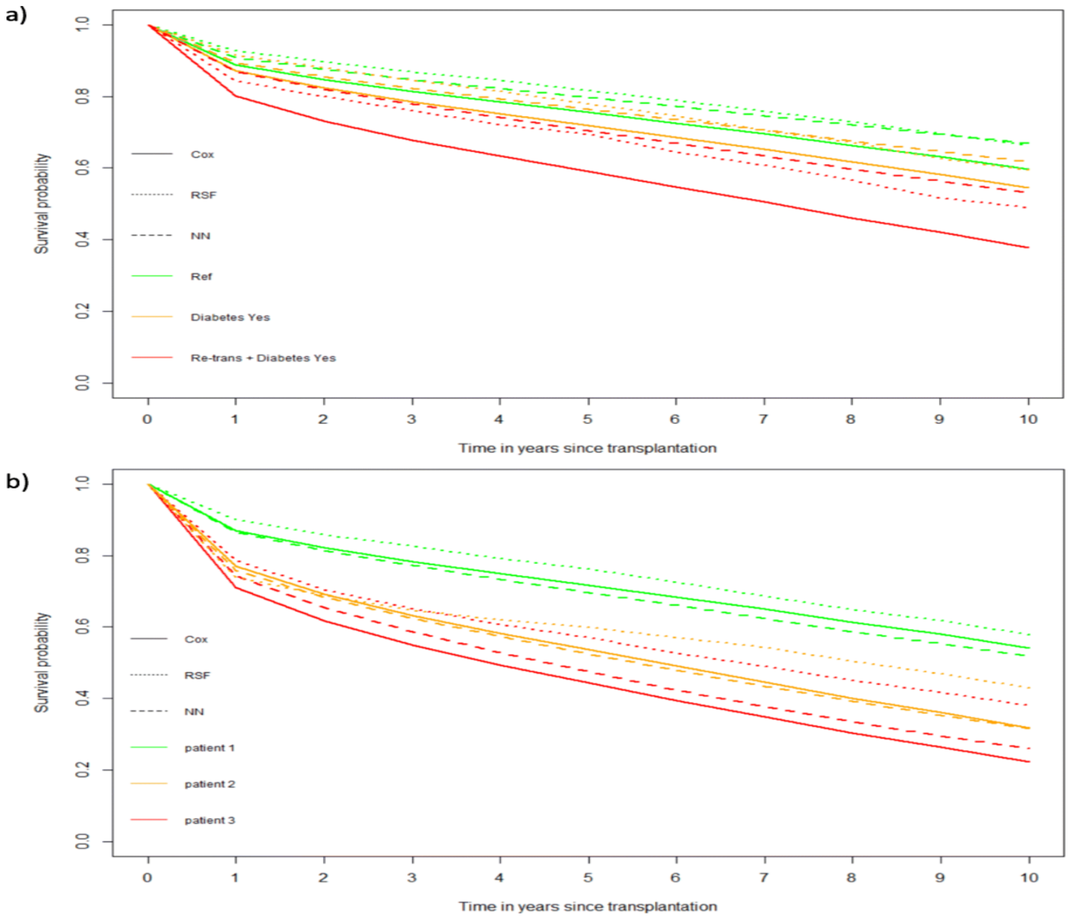


Figure 6.2: **a)** Predicted survival probabilities for 3 new hypothetical patients using the Cox model with all variables (solid lines), the tuned RSF (short dashed lines) and the tuned NN with 1 hidden layer (long dashed lines). The green lines correspond to a reference patient with the median values for the continuous and the mode value for categorical variables. The patient in the orange line has diabetes (the other covariates as in reference patient). The patient in the red line has been transplanted before and has diabetes simultaneously (the other covariates as in reference patient). Values for 10 prognostic variables for the reference patient are provided in Table 2 of Additional file 1.
b) Predicted survival probabilities for 3 patients selected from the test data based on the Cox model with all variables (solid lines), the tuned RSF (short dashed lines) and the tuned NN with 1 hidden layer (long dashed lines). Green lines correspond to a patient censored at 1.12 years. Patient in the orange line was censored at 6.86 years. Patient in the red line died at 0.12 years. Values for 10 prognostic variables for the patients are provided in Tables 3-5 of Additional file 1.

In general, the random survival forest provides the most optimistic survival probabilities whereas the most pessimistic survival probabilities are predicted by either the Cox model or the NN (more often by the Cox model). This may be related to the characteristics of the methods as RSF relies on recursive binary partitioning of predictors, whereas Cox models imply linearity, and NNs fit non-linear relationships.

6.3.6 Calibration

Here 4 methods are compared: Cox model with all variables, RSF, PLANN 1 hidden and 2 hidden layers based on the calibration on the test data. For each method, the predicted survival probabilities at each year are estimated and the patient data are split into 10 equally sized groups based on the deciles of the probabilities. Then the survival probabilities along with their 95% confidence intervals are calculated using the Kaplan-Meier methodology [47].

In figure 6.3 the results are showed at 2 years since LT. The Cox model with all variables and the PLANN with 1 hidden layer are both well calibrated. The RSF and the PLANN with 2 hidden layers tend to overestimate the survival probabilities for the patients at higher risk. Survival neural network with 1 hidden layer seems to be the most reliable for predictions between the ML techniques. Calibration plots at 5 and 10 years can be found in Additional file 3.

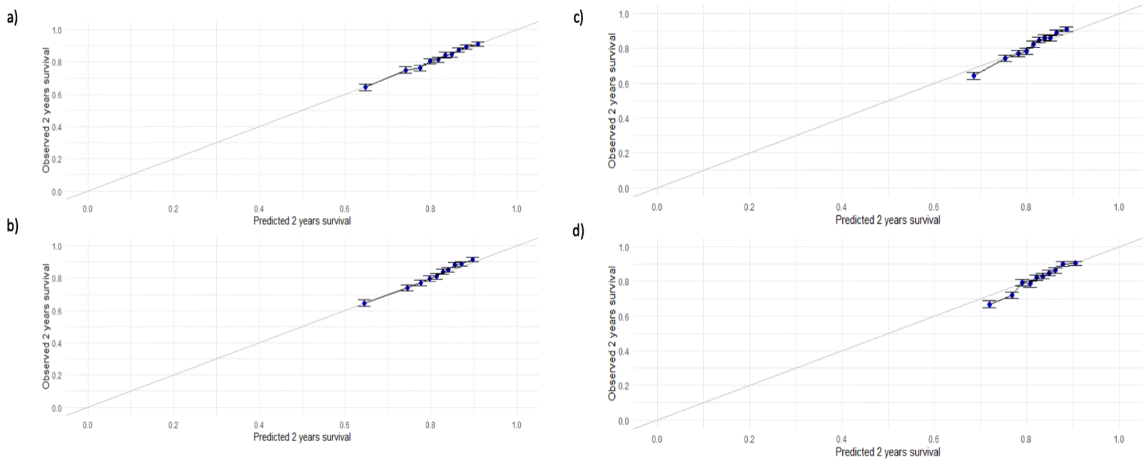


Figure 6.3: Calibration plots at 2 years on the test data: **a)** Cox model with all variables, **b)** Random Survival Forest, **c)** Partial Logistic Artificial Neural Network with 1 hidden layer, **d)** Partial Logistic Artificial Neural Network with 2 hidden layers.

6.4 Discussion

With the rise of computational power and technology on the 21^{st} century, more and more data have been collected in the medical field to identify trends and patterns which will allow building better allocation systems for patients, provide more accurate prognosis and diagnosis as well as more accurate identification of risk factors. During the past few years, machine learning (ML) has received increased attention in the medical area. For instance, in the area of LTs graft failure or primary non-function might be predicted at decision time with ML methodology [48]. Briceño *et al.* created a NN process for donor-recipient matching specifying a binary classification survival output (recipient or graft survival) to predict 3-month graft mortality [49].

In this study statistical and ML models were estimated for patients from the US post-transplantation. Random survival forest performed better than Cox models with respect to the C-index. This shows the ability of the model to discriminate between low and high risk groups of patients. The C-index was not estimated for NN because a natural ordering of subjects is not feasible. Therefore, the Brier score was measured each year for all methods. The RSF showed similar results to the Cox models having slightly smaller total prediction error (in terms of IBS). The NNs performed in general better than the Cox models or the RSF and had very similar performance over time. RSF and survival NN are ML techniques which have a different learning approach and model non-linear relationships between variables automatically. Both methods may be used in medical application but should be applied at present as additional analysis for comparison.

Special emphasis was given on the interpretation of the models. An indirect comparison was performed to examine which are the most prognostic variables for a Cox model with all variables, a RSF and NNs. Results showed that Cox model with all variables (via absolute z-score values) and the NNs with one/two hidden layer(s) (via relative importance) identified similar predictors. Both methods identified *re-transplantation* as the strongest predictor and *donor age*, *diabetes*, *life support* and *race* as relatively strong predictors. According to RSF, the most prognostic variables were *donor age*, *re-transplantation*, *life support* and *serology status of HCV*. *Aetiology* and *last serum creatinine* were selected as the 7^{th} and the 8^{th} most prognostic. This raises a known concern about the RSF bias towards continuous variables and categorical variables with multiple levels [50] (*aetiology* has 9 levels: metabolic, acute, alcoholic, cholestatic, HBV, HCV, malignant, other cirrhosis, other unknown). As continuous and multilevel variables incorporate larger amount of information than categorical, they tend to be favoured by the splitting rule of the forest during binary partitioning. Such bias was reflected in the variable importance results.

When comparing statistical models with machine learning techniques with respect to interpretability, Cox models offer a straightforward interpretation through the hazard ratios. On the contrary, for both neural networks and random survival forests the sign of the prediction is not provided (if the effect is positive or negative). Additionally, for NNs interpretation is possible for different variable levels (with the method of Garson and its extension), whereas for RSF only the total effect of a variable is shown. There is no common metric to directly compare Cox models with ML techniques in terms of interpretation. Future research in this direction is needed.

ML techniques are inherently based on mechanisms introducing randomisation and therefore very small changes are expected between different iterations of the same algorithm. To evaluate stability of performance, ML models were run several times under the same parametrisation. RSF were consistently stable after a certain number of trees (300 were selected). This was not the case for the NNs where instability is a common problem. It is challenging to tune a NN due to many hyper-parameter combinations available and the lack of a consistent global performance measure for survival data. IBS was used to tune the novel NNs, which may be the reason of instability for the NN with 2 hidden layers together with the large number of weights. Note also that the NN with 1 hidden layer is well calibrated whereas the NN with 2 hidden layers is less calibrated on the test data.

This is the first study where ML techniques are applied to transplant data where a comparison with the traditional Cox model was investigated. To construct the survival NN, the original problem had to be converted into a classification problem where exact survival times were transformed into (maximum) 10 time intervals denoting years since transplantation. On the other hand, for the Cox models and the RSF exact time to event was used. Recently, a new feed forward NN has been proposed for omics data which calculates directly a proportional hazards model as part of the output node using exact time information [51]. A survival NN with exact times may lead to better predictive performance. For UNOS data, 69.1% of the recipients were alive/censored and 30.9% had the event of interest. Results above were based on these particular percentages for censoring and events (for the NNs the percentages varied because of the reformulation of the problem).

It might be useful to investigate how the number of variables affects the performance of the models. Here 97 variables were pre-selected supported by clinical and statistical reasons (e.g. variables available before or during LT). It might be interesting to repeat the analyses on a smaller group of predictors, implementation time can be drastically reduced as the calculation complexity depends on sample size and predictors multiplicity. Alongside, predictive accuracy might be increased as some noisy factors will be removed from the dataset increasing the signal of potentially prognostic variables.

Both traditional Cox models and PLANNs allow for the inclusion of time-dependent covariates. For PLANNs, each patient is replicated multiple times during the transformation of exact times into a set of k non-overlapping intervals in long format. Thus, different values of a covariate can be naturally incorporated to increase the predictive ability of the networks. It would be interesting to apply and compare the predictive ability of time-dependent Cox models and PLANNs to liver transplantation data including explanatory variables whose values change over time. Such extension to more dynamic methods may increase predictive performance and help in decision making.

6.5 Conclusions

There is an increased attention to ML techniques beyond SM in the medical field with methods and applications being more necessary than ever. Utilization of these algorithmic approaches can lead to pattern discovery in the data promoting fast and accurate decision making. For time-to-event data, more ML techniques may be applied for prediction such as Support Vector Machines and Bayesian Networks. Moreover, deep learning with NN is gaining more and more attention and will likely be another trend in the future for these complex data.

In this work two alternatives to the Cox model from machine learning for medical data with large total sample size (62294 patients) and many predictors (97 in total) were discussed. RSF showed better performance than the Cox models with respect to C-index so it can be a useful tool for prioritisation of particular high risk patients. NNs showed better prediction performance in terms of Integrated Brier score. However, both ML techniques required a non-trivial implementation time. Cox models are preferable in terms of straightforward interpretation and fast implementation. Our study suggests that some caution is required when ML methods are applied to survival data. Both approaches can be used for exploratory and analysis purposes as long as the advantages and the disadvantages of the methods are presented.

List of abbreviations

BS, Brier score; CVPL, cross-validated log-partial likelihood; DCD, Donor Circulatory Dead; HBV, Chronic hepatitis B virus; HCV, Chronic hepatitis C virus; IBS, Integrated Brier score; IPCW, Inverse Probability of Censoring Weighting; LASSO, least angle and selection operator; LT, liver transplantation; LUMC, Leiden University Medical Center; ML, machine learning; NN(s), artificial neural network(s); OGS, overall graft-survival; OPO, Organ Procurement Organisations; OPTN, Organ Procurement and Transplantation Network; PLANN, partial logistic artificial neural network; PLANN-ARD, partial logistic artificial neural network - automatic relevance determination; PH, proportional hazards; Rel-Imp, relative importance; RSF, random survival forest; SM, statistical model; SRTR, Scientific Registry of Transplant Recipients; UNOS, United Network of Organ Sharing; VIMP, variable importance.

Declarations

Availability of data and materials

The research data for this project is private. Unauthorized use is a violation of the terms of the Data Use Agreement with the U.S. Department of Health and Human Services. More information and instructions for researchers to request UNOS data can be found at https://unos.org/data/. R-code developed to perform the analysis is available at https://github.com/GKantidakis/Survival-prediction-models-since-liver-transplantation.

Competing interests

The data reported here have been supplied by the Minneapolis Medical Research Foundation (MMRF) as the contractor for the Scientific Registry of Transplant Recipients (SRTR). The interpretation and reporting of these data are the responsibility of the author(s) and in no way should be seen as an official policy of or interpretation by the SRTR or the U.S. Government.

This study used data from the Scientific Registry of Transplant Recipients (SRTR). The SRTR data system includes data on all donor, wait-listed candidates, and transplant recipients in the US, submitted by the members of the Organ Procurement and Transplantation Network (OPTN). The Health Resources and Services Administration (HRSA), U.S. Department of Health and Human Services provides oversight to the activities of the OPTN and SRTR contractors.

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 as well as from the EORTC Cancer Research Fund (ECRF). 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

The authors would like to thank the United Network of Organ Sharing (UNOS) and Scientific Registry of Transplant Recipients (SRTR) for providing the data about liver transplantation to Leiden University Medical Center (LUMC) under DUA number 9477.

Online supplementary materials

The Additional files of this Chapter are available online at https://github.com/GKantidakis/Thesis_supplementary_materials/tree/main/Chapter6.

References

[1] J. M. Grinyó. Why is organ transplantation clinically important? *Cold Spring Harbor Perspectives in Medicine*, 3(6), 2013. doi: 10.1101/cshperspect.a014985.

[2] R. M. Merion, D. E. Schaubel, D. M. Dykstra, R. B. Freeman, F. K. Port, and R. A. Wolfe. The survival benefit of liver transplantation. *American Journal of Transplantation*, 5(2):307–313, 2005. doi: 10.1111/j.1600-6143.2004.00703.x.

[3] X. Song, A. Mitnitski, J. Cox, and K. Rockwood. Comparison of machine learning techniques with classical statistical models in predicting health outcomes. *Stud Health Technol Inform.*, 107(Pt 1):736–740, 2004.

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

[5] K. Shailaja, B. Seetharamulu, and M. A Jabbar. Machine Learning in Healthcare: A Review. In *Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 910–914, Coimbatore, 2018. doi: 10.1109/ICECA.2018.8474918.

[6] I. A. Scott, D. Cook, E. W. Coiera, and B. Richards. Machine learning in clinical practice: prospects and pitfalls. *Medical Journal of Australia*, 211(5):203–205, sep 2019. ISSN 13265377. doi: 10.5694/mja2.50294.

[7] R. J. Desai, S. V. Wang, M. Vaduganathan, T. Evers, and S. 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, 2020. doi: 10.1001/jamanetworkopen.2019.18962.

[8] 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.

[9] 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):1169–86, 1998. doi: 10.1002/(sici)1097-0258(19980530)17:10<1169::aid-sim796>3.0.co;2-d.

[10] P. Wang, Y. Li, and C. K. Reddy. Machine Learning for Survival Analysis: A Survey. *ACM Computing Surveys*, 51(6), 2019. doi: 10.1145/3214306.

[11] A. Xiang, P. Lapuerta, A. Ryutov, J. Buckley, and S. Azen. Comparison of the performance of neural network methods and Cox regression for censored survival data. *Computational Statistics & Data Analysis*, 34(2):243–257, 2000. doi: https://doi.org/10.1016/S0167-9473(99)00098-5. URL www.elsevier.com/locate/csda.

[12] D. Faraggi and R. Simon. A neural network model for survival data. *Statistics in Medicine*, 14(1):73–82, 1995. doi: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780140108.

[13] Liestøl K., Andersen P. K., and Andersen U. Survival analysis and neural nets. *Statistics in Medicine*, 13 (12):1189–1200, 1994. doi: 10.1002/sim.4780131202.

[14] J. Buckley and I. James. Linear regression with censored data. *Biometrika*, 66(3):429–436, 1979. doi: https://doi.org/10.1093/biomet/66.3.429.

[15] 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.

[16] E. Biganzoli, P. Boracchi, and E. Marubini. A general framework for neural network models on censored survival data. *Neural Networks*, 15(2):209–18, 2002. doi: 10.1016/s0893-6080(01)00131-9. URL `www.elsevier.com/locate/neunet`.

[17] A. Biglarian, E. Bakhshi, A. R. Baghestani, M. R. Gohari, M. Rahgozar, and M. Karimloo. Nonlinear survival regression using artificial neural network. *Journal of Probability and Statistics*, 2013, 2013. doi: https://doi.org/10.1155/2013/753930.

[18] A. S. Jones, A. G. F. Taktak, T. R. Helliwell, J. E. Fenton, M. A. Birchall, D. J. Husband, and A. C. Fisher. An artificial neural network improves prediction of observed survival in patients with laryngeal squamous carcinoma. *European Archives of Oto-Rhino-Laryngology*, 263(6):541–547, jun 2006. doi: 10.1007/s00405-006-0021-2.

[19] A. Taktak, L. Antolini, M. Aung, P. Boracchi, I. Campbell, B. Damato, E. Ifeachor, N. Lama, P. Lisboa, C. Setzkorn, V. Stalbovskaya, and E. Biganzoli. Double-blind evaluation and benchmarking of survival models in a multi-centre study. *Computers in Biology and Medicine*, 37(8):1108–1120, 2007. doi: 10.1016/j.compbiomed.2006.10.001.

[20] J. J. Blok, H. Putter, H. J. Metselaar, R. J. Porte, F. Gonella, J. De Jonge, A. P. Van den Berg, J. Van Der Zande, J. D. De Boer, B. Van Hoek, and A. E. Braat. Identification and validation of the predictive capacity of risk factors and models in liver transplantation over time. *Transplantation Direct*, 4(9), 2018. doi: 10.1097/TXD.0000000000000822.

[21] J. D. de Boer, H. Putter, J. J. Blok, I. P. J. Alwayn, B. van Hoek, and A. E. Braat. Predictive Capacity of Risk Models in Liver Transplantation. *Transplantation Direct*, 5(6):e457, 2019. doi: 10.1097/TXD.0000000000000896.

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

[23] G. Kantidakis, C. Lancia, and M. Fiocco. *Prediction models for liver transplantation – comparisons between Cox models and machine learning techniques [abstract OC30-4]*. 40th Annual Conference of the International Society for Clinical Biostatistics, 2019. URL `https://kuleuvencongres.be/iscb40/images/iscb40-2019-e-versie.pdf`.

[24] S. Van Buuren, H. C. Boshuizen, and D. L. Knook. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18(6):681–694, 1999. doi: 10.1002/(SICI)1097-0258(19990330)18:6<681::AID-SIM71>3.0.CO;2-R.

[25] 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.

[26] J. F. Lawless and K. Singhal. Efficient Screening of Nonnormal Regression Models. *Biometrics*, 34(2): 318–327, jun 1978. doi: 10.2307/2530022. URL https://www.jstor.org/stable/2530022?origin=crossref.

[27] R. Tibshirani. The lasso method for variable selection in the Cox model. *Statistics in medicine*, 16(4): 385–395, 1997. URL https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4{%}3C385::AID-SIM380{%}3E3.0.CO;2-3.

[28] P. J. M. Verweij and H. C. Van Houwelingen. Cross-validation in survival analysis. *Statistics in Medicine*, 12(24):2305–2314, dec 1993. doi: 10.1002/sim.4780122407. URL http://doi.wiley.com/10.1002/sim.4780122407.

[29] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *Annals of Applied Statistics*, 2(3):841–860, 2008. doi: 10.1214/08-AOAS169.

[30] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 08856125. doi: 10.1023/A:1010933404324. URL http://link.springer.com/10.1023/A:1010933404324.

[31] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, 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.

[32] M. R. Segal. Regression Trees for Censored Data. *Biometrics*, 44(1):35–47, 1988. URL http://www.jstor.org/stable/2531894.

[33] T. Hothorn and B. Lausen. On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis*, 43(2):121–137, 2003. doi: 10.1016/S0167-9473(02)00225-6.

[34] M. van Gerven and S. Bohte. Editorial: Artificial Neural Networks as Models of Neural Information Processing. *Frontiers in Computational Neuroscience*, 11:114–114, 2017. doi: 10.3389/fncom.2017.00114. URL http://journal.frontiersin.org/article/10.3389/fncom.2017.00114/full.

[35] M. Minsky and S. Papert. *Perceptrons; an introduction to computational geometry*. MIT Press, Cambridge, MA, 1 edition, 1969. ISBN 9780262130431.

[36] P. Lapuerta, Azen S. P., and LaBree L. Use of neural networks in predicting the risk of coronary artery disease. *Computers and Biomedical Research*, 28(1):38–52, 1995. doi: 10.1006/cbmr.1995.1004.

[37] G. D. Garson. Interpreting Neural Network Connection Weights. *AI Expert*, 6(4):46–51, 1991.

[38] 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.

[39] 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.

[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):2529–45, 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] J. C. van Houwelingen and H. Putter. *Dynamic prediction in clinical survival analysis*. CRC Press, 2012. ISBN 9781439835333. URL https://www.crcpress.com/Dynamic-Prediction-in-Clinical-Survival-Analysis/van-Houwelingen-Putter/p/book/9781439835333.

[42] A. T. C. Goh. Back-propagation neural networks for modeling complex systems. *Artificial Intelligence in Engineering*, 9(3):143–151, jan 1995. ISSN 0954-1810. doi: 10.1016/0954-1810(94)00011-S. URL https://www.sciencedirect.com/science/article/pii/095418109400011S.

[43] J. D. Olden and D. A. Jackson. Illuminating the ”black box”: a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, 154(1-2):135–150, 2002.

[44] H. Ishwaran, U. B. Kogalur, E. Z. Gorodeski, A. J. Minn, and M. S. Lauer. High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, 105(489):205–217, 2010. doi: 10.1198/jasa.2009.tm08622.

[45] H. Ishwaran and M. Lu. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in Medicine*, 38(4):558–582, 2019. doi: 10.1002/sim.7803.

[46] M. Schemper and T. L. Smith. A Note on Quantifying Follow-up in Studies of Failure Time. *Control Clin Trials*, 17(4):343–6, 1996. doi: 10.1016/0197-2456(96)00075-x.

[47] 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.

[48] L. Lau, Y. Kankanige, B. Rubinstein, R. Jones, C. Christophi, V. Muralidharan, and J. Bailey. Machine-Learning Algorithms Predict Graft Failure After Liver Transplantation. *Transplantation*, 101(4):e125–e132, apr 2017. ISSN 0041-1337. doi: 10.1097/TP.0000000000001600. URL http://insights.ovid.com/crossref?an=00007890-201704000-00025.

[49] J. Briceño, M. Cruz-Ramírez, M. Prieto, M. Navasa, J. O. De Urbina, R. Orti, M. Á. Gómez-Bravo, A. Otero, E. Varo, S. Tomé, G. Clemente, R. Bañares, R. Bárcena, V. Cuervas-Mons, G. Solórzano, C. Vinaixa, Á. Rubín, J. Colmenero, A. Valdivieso, R. Ciria, C. Hervás-Martínez, and M. De La Mata. Use of artificial intelligence as an innovative donor-recipient matching model for liver transplantation: Results from a multicenter Spanish study. *Journal of Hepatology*, 61(5):1020–1028, 2014. ISSN 16000641. doi: 10.1016/j.jhep.2014.05.039.

[50] W. Y. Loh and Y. S. Shih. Split selection methods for classification trees. *Statistica Sinica*, 7:815–840, 1997. URL https://www.jstor.org/stable/24306157.

[51] T. Ching, X. Zhu, and L. X. Garmire. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLOS computational biology*, 14(4), 2018. doi: 10.1371/journal.pcbi.1006076.

7

A simulation study to compare the predictive performance of survival neural networks with Cox models for clinical trial data

This chapter is based on: 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*, 2021:2160322, 2021.

Abstract

Background: Studies focusing on prediction models are widespread in medicine. There is a trend in applying machine learning (ML) by medical researchers and clinicians. Over the years, multiple ML algorithms have been adapted to censored data. However, the choice of methodology should be motivated by the real-life data and their complexity. Here, the predictive performance of ML techniques is compared with statistical models in a simple clinical setting (small/moderate sample size and small number of predictors) with Monte-Carlo simulations.

Methods: Synthetic data (250 or 1000 patients) were generated that closely resembled 5 prognostic factors pre-selected based on a European Osteosarcoma Intergroup study (MRC BO06/EORTC 80931). Comparison was performed between 2 partial logistic artificial neural networks (PLANNs) and Cox models for 20, 40, 61, and 80% censoring. Survival times were generated from a log-normal distribution. Models were contrasted in terms of the C-index, Brier score at 0-5 years, integrated Brier score (IBS) at 5 years, and miscalibration at 2 and 5 years (usually neglected). The endpoint of interest was overall survival.

Results: PLANNs original/extended were tuned based on the IBS at 5 years and the C-index, achieving a slightly better performance with the IBS. Comparison with Cox models showed that PLANNs can reach similar predictive performance on simulated data for most scenarios with respect to the C-index, Brier score, or IBS. However, Cox models were frequently less miscalibrated. Performance was robust in scenario data where censored patients were removed before 2 years or curtailing at 5 years was performed (on training data).

Conclusions: Survival neural networks reached a comparable predictive performance with Cox models but were generally less well calibrated. All in all, researchers should be aware of burdensome aspects of ML techniques such as data preprocessing, tuning of hyperparameters, and computational intensity that render them disadvantageous against conventional regression models in a simple clinical setting.

7.1 Introduction

Survival analysis (also called time-to-event analysis) is used to estimate the lifespan of a particular population under study. The most common problem that survival analysis addresses is right censoring; a form of missing data in which the time to event is not observed due to follow-up interruption before experiencing the event of interest or time limitations such as study termination (administrative censoring). The most popular statistical model (SM) for right censored data in clinical research is the Cox proportional hazards (PH) model [1] which is a semi-parametric as it makes a parametric assumption regarding the link of the predictors with the hazard function (PH), but it does not prespecify any distribution for the baseline hazard. Parametric regression methods for survival data include for instance models with the exponential, Weibull or log-normal distribution of survival time [2, 3].

The number of studies that focus on prediction models is rapidly expanding in the medical field. Furthermore, there is an increased interest in applying machine learning (ML) for prediction by medical researchers and clinicians [4]. Several ML algorithms have been developed and adapted to deal with censoring, as indicated in a recent comprehensive survey by Wang *et al.* in 2019 [5]. Choice of the appropriate methodology should be motivated by the available real-life data and their complexity. SMs 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 linearity and additivity are the dominant ways that predictors are associated with the outcome. 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 non-linearity and non-additivity are expected to be strong [6]. SM typically operate under a specific set of assumptions such as proportionality of hazards for the Cox model, whereas ML algorithms are data driven (non-parametric) without imposing any restrictions in the data structure.

Artificial neural networks (ANNs) have been one of the most widely used ML techniques in healthcare. Hence, over the years researchers have adapted them to time-to-event data [7–11]. A popular approach in the literature is that of Biganzoli *et al.* who proposed a partial logistic artificial neural network (PLANN) for flexible modelling of survival data [9]. The authors specified the time (in intervals) as an input feature in a longitudinally transformed feed forward network with logistic activation and entropy error function to estimate (smoothed) discrete hazards in the output layer for each time interval. A few years later, Lisboa *et al.* extended PLANN introducing a Bayesian framework that performs Automatic Relevance Determination (PLANN-ARD) [12]. PLANN and PLANN-ARD have been applied several times [13–17]. PLANN methodology has been developed for competing risks (PLANNCR) [18], and has also been employed under a Bayesian regularization framework (PLANNCR-ARD) [19]. Extensions of the PLANN in terms of architecture (i.e hyperparameters, activation functions, time interval specification) were recently discussed by Kantidakis *et al.* [20].

ML techniques are omnipresent in medicine as they can deal with complex data with many observations and different types of predictors (e.g. clinical and molecular) because of their data driven nature. In the previous study of our group [20], PLANN extended was developed and validated for complex liver transplantation data with large sample size and within a high dimensional setting (62294 patients, 97 risk factors). The method was compared with Cox models showing that it can be a useful tool for both prediction and interpretation. However, it is not uncommon to have a small number of patients recruited in clinical trials and a limited set of predictive features, for instance in cancer trials such as head and neck or sarcoma. Even so, there is an expectation by clinicians that ML models may perform better than SMs. Therefore, in this work, the focus is on ML techniques versus SM for non-complex clinical data to investigate a different real-life setting. A Monte-Carlo simulation study is performed to compare PLANN original or extended [9, 20] with Cox PH models for right censored survival data in terms of prediction. Hereto, real-life clinical data is mimicked to simulate synthetic data (5 predictors, 250 or 1000 observations) and to address different scenarios which are representative of the real disease (bone sarcoma). The dataset originates from a randomized phase III European Osteosarcoma Intergroup (EOI) study that investigated the effect of dose intense chemotherapy in patients with localised extremity osteosarcoma [21]. The endpoint of interest is overall survival (OS) defined as the time to death from any cause since the date of surgery.

The aims of this manuscript can be summarized as: i) investigation of the role of ML as a competitor of traditional methods for right-censored survival data in a simple setting using simulations (low-dimensional data with linear and additive dependence relations over covariates and time, small to medium sample size), ii) systematic evaluation of model predictive performance for two ML techniques (PLANNs) regarding discrimination and calibration for a number of scenarios (different censoring, sample size), iii) investigation of robustness for PLANN original (Biganzoli *et al.*) and PLANN extended (Kantidakis *et al.*) in scenarios with less observations or less information available (due to data truncation), and iv) practical relevance of findings.

The paper is organized as follows. In Section ”Methods”, details are presented about the clinical trial data and the simulation procedure. Further sections discuss the Cox model and the two ANNs, model training, and how the predictive performance was evaluated in simulated data. Section ”Results” presents neural networks tuned based on different measures, compares the predictive performance of all models, and examines the impact of scenarios for their predictive ability. The article ends with a ”Discussion” about findings, and advantages/disadvantages of the methods with respect to this particular clinical setting. All analyses were performed in R programming language version 4.0.1 [22].

7.2 Methods

This section is divided into different subsections with the necessary elements of this work. The clinical data and the simulation procedure are presented. Both Cox models and SNNs (PLANN original and PLANN extended) are discussed and it is described how the models were trained. This extensive section is concluded with the performance measures that were used to evaluate the predictive ability of the models. More technical details are provided in the supplementary material.

7.2.1 Clinical data and imputation technique

Osteosarcoma is the most common primary bone malignancy, and the third most frequent cancer in adolescents (only lymphomas and brain tumours are more prevalent) [23, 24]. In the 1970s, the introduction of adjuvant chemotherapy (administered after surgery) in the treatment of the disease increased survival rates dramatically with a current 5-year overall-survival (OS) rate above 65%. There are no significant advances in the treatment of the disease over the last 10+ years. Received dose, dose intensity and survival of chemotherapy have been investigated without evidence of difference in overall or progression-free survival [21, 25].

For this project, data was collected from a randomized controlled phase III trial of the EOI between 1993 and 2002 that investigated the effect of intensified chemotherapy on the OS of non-metastatic extremity osteosarcoma patients (MRC BO06/EORTC 80931). Treatment arm was randomly allocated to 497 eligible patients who had no prior chemotherapy before trial entry and were up to 40 years old. Treatment arms included the combination of cisplatin and doxorubicin (conventional or dose intense schedule with identical total doses). Surgery was planned at 6 weeks for both arms. The conventional two drug regimen (Regimen-C) consisted of six 3-week cycles with surgery planned between cycles 2 and 3. The dose intense regimen (Regimen-DI) consisted of six 2-week cycles with surgery planned between cycles 3 and 4. Results of the trial showed no evidence of difference in OS (primary outcome) between the two treatment arms, despite the statistically significant increase in histological response rate [21].

Five variables were pre-selected based on clinical reasoning: a) importance of particular prognostic factors in medical literature regarding osteosarcoma [21, 24, 26], b) clinical input from Leiden University Medical Center (LUMC). These were 4 categorical variables *treatment arm* (Regimen-C, Regimen-DI), *sex* (female, male), *histological response* (poor $\leq$ 90% tumour necrosis or good $>$ 90% tumour necrosis), and *excision of margin* (unknown/incomplete or complete). There was only one continuous variable *age at the date of surgery*. The clinical endpoint was OS defined as the time to death from any cause since the date of surgery. Note that only patients for whom surgery was performed after completing 2 cycles (in conventional arm) or 3 cycles (in dose-intense arm) were included. According to the study protocol, surgery was performed around 6 weeks since randomization in both treatment arms. Nevertheless, for 56 patients surgery was substantially delayed for more than 90 days due to toxicity, or it was never performed (28 patients each, respectively). These were excluded as well as 19 patients that did not fulfil the eligibility criteria or information was totally missing. Overall, 422 patients were included in the dataset.

Follow-up survival times ranged from 0.16 to 10.31 years with a median follow-up of 5.06 years (95% CI 4.45 - 5.60) estimated with reverse Kaplan-Meier [27]. There were 161/422 deaths (61.85% censoring). The dataset contained 4% missing data with 355/422 complete cases (84.1%) for the 5 variables. A visual overview of missing values is provided in Additional file 1. More specifically, there were missing values for two categorical variables: *histological response* (51/422, 12.1%) and *excision of margin* (67/422, 15.9%). The `missForest` algorithm was applied to reconstruct the missing values in order to make full use of the original data avoiding any waste of data (single imputation) [28]. This is a non-parametric imputation method which does not make assumptions about the data structure. A random forest is built for each variable with missing values (1000 trees were used to produce a stable model), testing all possible variable combinations as responses. It is the most exhaustive / accurate random forest algorithm for missing data. Poor *histological response* was imputed 27 times and good 24 times (242 poor vs 180 good in the final dataset). Unknown/incomplete *excision of margin* was imputed once and complete 66 times (49 unknown/incomplete vs 373 complete in the final dataset). The frequencies of the other 2 categorical variables were as follows: *drug regimen* (203 Regimen-C vs 219 Regimen-DI), *sex* (164 female vs 258 male). Mean *age at the date of surgery* was 16.15 years (range 3.60 - 40.85 years).

7.2.2 Simulations

This study is reported based on guidelines for simulation research in healthcare [29, 30]. The simulation procedure was repeated $B = 1000$ times to generate $N_1 = 250$ or $N_2 = 1000$ synthetic patients per dataset. Simulated data closely resembled the original osteosarcoma data described in ”Clinical data and imputation technique” following a 4-step approach:

1. Combinations were counted for the 4 categorical variables. As each variable consisted of 2 levels this led to $2^4 = 16$ unique combinations presented in Table 7.1. For all combinations, mean and standard deviation was calculated for variable *Age*.

2. Data was independently simulated according to the proportion of the occurrence of the 16 combinations in the original dataset. *Age* was sampled from a normal distribution with the mean and standard deviation determined by the combination.

3. Coefficients for the covariates were obtained with a log-normal regression in the original data [31]. These were then used to simulate survival times from a log-normal distribution. Survival time generation can be written as:

$$\log(T) = \mu + \beta^T \mathbf{x} + \sigma\epsilon, \tag{7.1}$$

   where T are the simulated survival times, μ is the intercept (part of coefficients), β is the vector of estimated coefficients for the 5 predictors in the original data, $\mathbf{x}$ is the covariate matrix for a given simulated dataset, σ scale parameter (part of coefficients), and ϵ random error with $\epsilon \sim N(0, 1)$.

4. Censoring times were generated with a Weibull distribution [32] with parameters (shape and scale) to create 20%, 40%, 61% (close to original data), or 80% censoring.

Treatment	Sex	Histological Response	Excision	Proportion	Mean age (sd)
Regimen-C	female	poor	unknown/incomplete	0.02	12.05 (4.44)
Regimen-C	female	poor	complete	0.12	17.04 (7.35)
Regimen-C	female	good	unknown/incomplete	0.01	11.06 (2.48)
Regimen-C	female	good	complete	0.05	13.72 (5.17)
Regimen-C	male	poor	unknown/incomplete	0.01	12.48 (1.30)
Regimen-C	male	poor	complete	0.17	16.70 (6.93)
Regimen-C	male	good	unknown/incomplete	0.02	14.30 (2.77)
Regimen-C	male	good	complete	0.09	16.07 (5.19)
Regimen-DI	female	poor	unknown/incomplete	0.01	14.60 (2.33)
Regimen-DI	female	poor	complete	0.08	15.35 (6.24)
Regimen-DI	female	good	unknown/incomplete	0.01	13.85 (6.12)
Regimen-DI	female	good	complete	0.09	14.34 (5.49)
Regimen-DI	male	poor	unknown/incomplete	0.03	15.87 (4.04)
Regimen-DI	male	poor	complete	0.14	18.54 (6.02)
Regimen-DI	male	good	unknown/incomplete	0.01	10.63 (2.98)
Regimen-DI	male	good	complete	0.14	17.11 (5.64)

Table 7.1: Proportions for the 16 unique combinations in the original data with 422 patients. Mean and standard deviation of *age at the date of surgery* are provided per combination.

Initially, censoring times were generated from a Weibull with shape = 2.03 and scale = 5.72 (parameters identified from censoring distribution of the original dataset). This led to simulated datasets with 61% censoring on average. Aiming to investigate the robustness of PLANNs' predictive ability, two (adverse) scenarios were defined with less patients or information on the training data: i) removing patients censored before the second year, or ii) curtailing patients survival at 5 years (administrative censoring at 5 years). Hereto, a Weibull distribution was used with shape 0.75 (set a priori) and appropriate scale parameter to reach on average 20, 40, 61 or 80% censoring on the simulated datasets (scale parameter 76, 20.5, 6.8, 2.4, respectively) and at the same time obtain a sufficient number of patients for these extra scenarios (for details see Section 5 of Additional file 3).

For 61% censoring, scenario 1 (Weibull with shape = 2.03, scale = 5.72) and 2 (Weibull with shape = 0.75, scale = 6.8) are presented with details in Section "Results" and in Additional file 3 (supplementary results). Predictive performance of the methods was not affected by the modification of Weibull parameters for the same censoring percentage. Therefore, it was reasonable to assume (a priori) a shape of 0.75 for the other simulated scenarios.

7.2.3 Cox proportional hazards model

The Cox proportional hazards (PH) regression model is commonly employed to estimate the effect of risk factors in models for time-to-event outcomes, on survival outcomes because of its simplicity [1]. This model assumes that each covariate has a multiplicative constant over time effect in the hazards function.

Suppose that data with sample size n consist of the independent observations from the triple (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 indicator ($d_i = 1$ if the event occurred, $d_i = 0$ if an observation is right censored) and $x_i = (x_1, \cdots, x_p)$ is the vector of predictors. The hazard function of the Cox model with time-fixed covariates is specified as:

$$h(t|X) = h_0(t) \exp(X^T \boldsymbol{\beta}), \tag{7.2}$$

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 the parameter vector.

The corresponding partial likelihood can be written:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp\left(\sum_{k=1}^{p} \beta_k X_{ik}\right)}{\sum_{j \in R(t_i)} \exp\left(\sum_{k=1}^{p} \beta_k X_{jk}\right)}, \tag{7.3}$$

where D is the set of failures, and $R(t_i)$ is the risk set at time t_i of all individuals who are still in the study at the time just before time t_i. This function is maximised over $\boldsymbol{\beta}$ to estimate the model parameters.

7.2.4 Survival neural networks

ANNs were inspired from the human brain activity and more specifically from the neurons that transmit information between different areas of the brain. ANNs 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. Feed forward neural networks, which are also called multilayer perceptrons, utilize a supervised learning technique called backpropagation for training [33, 34].

In the medical field, ANNs are popular ML methods and therefore their application has been extended to survival analysis. These are usually called survival neural networks (SNNs). Different approaches have been considered; some model the survival probability $S(t)$ directly or the unconditional probability of death $F(t)$ [7, 8, 10] whereas other approaches estimate the conditional hazard $h(t)$ [9, 11, 12]. For this work, the partial logistic artificial neural network (PLANN) approach was applied as developed originally by Biganzoli *et al.* [9] and its extensions by Kantidakis *et al.* [20] for a simple feed forward ANN with one hidden layer. PLANN 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 PH 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. Here, without loss of generality, each subject was replicated for a maximum of 8 yearly intervals for the main analyses. The last interval included survival times longer than 7 years (as it was not of interest to specify follow-up times longer than 8 years in separate intervals). Similarly, for supplementary analyses, $4 \cdot 8 = 32$ or $2 \cdot 8 = 16$ time intervals were defined representing 3 or 6-month periods (no separate intervals for follow-up longer than 8 years).

Activation function of both hidden and output layers is the logistic (sigmoid) function:

$$f(\theta) = \frac{1}{1 + e^{-\theta}}. \tag{7.4}$$

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 $S(t) = \prod_{l:t_l \leq t}(1 - h_l)$.

The contribution to the log-likelihood is calculated for all intervals one is at risk. Following Biganzoli *et al.* (1998) [9], the dependence of hazards can be jointly modelled from the time variable α_l and the vector of covariates $\mathbf{x}_i$ using as event indicator d_{il} (with $d_{il} = 1$ in the interval A_l containing the event and $d_{il} = 0$, otherwise) for discrete survival data as:

$$E = -\sum_{i=1}^{n} \sum_{l=1}^{l_i} \{d_{il} \log h_l(\mathbf{x}_i, \alpha_l) + (1 - d_{il})[1 - \log h_l(\mathbf{x}_i, \alpha_l)]\}, \tag{7.5}$$

where $h_l(\mathbf{x}_i, \alpha_l)$ are discrete hazard rates which are estimated by the output values $\lambda(\mathbf{x}_i, \alpha_l, w)$ with w the weights matrix. The error (loss) function in Equation (7.5) is summed both over the n subjects and the time intervals $l = 1, \cdots l_i$ in which the subject i is observed. It is equivalent to the cross-entropy error function used for binary classification problems. By using this error function in an ANN with no hidden layers and the logistic activation function (Equation (7.4)), a linear logistic regression model is obtained.

The PLANN original model can be mathematically framed as:

$$\lambda(\mathbf{x}_i, \alpha_l, w) = f\left[w'_{0k} + \sum_{h=1}^{H} w'_{hk} g_h\left(w_{0h} + w_{1h}\alpha_l + \sum_{j=1}^{j=p} w_{(j+1)h} x_{ij}\right)\right], \tag{7.6}$$

where $j = 1, 2, \cdots, J$ are the nodes in the input layer, $h = 1, 2, \cdots, H$ are the nodes in the hidden layer, $k = 1$ is the unique node in the output layer, x_{ij} are the p elements of covariate vector $\mathbf{x}_i$. In addition, w_{jh} are the weights from the input to the hidden layer, w'_{hk} are the weights from the hidden to the output layer (w_{0h}, w'_{0k} are the weights of the bias nodes for the input-hidden and the hidden-output layer, respectively) and $g_h(\cdot)$, $f(\cdot)$ are the activation (transformation) functions for the hidden and the output layers ($f(\cdot)$ is given in Equation (7.4)). In Section "Evaluation of predictive performance" a prognostic score is defined from Equation (7.6), which is used to construct C-index for PLANNs.

Extensions of the specification of the PLANN approach were applied as described by Kantidakis *et al.* (2020) [20]. SNNs were tuned investigating two more activation functions for the hidden layer the rectified linear unit (ReLU):

$$g_h(\eta) = \eta^{+} = \max(0, \eta), \tag{7.7}$$

which is the most frequently used activation function for ANNs and the hyperbolic tangent (tanh):

$$g_h(\eta) = \frac{1 - e^{-2\eta}}{1 + e^{-2\eta}}. \tag{7.8}$$

These activation functions can be seen as different modulators of the non-linearity transferred to the hidden layer from the input features. Note that the activation function in the output layer was strictly the logistic (sigmoid) shown in Equation (7.4). The L non-overlapping intervals of the discrete survival times were treated as L separate variables ($1 + L + p$ nodes in the input layer instead of $1 + 1 + p$ for PLANN original). However, the extension of PLANN with 2 hidden layers was not applied due to substantial danger for overfitting in this clinical setting with small data (small to moderate sample size, 5 predictors only).

7.2.5 Model training

Each simulated dataset was randomly split into 2 complementary parts (50% training, 50% test data) under the same event/censoring proportions. To tune the hyperparameters of SNN (PLANN original or extended), 5-fold cross validation was performed with grid search on the training part of a simulated dataset with 1000 synthetic patients according to the censoring rate of interest. Training data was divided into 5 folds. Each time 4 folds were used to train a model and the remaining fold was used to validate its performance (the same folds were used for PLANN original and extended). This procedure was repeated 5 times to take into account all combinations of folds. Performance of the final models with the hyperparameters selected was assessed on the test sets (for each simulated dataset). Packages of implementation for PLANN original [9] and PLANN extended [20] and technical details such as the choice of tuning parameters are provided in Additional file 2. Parameters were tuned on the training data based either on the C-index [35] or the integrated Brier score (IBS) at 5 years (time-point of major clinical interest) [36]. These measures are described in the next section. All analyses were performed in R programming language version 4.0.1 [22].

7.2.6 Evaluation of predictive performance

The predictive performance of the models was assessed in terms of discrimination and calibration. The C-index, the Brier score, the integrated Brier score (IBS), and the miscalibration (in terms of absolute accuracy error) were estimated in the simulated test datasets.

In survival analysis, a well known measure of model performance is Harrell's C-index [35] as an extension of the concept of the receiver operating characteristic (ROC) area [37]. It measures the proportion of all usable pairs of observations (at least one of them has the event of interest) for which the survival times and model predictions are concordant taking into account censoring. Typically, it takes values between 0.5 to 1 with higher values indicating higher ability of the model to discriminate well. Nevertheless, good discrimination does not imply good calibration and vice versa.

For a Cox model, the predicted survival time of an individual is longer if the linear prognostic index (PI) defined as $X^T\boldsymbol{\beta}$ is lower (opposite ranking). This relationship can then be used to calculate the Harrell's C-index to quantify the ability of the model to discriminate among subjects with different event times [35, 37]. For the PLANN, the equivalent is a non-linear time-dependent PI defined as $\theta = w'_{0k} + \sum_{h=1}^{H} w'_{hk} g_h \Big(w_{0h} + w_{1h}\alpha_l + \sum_{j=1}^{j=p} w_{(j+1)h} x_{ij} \Big)$ in Equation (7.6) inside the logistic (sigmoid) activation function of the output layer $f(\theta)$ (see Equation (7.4)). Therefore, Equation (7.6) can be re-written as:

$$\lambda(\mathbf{x}_i, \alpha_l, w) = \frac{1}{1+e^{-\theta}}. \tag{7.9}$$

By solving this equation with respect to θ this non-linear PI can be estimated as:

$$\theta(\mathbf{x}_i, \alpha_l, w) = \log \left[\frac{\lambda(\mathbf{x}_i, \alpha_l, w)}{1-\lambda(\mathbf{x}_i, \alpha_l, w)} \right], \tag{7.10}$$

which is the log-odds ratio of the conditional hazard probabilities.

The non-linear time-dependent PI in Equation (7.10) depends on the covariates, the time interval and the weights of the network. A simple non-linear PI that is not time-dependent can be obtained by averaging these indexes over all the time intervals

$$\theta(\mathbf{x}_i, w) = \frac{\sum_{l=1}^{L}(\theta(\mathbf{x}_i, \alpha_l, w))}{L}. \tag{7.11}$$

Then this non-linear PI was used to calculate the C-index for PLANN original. Similarly, a simple non-linear PI was obtained for PLANN extended by averaging the time-dependent non-linear PIs over all intervals.

The C-index provides a rank statistic that is not time-dependent. Following van Houwelingen and le Cessie [38] a time-dependent prediction error [36] is defined as

$$\mathit{Brier}(y, \hat{S}(t_0|x)) = (y - \hat{S}(t_0|x))^2, \tag{7.12}$$

where $\hat{S}(t_0|x)$ is the model-based survival probability 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 in simulated data, censored observations before time t_0 have to be considered. To calculate Brier Score when censored observations are present, Graf *et al.* proposed the use of inverse probability of censoring weighting [36]. Hence, an estimate of the average prediction error of the prediction model $\hat{S}(t|x)$ at time $t = t_0$ is

$$Err_{Score}(\hat{S}, t_0) = \frac{1}{n}\sum_i 1\{d_i = 1 \vee t_i > t_0\}\frac{Score(1\{t_i > t_0\}, \hat{S}(t_0|x_i))}{\hat{C}(\text{min}(t_i-, t_0)|x_i)}. \tag{7.13}$$

In (7.13), the term $\frac{1}{\hat{C}(\text{min}(t_i-,t_0)|x_i)}$ is a weighting scheme known as inverse probability of censoring weighting (IPCW) and $Score$ is the Brier Score. It ranges (typically) from 0 to 0.25 with lower values indicating smaller prediction error. Brier score was calculated at 0-5 years (time period of clinical interest).

An overall measure of prediction error is the Integrated Brier Score (IBS) which can summarise the prediction error over the whole range up to a time horizon of interest $\int_0^{t_{hor}} Err_{Score}(\hat{S}, t_0)dt_0$ (here t_{hor} = 5 years) [2]. IBS provides the cumulative prediction error up to t_{hor} at all available times (e.g. $t_0 = 1, 2, 3, 4, 5$ years). As the Brier score, it ranges (typically) from 0 to 0.25.

Last, the predictive ability of the models was evaluated based on their calibration on the test data, which is usually neglected for ML techniques. Calibration refers to the agreement between observed outcomes and predictions [39, 40]. For each method (Cox model, PLANN original, PLANN extended) the predicted survival probabilities are estimated, and the synthetic 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 the Kaplan-Meier (KM) methodology [27]. Miscalibration on test sets for each group is defined as the mean squared error (MSE) of the difference between the observed and the predicted survival probabilities:

$$MSE(t_0) = \frac{\sum_{m=1}^{4}\left[S_{KM}^{(m)}(t_0) - \hat{S}^{(m)}(t_0)\right]^2}{4}, \tag{7.14}$$

at $t_0 = 2$ and $t_0 = 5$ years.

7.3 Results

In this section the findings are presented. The following models were compared: a) Cox model, b) PLANN original and c) PLANN extended in terms of predictive performance in the simulated data with 5 prognostic factors under different percentages of censoring / sample size per dataset. Additional file 3 provides supplementary results for the scenarios and extra details not shown here (e.g. hyperparameters selected for the ML techniques, more tables and plots for predictive performance).

7.3.1 Proportional hazards assumption

The PH assumption was tested in the original clinical data ($n = 422$) detailed in "Clinical data and imputation technique". Plots are provided in Additional file 1. The global test for the Schoenfeld residuals was not significant (p-value = 0.244) and the Schoenfeld residuals showed random patterns against time with coefficients close to 0 suggesting that the proportionality of hazards is not violated [41]. Individual Schoenfeld test for the 5 variables was only significant for *age since the date of surgery* (p-value = 0.035). Nevertheless, investigation of the plotted Schoenfeld residuals values did not show any systematic divergence for *age* from the straight line with residual value 0 (no time-dependent effect, Figure S3 of Additional file 1). Moreover, the linear assumption was examined plotting *age* against the martingale residuals of the null Cox model. The log and square root transformations were tested but did not improve its functional form (see Additional file 1). Non-linearity for *Age* seemed to be small. There was no statistical evidence for interactions between risk factors (all p-values > 0.10 in the multivariate Cox model). For the rest of the analyses, Cox models without interactions between the 5 predictors or time-dependent effects were considered.

7.3.2 SNNs tuned with IBS or C-index

The hyperparameters selected for PLANN original and PLANN extended are provided in Section 2 of Additional file 3. Optimal combinations are reported separately for IBS at 5 years or C-index. For PLANN original (2 hyperparametes (node) `size` and `decay`), a small `size` was selected for the majority of scenarios by both performance measures. Nevertheless, a larger `decay` parameter was suggested in general by the networks tuned for IBS. For PLANN extended, tuning was performed on a 5-D space for parameters `nodesize`, `dropout rate`, `learning rate`, `momentum` and `weak class weight` (see details in Additional file 2). Three activation functions were tested for the input-hidden layer: the ”sigmoid” (logistic), the ”relu” (rectified linear unit) and the ”tanh” (hyperbolic tangent). Overall, ”tanh” and ”relu” provided the best performance on the training data for each scenario (IBS or C-index). Optimal parameters for `nodesize`, `dropout rate`, `learning rate` or `momentum` differed between the scenarios. A `weak class weak` of 1 or 1.05 (small adjustment in favor of the weak class) was generally selected.

Measure	PLANN original IBS	PLANN original C-index	PLANN extended IBS	PLANN extended C-index
Brier score 2 years (sd)	0.145 (0.012)	0.146 (0.012)	0.144 (0.011)	0.144 (0.011)
Brier score 5 years (sd)	0.229 (0.010)	0.232 (0.011)	0.229 (0.011)	0.230 (0.010)
IBS 5 years (sd)	0.124 (0.007)	0.125 (0.007)	0.123 (0.006)	0.124 (0.007)
C-index (sd)	0.633 (0.022)	0.628 (0.023)	0.637 (0.021)	0.631 (0.024)
Miscalibration 2 years (sd)	0.003 (0.003)	0.004 (0.003)	0.003 (0.002)	0.003 (0.002)
Miscalibration 5 years (sd)	0.006 (0.004)	0.007 (0.004)	0.008 (0.006)	0.006 (0.006)

Table 7.2: Performance of PLANN original and PLANN extended tuned for IBS at 5 years or C-index for 61% censoring (scenario 1) and 1000 synthetic patients per dataset. The standard deviation (sd) based on 1000 datasets is provided in parentheses.

The performance of tuned SNNs was compared with either IBS at 5 years or C-index. Results for scenario 1 with 61% censoring are presented in Table 7.2. It can be observed that both SNNs had a slightly better predictive performance tuned for IBS at 5 years. This pattern was consistent for the other scenarios (Tables S10-S13 in Additional file 3). This might be related with the nature of these neural networks, which both predict conditional hazard (death) probabilities h_l for each time interval in a single output node. Then, survival probabilities can be directly estimated at each interval as $S(t) = \prod_{l:t_l \leq t}(1 - h_l)$. From the 2 predictive performance measures considered here (to train the networks), IBS was calculated through the model-based survival probabilities of an individual beyond t_0 (Equation 7.13) whereas the C-index was estimated indirectly after the calculation of a non-linear PI for each individual (Equation 7.11). Taking everything into account, IBS at 5 years seemed to be more reliable than C-index to tune PLANNs. Hence, in the analyses shown below, optimal combinations for IBS at 5 years were selected for SNNs (PLANN original and extended).

7.3.3 Comparison of predictive performance for the methods

The simulation procedure was repeated $B = 1000$ times to generate $N_1 = 250$ or $N_2 = 1000$ synthetic patients per dataset (50% training and 50% test set). In this section, the 3 methods (Cox model, PLANN original, PLANN extended) are compared based on different predictive performance measures on test data: i) Brier score from 0-5 years, ii) Harrell's C-index, iii) Integrated Brier Score (IBS) at 5 years and iv) miscalibration at either 2 or 5 years. Measures are detailed in Section "Evaluation of predictive performance". For the sake of simplicity, the focus is on 61% censoring on average (scenario 1). Plots for the other scenarios: 61% scenario 2, 20%, 40% and 80% are included in Section 4 of Additional file 3.

Figure 7.1 shows the Brier score corresponding to each method per year (0-5 years). For small sample size ($N_1 = 250$), the performance largely overlapped (the standard deviations [sds] were very similar). For larger sample size ($N_2 = 1000$), the Cox model performed slightly better than the SNNs (sd over 1000 datasets was also smaller for the Cox model). For all methods the predictive performance improved when the sample size increased (smaller Brier scores, higher C-indexes). For 61% censoring scenario 2, results were very similar - especially for large sample size. For smaller sample size, PLANN original performed slightly worse than PLANN extended or Cox and had the largest sd. For 80% censoring results were to the same direction as these. Interestingly, for 20 and 40% censoring PLANN original and PLANN extended performed as good as the Cox model for both sample sizes examined.

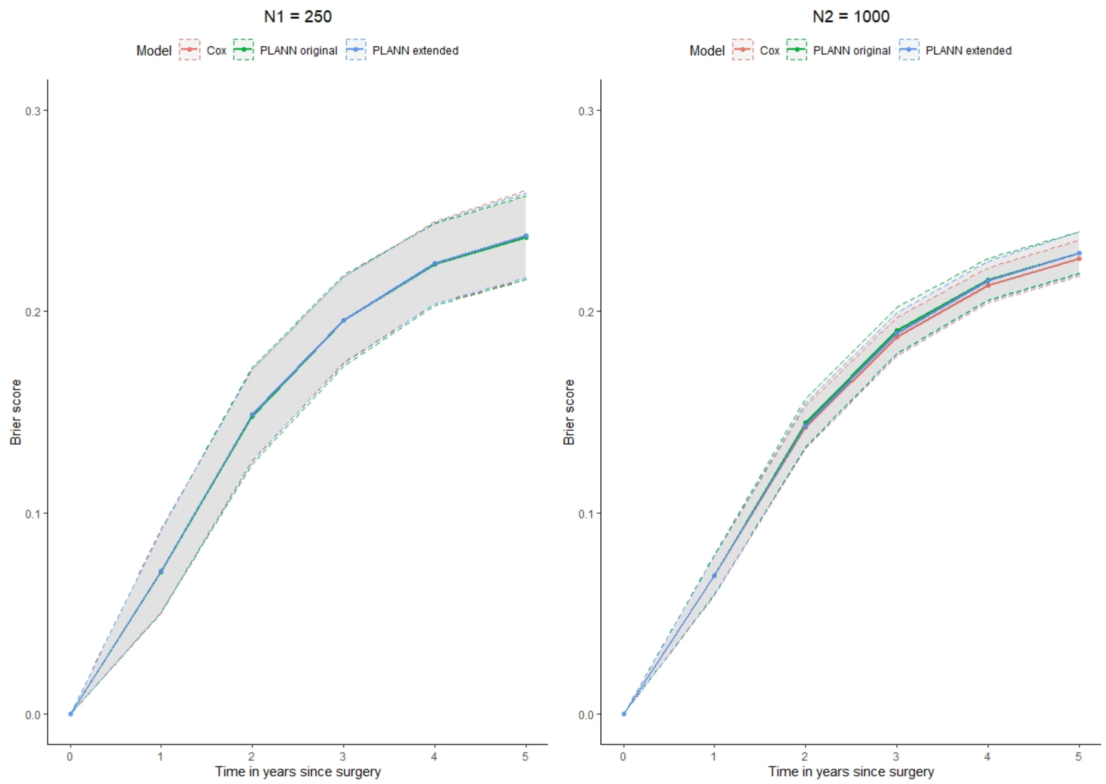


Figure 7.1: Brier score for Cox, PLANN original, and PLANN extended ± one standard deviation for 61% censoring scenario 1. Left panel: 250 patients, right panel: 1000 patients.

The C-index and IBS at 5 years are illustrated in Figure 7.2 for 61% censoring scenario 1. Regarding the C-index, performance was very similar for N_1 whereas the Cox model achieved (marginally) the best performance for N_2 very close to PLANN extended. For IBS at 5 years, performance was very similar between the methods. The Cox model provided the smallest error for larger sample size (largest sd by PLANN original). For all methods, performance improved as the sample size increased. Examining the other scenarios, the performance of the methods was very close in terms of C-index or IBS. Cox models achieved the best performance (and the smallest sds for N_2). For some scenarios (and different sample sizes) PLANN original performed better than PLANN extended and vice versa. This is likely to be related with the optimal parameters selected in each case. The PLANNs fitted might have been out of control for smaller sample size due to insufficient amount of regularization implied by the parameters, but the results improved for larger sample size.

Furthermore, the miscalibration of the methods was compared with boxplots in Figure 7.3 for 61% censoring (scenario 1). For $N_1 = 250$ PLANN original achieved a slightly better performance at 2 and 5 years. Nevertheless, for $N_2 = 1000$ patients the Cox models had by far the lowest miscalibration error (defined as the MSE for 4 groups

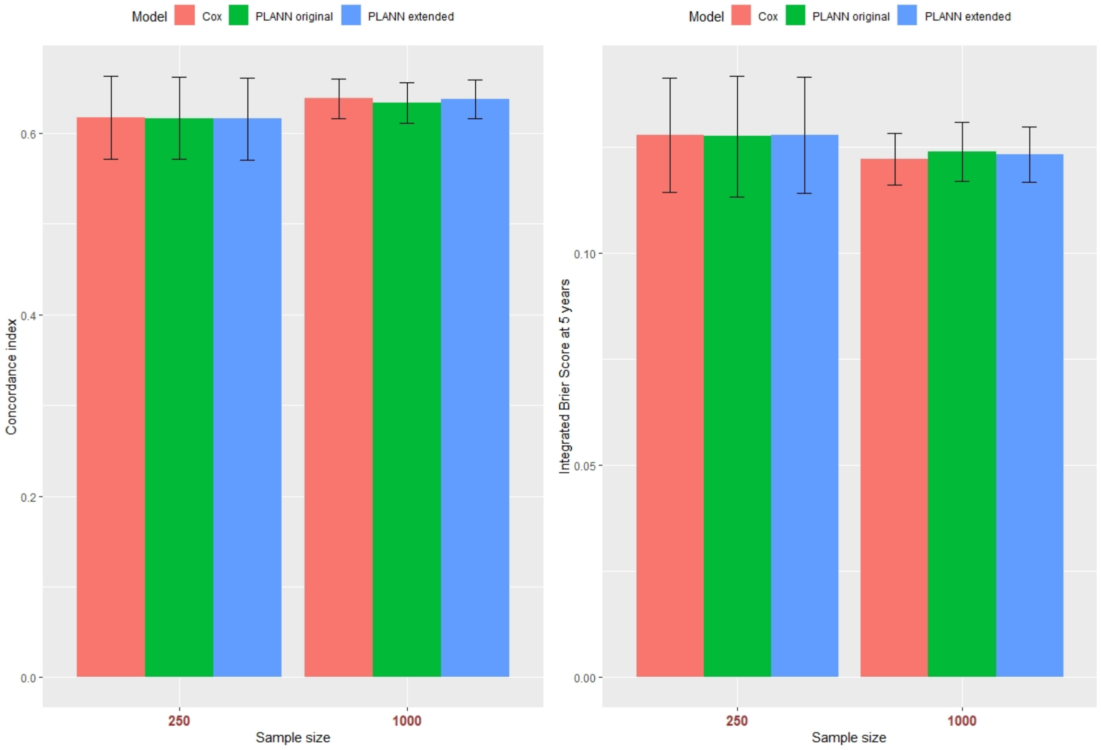


Figure 7.2: Predictive performance for Cox, PLANN original, and PLANN extended for sample size 250 and 1000 ± one standard deviation for 61% censoring scenario 1. Left panel: C-index, right panel: IBS at 5 years.

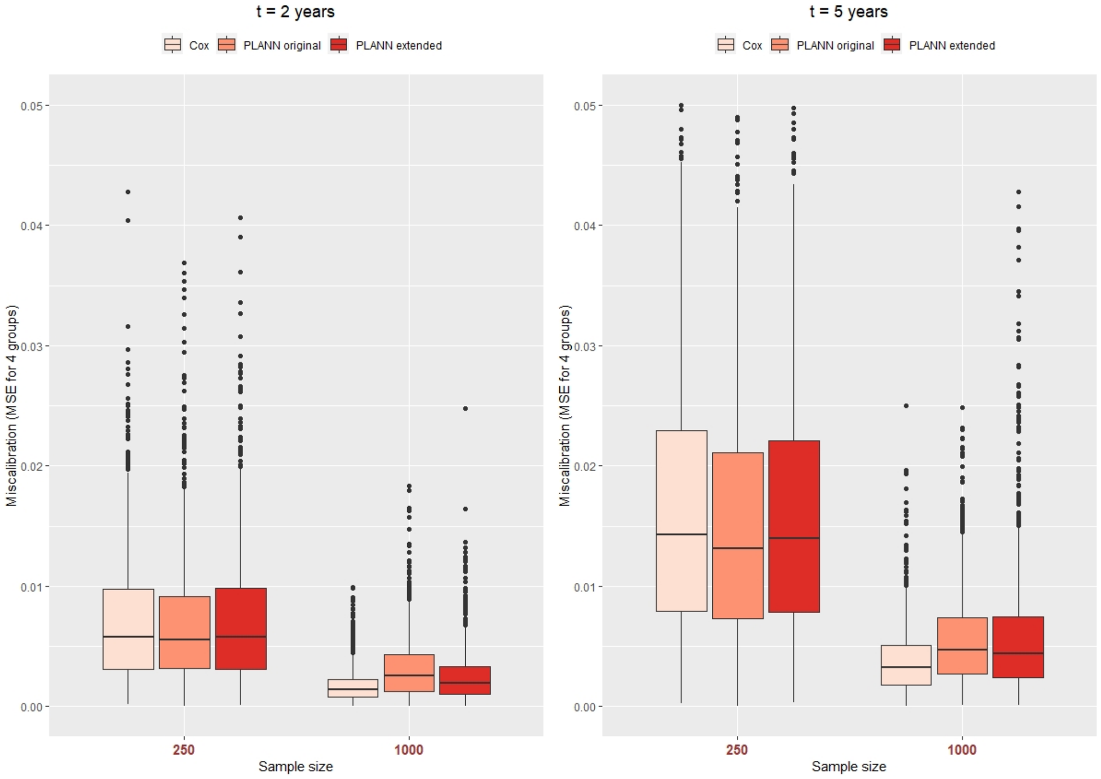


Figure 7.3: Miscalibration for Cox, PLANN original, and PLANN extended per sample size and 61% censoring (scenario 1). Left panel: 2 years, right panel: 5 years.

on test data). PLANN extended showed the highest number of outliers here. Miscalibration error decreased for datasets with larger sample size. For the rest of the scenarios a similar pattern was observed. For N_1 miscalibration error was almost the same between the methods (at 2 or 5 years) but for larger datasets in size (N_2 = 1000) Cox model was clearly better calibrated than PLANNs (for 20% censoring differences were minimal). In all boxplots, PLANN original or extended had more outliers than Cox models for N_2. This indicates that both were less stable than Cox. Moreover, both SNNs were less calibrated for larger percentages of censoring (less events).

In Section 7 of Additional file 3, the effect of interval length (3-monthly or 6-monthly intervals) is reported for 61% censoring (scenario 1). Performance of monthly intervals versus yearly intervals was very similar for PLANN original. This is consistent with the absence of relevant time dependent effects in the simulated datasets, since an increase of the binning over time intervals should improve predictive performance if such effects were present [42]. For PLANN extended, performance slightly deteriorated for monthly intervals. This can be explained by the increase in the number of input features. For PLANN extended, the L non-overlapping intervals were treated as L separate variables. Therefore, as 3-monthly and 6-monthly intervals corresponded to 32 and 16 variables (versus 8 for yearly intervals), complexity of the network increased and its predictive ability decreased. A different parametrization of the time intervals into 1 prognostic factor (input feature) as in PLANN original instead of dummy coding for each interval would effectively deal with this issue, if monthly intervals are to be considered.

7.3.4 Impact of adverse scenarios for predictive ability

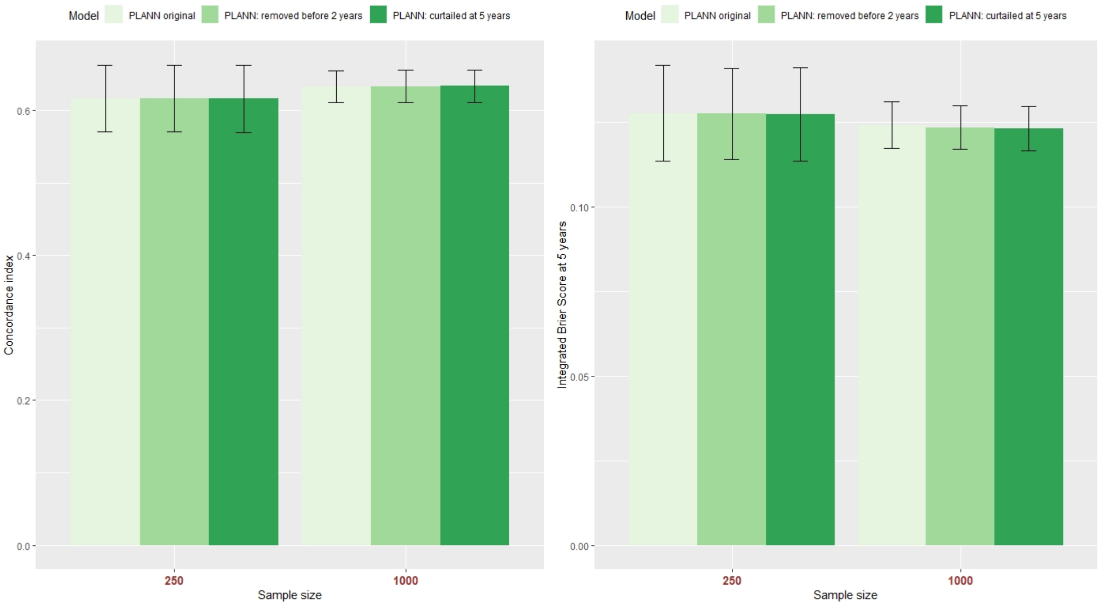


Figure 7.4: Predictive performance of PLANN original ± one standard deviation for sample size 250 or 1000 and 61% censoring (scenario 1). Darker green palette colours correspond to the 2 adverse scenarios a) removing patients censored before the second year, or b) curtailing patients' survival at 5 years. Left panel: C-index, right panel: IBS at 5 years.

To investigate the robustness of the methods, the following 2 scenarios were defined on the training part of the simulated data: a) removing patients censored before the second year, or b) curtailing patients' survival at 5 years. The number of patients affected by these scenarios for different % of censoring is provided in Additional file 3 Section 5 (Tables S14-S18).

Results for PLANN original and PLANN extended for 61% censoring scenario 1 are illustrated in Figures 7.4 and 7.5, respectively. The predictive performance for both SNNs did not seem to be affected in terms of C-index or IBS at 5 years. More plots for SNNs (green and blue palette colours) and the Cox model (red palette colours) can be found in Additional file 3 Section 5 for different censoring scenarios. Overall, all methods were quite robust to the adverse scenarios investigated. PLANN extended was less robust that PLANN original for 20% censoring scenario b (administrative censoring) and 80% censoring scenario a (removing patients) for N_1 = 250 (Figures S23 and S29 in Additional file 3).

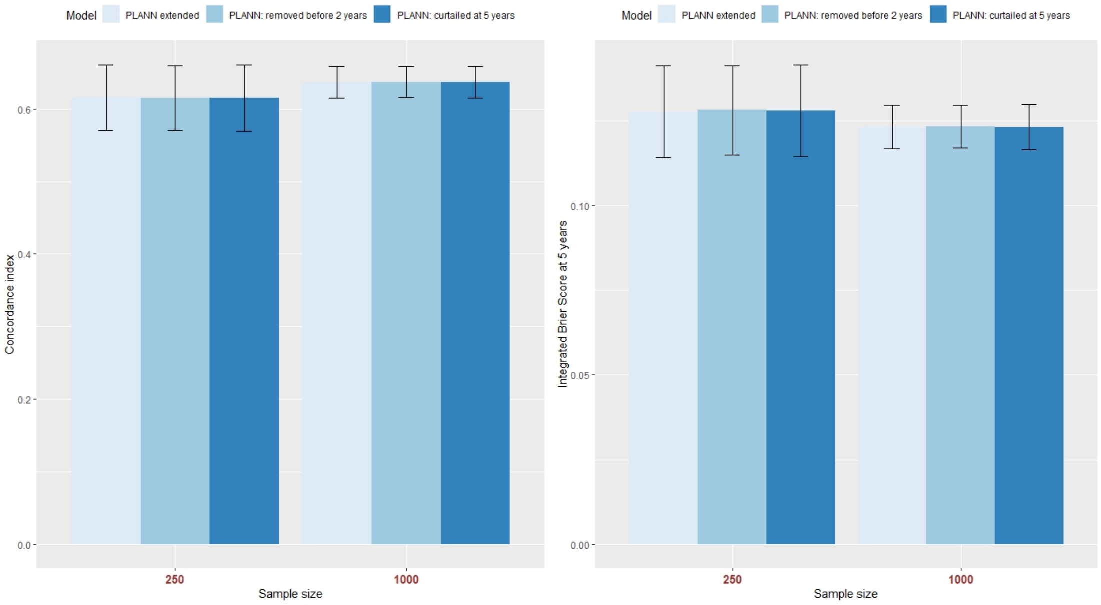


Figure 7.5: Predictive performance of PLANN extended ± one standard deviation for sample size 250 or 1000 and 61% censoring (scenario 1). Darker blue palette colours correspond to the 2 adverse scenarios a) removing patients censored before the second year, or b) curtailing patients' survival at 5 years. Left panel: C-index, right panel: IBS at 5 years.

7.4 Discussion

Nowadays, there is an increased interest in applying ML techniques to create prediction models, because of their intrinsic capability in extracting and modelling the relevant information underlying the available data. This trend is pertinent with the collection of large volume of patient data in electronic health records (EHR). However, concerns have been raised that the employment of Artificial Intelligence (AI) for clinical prediction is overhyped in some contexts. Some points of criticism include the use of unsuitable performance measures, overfitting the training data, and the lack of extensive assessment of predictive accuracy (for instance absence of calibration curves). Hence, appropriate development/evaluation and transparent reporting of such prediction models is of paramount importance to avoid research waste [43, 44].

Two simulation studies compared PLANN original with Cox models for prediction investigating linear and non-linear effects for the hazards and several censoring rates [45, 46]. In the first study, Biglarian *et al.* (2013) proposed PLANN for high censoring or when complex interactions are present [45]. In simple models, differences in predictive ability were negligible. Gong *et al.* (2018) found that PLANN is less sensitive to data size and censoring rates than Cox regression and achieved the best performance when predictor variables assumed non-linear relationships (or a similar performance elsewhere). ANN extensions of the Cox PH model have been considered as alternatives to PLANN models in the literature for prediction. In 1995, Faraggi and Simon replaced the linear function of the Cox model with the non-linear output of a feed forward ANN with `logistic` hidden and `linear` output layers [47]. Modern deep networks utilise the framework by Faraggi-Simon to extend the Cox model for low- or high-dimensional data [48–50].

In this simulation study, PLANN original and its extensions were compared with traditional regression models in a simple setting with small to moderate sample size and 5 predictors with synthetic data generated from a clinical trial (MRC BO06/EORTC 80931) in the absence of complex functional dependence relationships involving time and covariates (i.e. non-linear and non-additive). Different percentages of censoring and sample sizes were investigated based on well-established performance measures for survival analysis. Both aspects of model discrimination and calibration were evaluated. It was shown that SNNs may reach a comparable performance in terms of C-index, Brier score or IBS. The standard deviations (over 1000 repetitions) overlapped to a large extent for all scenarios. Predictive ability was adequately robust to predefined adverse scenarios. However, the Cox models were usually better calibrated (predicted survival probabilities closer to the observed) even though data were not generated from a Cox model. This result in particular shows the relevance of reporting calibration of ML techniques to obtain a neutral comparison with SMs (not reported in the aforementioned articles by Biglarian and Gong). In the paper by Taktak *et al.* (2007) [16], an extensive comparison of different ML models was performed on a large clinical dataset resorting both to discrimination and calibration measures. The Bayesian extension of the PLANN model (PLANN-ARD) achieved a slightly better performance with respect to the other models. Overall, these results and conclusions are consistent with the present findings, which indicate an urgent need of more attention to model calibration.

Both SNNs were tuned based on global performance measures (IBS at 5 years, C-index) on training data according to the amount of censoring. These measures were chosen as they can summarize the predictive ability of a model in one value, in contrast with Brier score that is time-dependent. For the calculation of the C-index for the PLANNs, it was assumed that there is a monotonic relationship between the predicted survival times and the non-linear PI obtained (opposite ranking). Such a relationship holds for the Cox model under the PH assumption [37] (between predicted survival times and the linear PI), but is not guaranteed for ML techniques if there are time-dependent effects between the covariates [51]. Nevertheless, the examination of the PH assumption in the original data (from which the data was generated) and the implementation of PLANNs with time coded in 3-monthly or 6-monthly intervals instead of yearly intervals did not improve the performance of the networks (for 61% censoring first scenario) which supports the evidence that no relevant time-dependent effects were present. To explain this, PLANN can estimate complex functional relationships between time and covariates (if present) to improve predictive ability due to the necessary data transformation into a long format with the time split into a set of non-overlapping intervals. Nonetheless, in the absence of such complex relationships, assuming a monotonic relationship between the predicted survival times and the non-linear PI is reasonable.

ML techniques such as the SNNs considered in this work have both advantages and disadvantages in the application of the considered clinical data. Some of their most appealing characteristics are the minimal assumptions, and the fact that they can model automatically complex (usually high dimensional) data which exhibit non-linearities and higher order interactions between predictors. Meanwhile, model optimisation is a delicate task requiring robust numerical methods and skillful use. Actually, it should not be neglected that ANNs might converge in suboptimal minima in the error function or not converge on a true and stable local minimum [52], require non-trivial implementation time, and have limited interpretability. More specifically from the two PLANNs examined, PLANN extended required more time and effort for model fine-tuning because of the larger number of hyperparameters (5 versus 2 for PLANN original) and the inclusion of time intervals as multiple input features. Therefore, PLANN extended was a more complicated and harder to control network. On the other hand, the standard Cox model makes the PH assumption and implies additivity of effects between the predictors (as any regression model), but offers fast implementation and straightforward interpretation of the estimated coefficients via hazard ratios - which is helpful for clinicians to take informed decisions. However, the shape of the hazard function over time can also be extracted from PLANN models allowing for visualisation of their results. An example of this application for breast cancer clinical data is in ref. [14]. Regarding the practical utility in a simple clinical setting, the Cox model has a advantage over ML techniques such as PLANN original or extended. These methods require significantly more resources and time (such as data pre-processing, tuning of parameters, computational intensity) for merely a comparable predictive performance but also (usually) a suboptimal calibration and a less straightforward interpretation.

The increasing demand for modern methods to improve predictions with survival data has led into the development of several ML algorithms for time-to-event data [5]. Application of such ML techniques should not be pointless but ought to be motivated by exploration of the collected medical data. Building an advanced prediction model powered by AI tools does not necessarily entail a better predictive performance, especially when the sample size and/or the number of features are limited with respect to the complexity of the modeled effects, or when the data are not informative enough [6]. A conventional regression model might still provide more accurate survival probabilities and better generalizability on new data in comparison with ML models not developed appropriately to control model complexity. Therefore, in simple clinical settings, ML methods should only be recommended as exploratory tools to assess linear and additive model assumptions.

7.5 Conclusions

Ultimately, the choice of methodology should be based on a combination of factors such as the types of data collected, their size, computational intensity together with the skills in model implementation, as well as software availability. For this paper, simulated data closely resembled real-life data in a specific clinical setting (low to moderate sample size, small number of predictors) for which the Cox model was expected to be the frontrunner. ML techniques were comparable for a number of suitable predictive performance measures (C-index, Brier score, IBS), but fall short in terms of calibration. Hereto, there is an urgent need to pay closer attention to calibration (absolute predictive accuracy) of ML techniques to achieve a complete comparison with SMs in medical research. Researchers should also be aware of burdensome aspects of ANNs (data pre-processing, tuning of hyperparameters, computational intensity), which are not affordable for most non specialized researchers, that may render them disadvantageous for survival analysis in a simple clinical setting against conventional regression models.

List of abbreviations

AI, Artificial Intelligence; ANN, artificial neural network; EHR, electronic health records; EOI, European Osteosarcoma Intergroup; IBS, integrated Brier score; IPCW, inverse probability of censoring weighting; KM, Kaplan-Meier; ML, machine learning; MSE, mean squared error; OS, overall survival; LUMC, Leiden University Medical Center; PH, proportional hazards; PI, prognostic index; 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; ROC, receiver operating characteristic; sd, standard deviation; SM, statistical model; SNN, survival neural network.

Declarations

Data availability statement

The research data for this project is private. Access to the full dataset of MRC BO06 trial can be requested to MRC Clinical Trials Unit, Institute of Clinical Trial and Methodology, UCL, London, UK. The R-code developed to perform this simulation study is provided in the following GitHub repository `https://github.com/GKantidakis/Simulations-SNNs-vs-Cox`. There, the reader will also find additional files including (1) the R environment (R objects) to generate all data during this study, (2) a zip file which provides randomly generated synthetic data (n = 1000) for 20, 40, 61 (as original, user defined), and 80% censoring, (3) a zip file which is a comprehensive example of how to run the simulations for synthetic data with 61% censoring (as original), and (4) a word document that provides details about the files and a step-by-step tutorial of how to run the R-code.

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

The authors would like to thank the Medical Research Council (MRC) for sharing the dataset used in this manuscript.

Online supplementary materials

The Additional files of this Chapter are available online at https://github.com/GKantidakis/Thesis_supplementary_materials/tree/main/Chapter7.

References

[1] 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.

[2] 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.

[3] F. E. Harrell Jr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer, 2nd edition, 2015. ISBN 978-3-319-19425-7. doi: https://doi.org/10.1007/978-3-319-19425-7. URL http://www.springer.com/series/692.

[4] 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.

[5] 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.

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

[7] P. M. Ravdin and G. M. Clark. A practical application of neural network analysis for predicting outcome of individual breast cancer patients. *Breast Cancer Research and Treatment*, 22(3):285–293, 1992.

[8] K. Liestol, P. K. Andersen, and U. Andersen. Survival analysis and neural nets. *Statistics in Medicine*, 13 (12):1189–1200, 1994. doi: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780131202.

[9] 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.

[10] R. M. Ripley, A. L. Harris, and L. Tarassenko. Non-linear survival analysis using neural networks. *Statistics in Medicine*, 23(5):825–842, 2004. doi: 10.1002/sim.1655.

[11] T. Ching, X. Zhu, and L. X. Garmire. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Computational Biology*, 14(4), 4 2018. ISSN 15537358. doi: 10.1371/journal.pcbi.1006076.

[12] 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.

[13] E. Biganzoli, P. Boracchi, and E. Marubini. A general framework for neural network models on censored survival data. *Neural Networks*, 15(2):209–218, 2002. doi: 10.1016/s0893-6080(01)00131-9. URL www.elsevier.com/locate/neunet.

[14] E. Biganzoli, P. Boracchi, D. Coradini, M. G. Daidone, and E. Marubini. Prognosis in node-negative primary breast cancer: A neural network analysis of risk profiles using routinely assessed factors. *Annals of Oncology*, 14(10):1484–1493, 2003. doi: 10.1093/annonc/mdg422.

[15] A. S. Jones, A. G. F. Taktak, T. R. Helliwell, J. E. Fenton, M. A. Birchall, D. J. Husband, and A. C. Fisher. An artificial neural network improves prediction of observed survival in patients with laryngeal squamous carcinoma. *European Archives of Oto-Rhino-Laryngology*, 263(6):541–547, 6 2006. doi: 10.1007/s00405-006-0021-2.

[16] A. Taktak, L. Antolini, M. Aung, P. Boracchi, I. Campbell, B. Damato, E. Ifeachor, N. Lama, P. Lisboa, C. Setzkorn, V. Stalbovskaya, and E. Biganzoli. Double-blind evaluation and benchmarking of survival models in a multi-centre study. *Computers in Biology and Medicine*, 37(8):1108–1120, 2007. doi: 10.1016/j.compbiomed.2006.10.001.

[17] L. Spelt, J. Nilsson, R. Andersson, and B. Andersson. Artificial neural networks-A method for prediction of survival following liver resection for colorectal cancer metastases. *European Journal of Surgical Oncology*, 39(6):648–654, 2013. doi: 10.1016/j.ejso.2013.02.024.

[18] 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.

[19] 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.

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

[21] 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.

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

[23] G. Ottaviani and N. Jaffe. The epidemiology of osteosarcoma. *Cancer Treatment and Research*, 152:3–13, 2009. ISSN 09273042. doi: 10.1007/978-1-4419-0284-9. URL https://pubmed.ncbi.nlm.nih.gov/20213383/.

[24] A. Misaghi, A. Goldin, M. Awad, and A. A. Kulidjian. Osteosarcoma: a comprehensive review. *SICOT-J*, 4:12, 2018. ISSN 2426-8887. doi: 10.1051/sicotj/2017028. URL https://pubmed.ncbi.nlm.nih.gov/29629690/.

[25] I. J. Lewis, S. Weeden, D. Machin, D. Stark, and A. W. Craft. Received dose and dose-intensity of chemotherapy and outcome in nonmetastatic extremity osteosarcoma. *Journal of Clinical Oncology*, 18(24):4028–4037, 2000. ISSN 0732183X. doi: 10.1200/JCO.2000.18.24.4028.

[26] J. S. Whelan, R. C. Jinks, A. McTiernan, M. R. Sydes, J. M. Hook, L. Trani, B. Uscinska, V. Bramwell, I. J. Lewis, M. A. Nooij, M. Van glabbeke, R. J. Grimer, P. C. W. Hogendoorn, A. H. M. Taminiau, and H. Gelderblom. Survival from high-grade localised extremity osteosarcoma: Combined results and prognostic factors from three European osteosarcoma intergroup randomised controlled trials. *Annals of Oncology*, 23(6):1607–1616, 2012. ISSN 15698041. doi: 10.1093/annonc/mdr491.

[27] 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.

[28] 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.

[29] A. Cheng, D. Kessler, R. Mackinnon, T. P. Chang, V. M. Nadkarni, E. A. Hunt, J. Duval-Arnould, Y. Lin, D. A. Cook, M. Pusic, J. Hui, D. Moher, M. Egger, and M. Auerbach. Reporting guidelines for health care simulation research. *Simulation in Healthcare*, 11(4):238–248, 2016. ISSN 1559713X. doi: 10.1097/SIH.0000000000000150.

[30] A. L. Boulesteix, R. H. H. Groenwold, M. Abrahamowicz, H. Binder, M. Briel, R. Hornung, T. P. Morris, J. Rahnenführer, and W. Sauerbrei. Introduction to statistical simulations in health research. *BMJ Open*, 10 (12), 2020. ISSN 20446055. doi: 10.1136/bmjopen-2020-039921.

[31] P. Royston. The lognormal distribution as a model for survival time in cancer, with an emphasis on prognostic factors. *Statistica Neerlandica*, 55(1):89–104, 2001. ISSN 00390402. doi: 10.1111/1467-9574.00158.

[32] J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics, 2nd edition, 2002. ISBN 978-0-471-36357-6. URL https://www.wiley.com/en-ag/The+Statistical+Analysis+of+Failure+Time+Data,+2nd+Edition-p-9780471363576.

[33] M. Minsky and S. Papert. *Perceptrons; an introduction to computational geometry*. MIT Press, Cambridge, MA, 1 edition, 1969. ISBN 9780262130431.

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

[35] 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.

[36] 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.

[37] M. J. Pencina and R. B. D'Agostino. Overall C as a measure of discrimination in survival analysis: Model specific population value and confidence interval estimation. *Statistics in Medicine*, 23(13):2109–2123, 2004. ISSN 02776715. doi: 10.1002/sim.1802.

[38] 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.

[39] 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/.

[40] C. I. Bharat, K. Murray, E. Cripps, and M. R. Hodkiewicz. Methods for displaying and calibration of Cox proportional hazards models. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 232(1):105–115, 2018. ISSN 17480078. doi: 10.1177/1748006X17742779. URL https://doi.org/10.1177/1748006X17742779.

[41] 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.

[42] Bradley E. Logistic regression, survival analysis, and the Kaplan–Meier curve. *Journal of the American Statistical Association*, 83(402):414–425, 1988. ISSN 1537274X. doi: 10.1080/01621459.1988.10478612.

[43] 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.

[44] 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.

[45] A. Biglarian, E. Bakhshi, A. R. Baghestani, M. R. Gohari, M. Rahgozar, and M. Karimloo. Nonlinear survival regression using artificial neural network. *Journal of Probability and Statistics*, 2013, 2013. doi: https://doi.org/10.1155/2013/753930.

[46] X. Gong, M. Hu, and L. Zhao. Big Data Toolsets to Pharmacometrics: Application of Machine Learning for Time-to-Event Analysis. *Clinical and Translational Science*, 11(3):305–311, 2018. doi: 10.1111/cts.12541.

[47] D. Faraggi and R. Simon. A neural network model for survival data. *Statistics in Medicine*, 14(1):73–82, 1995. doi: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780140108.

[48] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1), 2018. doi: 10.1186/s12874-018-0482-1.

[49] H. Wang and G. Li. Extreme learning machine Cox model for high-dimensional survival analysis. *Statistics in Medicine*, 38(12):2139–2156, 2019. ISSN 10970258. doi: 10.1002/sim.8090. URL https://doi.org/10.1002/sim.8090.

[50] T. Sun, Y. Wei, W. Chen, and Y. Ding. Genome-wide association study-based deep learning for survival prediction. *Statistics in Medicine*, 39(30):4605–4620, 2020. ISSN 10970258. doi: 10.1002/sim.8743.

[51] L. Antolini, P. Boracchi, and E. Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, 2005. doi: 10.1002/sim.2427.

[52] C. M. Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995. ISBN 978-0-19-853864-6.

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

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

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]:

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

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:

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

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 $\alpha_h(\mathbf{z}_k, \boldsymbol{\beta}_h) = \frac{\exp(\beta_{0h} + \beta_h^T \mathbf{z}_k)}{1 + \exp(\beta_{0h} + \beta_h^T \mathbf{z}_k)}$.

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, $\alpha_h(\mathbf{z}_k, \boldsymbol{\beta}_h) = \max(0, \beta_{0h} + \beta_h^T \mathbf{z}_k)$, and (2) the hyperbolic tangent (tanh), $\alpha_h(\mathbf{z}_k, \boldsymbol{\beta}_h) = \frac{1 - \exp(-2(\beta_{0h} + \beta_h^T \mathbf{z}_k))}{1 + \exp(-2(\beta_{0h} + \beta_h^T \mathbf{z}_k))}$. 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.

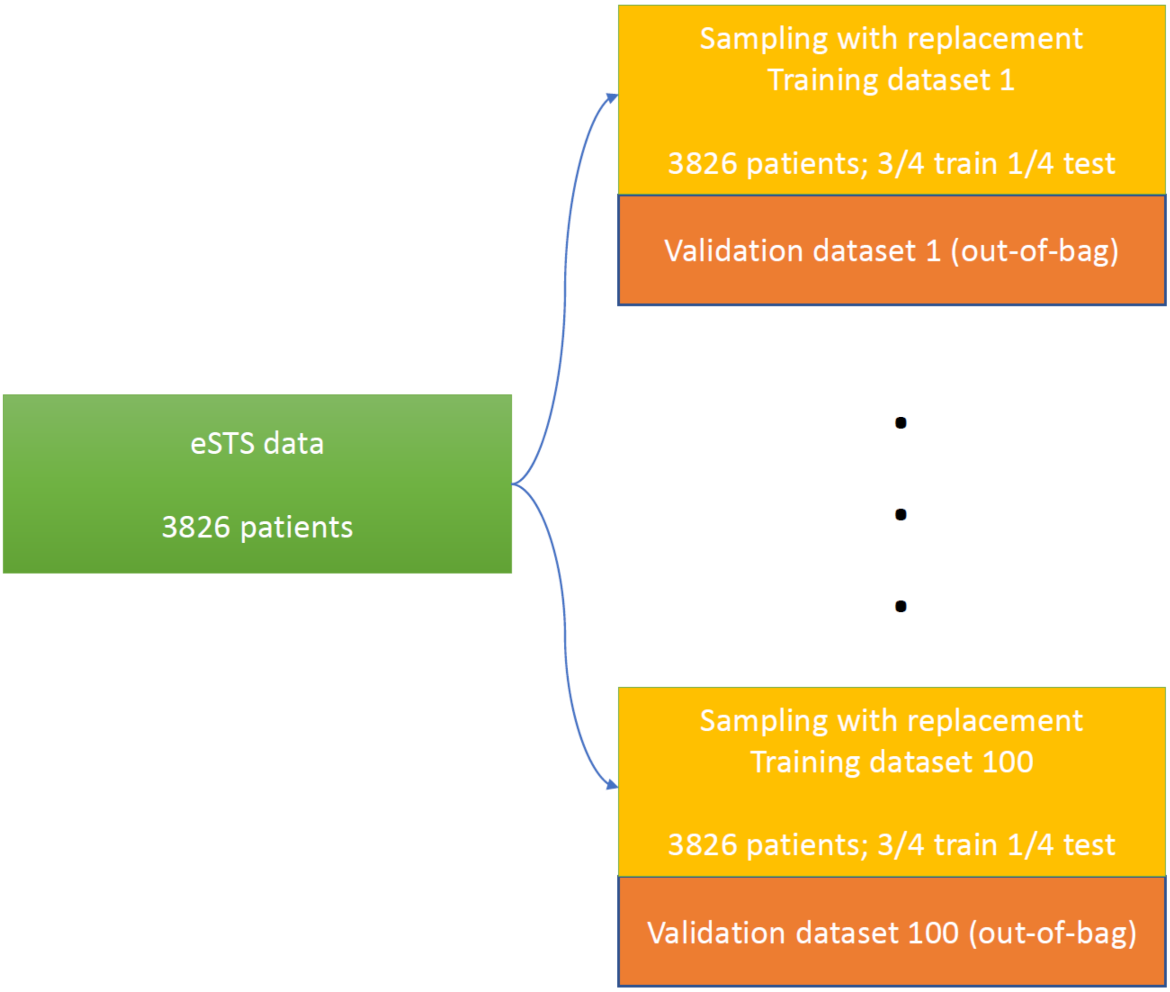

Sampling with replacement
Training dataset 1
3826 patients; 3/4 train 1/4 test
Validation dataset 1 (out-of-bag)
eSTS data
3826 patients
Sampling with replacement
Training dataset 100
3826 patients; 3/4 train 1/4 test
Validation dataset 100 (out-of-bag)

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.

	PLANNCR extended with Brier score	**PLANNCR extended with AUC**
Performance		
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.

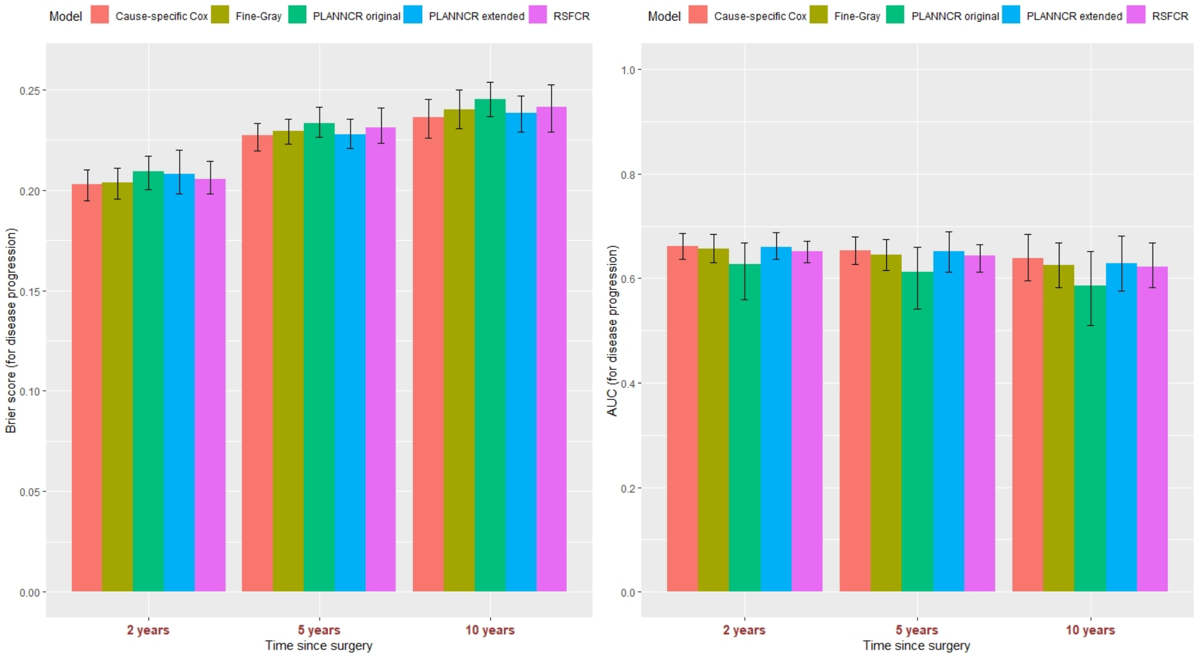


Figure 8.2: Predictive performance of cause-specific Cox model, Fine-Gray model, PLANNCR original, PLANNCR extended (tuned with Brier score at 5 years), and RSFCR for the event of interest: disease progression ± 95% percentile confidence intervals based on 100 validation datasets. Left panel: Brier score,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

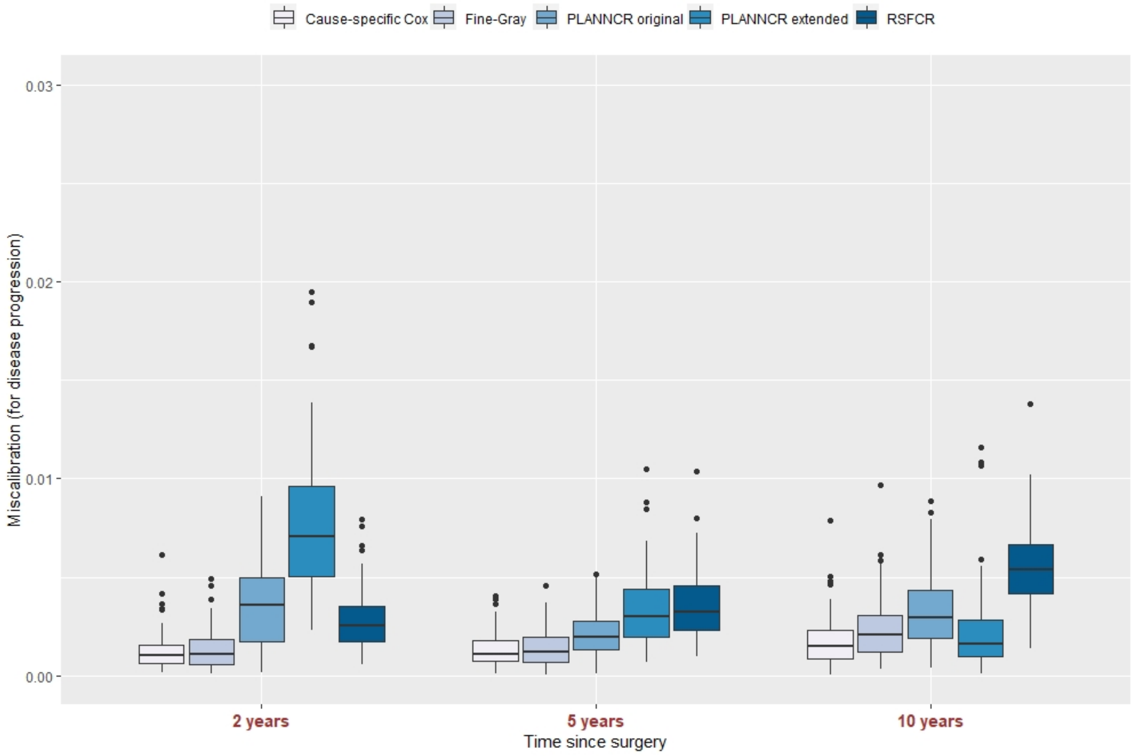


Figure 8.3: Miscalibration of cause-specific Cox model, Fine-Gray model, PLANNCR original, PLANNCR extended (tuned with Brier score at 5 years), and RSFCR at 2, 5, and 10 years for the event of interest: disease progression based on 100 validation datasets. Miscalibration was calculated as the mean squared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

9 General discussion

This thesis focused on statistical analyses aimed at improving clinical trial design on behalf of the European Organisation for Research and Treatment of Cancer - Soft Tissue and Bone Sarcoma Group (EORTC - STBSG) and on investigating the potential of survival prediction models with machine learning techniques compared with statistical models. Sarcoma and non-sarcoma clinical data were used to compare the performance of the different prediction models. Results were presented in two different parts.

9.1 Part I: Clinical trials in soft-tissue sarcomas

In **Part** I, modern benchmarks were estimated to design new phase II clinical trials for common histotypes of locally advanced or metastatic soft-tissue sarcoma (STS) patients based on two meta-analyses which originated from an in-house literature review of STS (**Chapters** 2, 3). As an extra step, the prognostic significance of bone metastasis at study entry in STS was investigated to identify high-risk patient populations based on a pooled analysis of five EORTC - STBSG clinical trials (**Chapter** 4).

9.1.1 Designing new phase II trials

The historical benchmarking analysis by Van Glabbeke *et al.* (2002) provided pooled progression-free rates for various STS patients who participated in phase II trials of EORTC - STBSG (see table 1.1) [1]. These have been used to design a large number of new studies (currently more than 420 citations). This is the first gap that this thesis aimed to cover. We performed an extensive in-house literature search to identify all phase II, III, or IV clinical trials of advanced or metastatic STS (from 2003 to 2018). Due to the presence of heterogeneity among clinical trials, we decided to provide an update focusing on the most prevalent STS identified during the review. In **Chapter** 2 [2], we re-evaluated thresholds with a meta-analysis for first-line and pre-treated leiomyosarcoma patients (all or uterine only) using progression-free survival rate (PFSR) as the primary endpoint. Information was acquired from publications and sponsors. Reference values for parameter P_0 (null hypothesis) were calculated as the overall pooled PFSR at 3 and 6 months ((based on inactive and active agents), whereas minimum values to target for parameter P_1 (alternative hypothesis) were calculated using the recommended treatment effect for PFS by the ESMO Magnitude of Clinical Benefit Scale [3]. In **Chapter** 3, another meta-analysis was performed by using the same methodology, for advanced or metastatic liposarcoma and synovial sarcoma patients to provide benchmarks to design new phase II clinical trials with PFSR. Table 9.1 summarises P_0 and P_1 parameters based on the meta-analyses.

A need to raise the bar of thresholds is indicated for the commonest types of STS in future phase II trials by both of our studies, which aligns with the perspective of the American Society of Clinical Oncology [4]. The cost-benefit of new systemic therapies for cancer should be balanced against the societal resources in this era of rapidly rising

	3 months		6 months	
Treatment line and analysed group	**Ref (P_0)**	**Min target (P_1)**	**Ref (P_0)**	**Min target (P_1)**
First-line uterine LMS	71%	80%	49%	63%
First line all LMS	74%	82%	58%	70%
First-line LPS	69%	79%	56%	69%
First-line SS	74%	82%	56%	69%
Pre-treated uterine LMS	53%	66%	42%	57%
Pre-treated all LMS	48%	62%	28%	44%
Pre-treated LPS	49%	63%	28%	44%
Pre-treated SS	45%	60%	25%	41%

Table 9.1: **Treatment effect (PFSR) for the parameter P_0 (null hypothesis) and the parameter P_1 (alternative hypothesis) of a study**. LMS, leiomyosarcoma; LPS, liposarcoma; SS, synovial sarcoma. The meta-analysis for all LMS excluded trials designed for uterine LMS patients.

healthcare costs.

Strengths and limitations of the meta-analyses

To the best of our knowledge, these are the first attempts to meta-analyse the outcome of patients with advanced/metastatic leiomyosarcoma, liposarcoma, or synovial sarcoma for both first or further lines of treatment. Overall, a key strength of these two studies (**Chapter** 2, **Chapter** 3) is the collection of summary estimates for 1500 leiomyosarcoma (and 421 extra patients from clinical trials designed exclusively for uterine leiomyosarcoma), 1030 liposarcoma, and 348 synovial sarcoma patients from phase II, III, or even IV clinical trials.

Our results cannot be directly compared with the Van Glabbeke study for several reasons: (i) our meta-analyses used patients from phase II, III or IV clinical trials whereas the historical benchmarking analysis used patients who participated in phase II trials only, (ii) the primary endpoint shifted from progression-free rate to PFSR (counting any death as an event), (iii) the 2002 publication exploited individual patient data from the EORTC - STBSG database whereas we used summary estimates retrieved from the publications or the sponsors, (iv) Van Glabbeke *et al.* defined thresholds mixing all STS subgroups in the pre-treated setting, (v) the historical study calculated reference values for drug inactivity (P_0) and separately for activity (P_1) whereas we defined P_0 as the overall pooled rate (based on inactive and active agents) and estimated the values to target for P_1 based on the ESMO Magnitude of Clinical Benefit Scale.

Our studies have some limitations. First, the large majority of the trials were designed for several STS types and were therefore underpowered for specific subgroup analyses (i.e., for leiomyosarcoma). Moreover, effects seen in subgroups may not necessarily be visible in the overall study, thus not prompting an update of the recommendations for standard of care. In addition, some treatments were still in the stage of validation and were not part of the recommended treatments. This might explain the non-significant difference between recommended and non-recommended treatments based on the standard ESMO 2018 or 2021 guidelines [5, 6] for all leiomyosarcoma (first-line or pre-treated), first-line liposarcoma / synovial sarcoma, and pre-treated synovial sarcoma patients. Therefore, we used the overall pooled PFSR to define parameter P_0 in our meta-analyses for the sake of consistency. Secondly, the condition of any meta-analysis that the effect sizes between drugs of the same trial are independent may be violated in the randomised studies, as a random-effect model was used for each treatment regimen. Here, we observed high unexplained overall heterogeneity indicative of a large variation between effect sizes, which may limit our meta-analytic results. Thirdly, PFSRs were calculated based on summary estimates per treatment arm and treatment line that are less reliable than individual patient data but require a smaller amount of time to be collected from the different study sponsors. Moreover, in **Chapter** 3, liposarcomas were addressed as a single disease while it is known that there are three different histologic subtypes (e.g., well differentiated/dedifferentiated, myxoid, or pleomorphic) that exhibit different clinical behavior and sensitivity to treatments. Yet, in older studies such information might have not been collected at the subtype level.

Towards a histology-tailored research

In the last decades, STS studies were designed based on the one-size-fits-all principle mixing several histologic subtypes. However, more recently research has shifted to a more histology-specific approach to better diversify the eligibility criteria of clinical trials [7–10]. This is also something we have noticed in our literature review (i.e, some studies designed only for uterine leiomyosarcoma / leiomyosarcoma, dedifferentiated liposarcoma / liposarcoma, angiosarcoma). An urgent need remains for the development of individualised treatment plans such as targeted therapy or immunotherapy to move away from the conventional chemotherapy options. Hence, new studies tailoring therapy to specific histological types should be based on modern thresholds for drug activity. With this work, we suggested a new benchmark to aid the design of phase II studies for leiomyosarcoma, liposarcoma, or synovial sarcoma patients using PFSR at 3 or 6 months as the primary endpoint.

The choice of primary endpoint in phase II

In general, the choice of the endpoint in phase II exploratory studies should be tailored on the disease and the drugs under investigation (mechanism of action, potential toxicity). Response-based endpoints defined by RECIST 1.1 [11] - such as the objective response rate - might be appropriate primary endpoints if unambiguous and clinically relevant antitumor activity (tumor shrinkage) is hypothesized by a drug or combination [12]. However, in contrast, if response-based endpoints are not appropriate, PFS (and/or time-to-progression) can be considered the primary endpoint as biological activity is frequently not expected to result in shrinkage of lesions, but rather in stabilisation of disease. In our meta-analysis databases, rather low response rates were observed (frequently < 15%) for the majority of the drugs / drug combinations. This was expected for these STS types (leiomyosarcoma, liposarcoma, synovial sarcoma), as a decrease of tumor volume > 30% is unlikely with current standard of care for STS. Hence, PFS (or PFSR at 3 and 6 months) is a valuable alternative endpoint for the estimation of the biological antitumor activity of a new treatment (e.g., a targeted therapy) and to justify further investigation in confirmatory phase III trials [1].

The ultimate aim of trials

Clinical research has as an ultimate aim is to provide evidence of improved OS or improved quality of life. Nonetheless, strong surrogacy properties between PFS and OS are questionable based on two meta-analyses of randomised clinical trials with advanced STS [13, 14]. Therefore, PFS carries the risk of misleading conclusions because of erroneous extrapolation of the results, which might lead to exaggerated enthusiasm for a new anticancer therapy [15, 16]. On the other hand, PFS remains an attractive endpoint to identify benefit earlier than OS, and phase II trials are not intended to provide definite proof of the new treatment but rather a justification to further investigation. PFS (or the 3- or 6-month PFSR) can thus be used as the primary endpoint in phase II trials or as futility endpoint in phase III trials, but OS should remain the primary endpoint in phase III trials (whenever possible).

Challenges in rare cancers

Rare cancers account approximately for one-fifth of new cancer cases and patients typically have a poor median survival time [17, 18]. STS - the rare malignancies discussed in this thesis - account for about 1% of all adult tumours [19]. Over the years, more than 100 histologic subtypes have been recognised with widely varying presentation, sensitivity to treatment, and long-term outcomes [7]. Hence, each of these histological types and subtypes is a rare or even an ultra-rare indication, which brings challenges to design new studies and to build clinical evidence to advance treatment options.

Some important challenges can be summarised as [20]:

- Lack of clinical expertise in hospitals. For instance, it is well-known that management of STS should be carried out in sarcoma reference centers and/or within reference networks that share multidisciplinary expertise [6].
- Smaller incentive by the drug developers due to market limitations. It is very hard to convince a pharmaceutical company to invest money for a histology-specific trial on an ultra-rare cancer subtype (e.g. myxoid liposarcoma in STS).
- A rigorous study design requires a large number of patients. According to conventional methodologies, clinical trials need considerable numbers of patients that are difficult to collect in rare cancers such as STS or their ultra-rare subtypes. Such studies require extensive international collaboration.
- Selection of experimental treatments is often based on inadequate evidence. Consequently, accommodating a higher than average degree of uncertainty is necessary for clinical as well as population-based decision making.
- Randomised controlled trials - which are usually the gold standard - might not be feasible due to limitations in patient availability. This is also a typical problem for sarcoma studies.
- Data collection to perform meta-analysis can be very time-consuming. Our meta-analyses in **Chapter** 2 and **Chapter** 3 provide useful information to aid the design of new single arm phase II exploratory studies for leiomyosarcoma, liposarcoma, and synovial sarcoma patients. For this work, we collected summary estimates from publications and sponsors ($n \geq 10$) which is typically less problematic than collecting individual patient data due to the General Data Protection Regulation (GDPR) restrictions. However, collecting data based on published literature might not be feasible for rarer STS entities. Collection of individual patient data is likely to be needed, which will increase the complexity and the time required to complete such a project.

Methodological recommendations for clinical trials in rare cancer are discussed in the paper by Casali *et al.* [17]. Methodologies should be refined to combine all the available evidence.

9.1.2 Patients with bone metastasis at diagnosis

We were also interested to expand on two previous studies of the EORTC - STBSG [21, 22] to identify high-risk patient populations in clinical trials performed by our group examining patient characteristics. 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 [23]. Hence, in **Chapter** 4 [24], we investigated whether bone metastases at presentation affect overall survival (OS) and progression-free survival (PFS) per treatment line (first versus second line or higher) based on a pooled analysis of five EORTC - STBSG clinical trials [25–29]. Patients were enrolled from April 2003 to June 2015. For the subgroup of patients with bone metastasis presence (n = 140, 13.5%), the strongest metastatic combination was identified between bone and another site (liver, lymph node, lung, soft-tissue, other).

There is an increased occurrence of bone metastasis in advanced-stage cancers [30]. A metastatic bone profile may be part of STS patients' natural history, which negatively affects their prognosis. In this pooled analysis, patients with STS of the extremities, abdomen, thorax, visceral or other sites of primary tumour were included (n = 1034). The unadjusted difference in OS/PFS for patients with or without bone metastasis was statistically significant for first-line treatment (hazard ratios 1.55 [95%-CI: 1.19–2.01] and 1.43 [95% CI: 1.12–1.84] for OS and PSF, respectively). However, this difference was not significant when adjusting for 12 known prognostic factors (hazard ratios 1.33 [95%-CI: 0.99–1.78] and 1.31 [95%-CI: 1.00–1.73]). For second line or further, the unadjusted hazard ratios for OS / PFS were 1.10 (95%-CI: 0.81–1.49) / 1.06 (95% CI: 0.80–1.40) and the adjusted 1.11 (95%-CI: 0.81–1.52) / 1.07 (95%-CI: 0.80–1.43), respectively. An overall worse status is suggested for patients suffering from bone metastasis in any treatment line, although the effect was not statistically significant.

In our database, 6/1034 patients had exclusively bone metastasis at diagnosis, and therefore a separate analysis of this small subgroup could not be performed. Tentative explanations of this small number could be that (i) bone lesions alone are typically challenging to measure and most trials require a measurable disease to assess response/progression per RECIST 1.1 criteria [11], (ii) bone metastasis at diagnosis is a sign of extensive disease. A limitation of this work is the retrospective exploratory nature. Both randomised and nonrandomised studies were pooled together from the EORTC - STBSG database to increase the statistical power, which is likely to have introduced some selection bias in the population.

A subgroup analysis was performed for 140 patients (13.5%) with bone metastasis at presentation to identify the metastatic organ site combination that is the most detrimental for OS and PFS after adjusting for demographic characteristics, histological entity, tumour grade, site of primary tumour, and time between histological diagnosis and registration/randomisation. A combined bone/lymph nodes metastatic presentation had the worst OS prognosis (hazard ratios 2.97 [1.53, 5.78] for first-line and 1.59 [0.55, 4.54] for second-line or higher treatment). For PFS, bone plus lung metastasis was the most detrimental combination (hazard ratios 2.80 [1.10, 7.09] and 2.45 [0.61, 9.84], respectively). Of note, such combinations were statistically significant for first-line treatment. Due to the limited number of patients per treatment line (77 first-line, 63 second-line or later), results should be interpreted with caution.

Our findings may have some implications on managing advanced/metastatic STS patients with bone metastasis at diagnosis. The impact of skeletal metastases was more profound in first-line treated population. Presence of bone metastasis at study entry was not ascertained (statistically) as a sufficiently important risk factor on its own for first or second line or further to justify stratification in randomised studies for these patients. As individuals with metastatic STS survive these days somewhat longer than 20–25 years ago due to advances in supportive and multidisciplinary care, the prevalence of bone invasion is difficult to be verified and an increase is likely to be observed in the future.

9.1.3 Future perspectives

In this subsection, we provide some recommendations for future research in STS.

The design of new studies

STS subtypes are very heterogeneous rare malignancies with widely varying presentation, sensitivity to treatment, and long-term outcomes [7, 31]. The work provided in **Chapter** 2 and **Chapter** 3 will help to optimise the design of new histology-tailored phase II trials for common histotypes (: leiomyosarcoma, liposarcoma, synovial sarcoma) in advanced or metastatic setting. Further meta-analyses should be designed and performed by using summary estimates or individual patient data for other frequent types such as undifferentiated pleomorphic sarcoma or angiosarcoma. Efficient data sharing could be key to simplify this process, as well as an international collaboration between data centers. We would like to stress the importance of designing robust and efficient phase II studies especially in rare cancers, where there is lack of evidence to support clinical decisions. Our projects have yielded a good historical control database with updated reference values which can be used to design new single arm STS studies (with a sufficient sample size and within a reasonable time-frame). This will lead to a better early evaluation of drug activity as an urgent need remains for the development of personalised treatment plans such as immunotherapy to move away from the conventional chemotherapy options.

Randomised controlled trials are considered to be the gold standard for clinical development, but are challenging to run for simple screening studies in STS mainly due to sample size requirements, and small incentive by pharmaceuticals. Too many randomised controlled phase III trials have recently failed in a rush for success [32]. The development of a new drug might be easier and faster with adaptive randomised trials such as the transformation of a phase II into a phase III (”seamless phase II/III designs”) in case of a positive early stage of a study, or the use of ”drop-the-loser” or ”play-the-winner” designs [17]. Rare tumors require extremely good and innovative trial design to make optimal use of every patient. Importantly, in order to make progress in this context, practice changing decisions will have to be based on less direct evidence than those in common malignancies and all available evidence should be evaluated. With our meta-analyses, we provided new robust benchmarks to speed up the process through single arm studies which are faster to run, and require fewer patients than randomised phase II trials.

As an example, the Pautier 2015 non-randomised phase II study [33] investigated the potential of adding trabectedin to doxorubicin for first-line advanced uterine or soft-tissue leiomyosarcoma patients. For soft-tissue leiomyosarcoma PFSRs were 90% and 81% at 3 and 6 months (for uterine leiomyosarcoma 87% and 72%), respectively. Therefore, this drug combination even if it was not recommended based on the ESMO 2018 guidelines [5] would qualify according to our updated rules for all or uterine first-line leiomyosarcoma patients (see minimum targets in table 9.1) suggesting a significant and relevant improvement over a standard of care in a prospective randomised phase III trial. As a verification, the results of the phase III trial with 67 uterine leiomyosarcoma and 83 soft-tissue leiomyosarcoma patients are positive [34]. There is a significant increase in PFS compared with the single doxorubicin arm, as well as a benefit in terms of OS. This highlights the pertinence of good historical benchmarks for simple screening phase II studies, which will lead to successful phase III confirmatory trials.

The importance of collaboration and funding

The rarity and heterogeneity of STS highlights the importance of a multidisciplinary approach (i.e., a specialised team of radiologists, pathologists, surgical and medical oncologists) for the management of the disease [7]. A major strength of sarcoma research is collaboration in an academic climate. Progressive multinational collaborative efforts (many sites in several countries) will allow both sufficiently large and sufficiently focused studies to generate high-quality evidence in specific STS patient populations. Such efforts are more than necessary when fighting rare cancers. The EORTC offers an integrated approach to drug development and drug evaluation programs through translational and clinical research of its rich database. The EORTC - STBSG can connect people and sites, as well as start working groups with other national sarcoma societies / groups to reinforce the academic climate and lead to major breakthroughs. The STBSG and other national sarcoma groups should strengthen the relationship with the European Medical Agency (EMA) and U.S. Food and Drug Administration (FDA) and discuss with the regulators regarding the conditions for a new drug approval. In addition, the development of new statistical methodology is more than welcomed to address research questions, and strong collaborations between clinicians and statisticians are key to bring new project ideas into maturity. Overall, in rare cancers such as STS, flexibility and out-of-the-box thinking are required to advance research.

Simultaneously, it is necessary to initiate high value and high-quality company sponsored trials through the interaction with pharmaceutical companies. Major congresses e.g. the European Society of Medical Oncology (ESMO), the European Society for Radiation Oncology (ESTRO) can be used as platforms for this interaction. Next to that, it is also of great importance to secure funding for academic studies and search for national grants or support from industry. The EORTC 1809 STRASS II trial in patients with high-risk retroperitoneal sarcoma belongs to an alternative pipeline of purely academic trials. It addresses a significant unmet need (chemotherapy plus surgery versus surgery only) in this population [35]. The trial recently received support by the Anticancer Fund (ACF) a Belgian foundation of public utility dedicated to expanding the range of treatment options available to cancer patients regardless of commercial value. This shows how essential is the allocation of research funds to areas with limited interest from the profit-driven cancer industry to complement progress.

9.1.4 In conclusion

The first part of this thesis yielded modern benchmarks of efficacy to design new phase II clinical trials for locally advanced or metastatic leiomyosarcomas, liposarcomas, and synovial sarcomas. These meta-analyses were essential to update and re-evaluate well-established historical thresholds by the EORTC - STBSG. Expanding previous STBSG studies, skeletal metastasis at baseline was found to be detrimental for both OS and PFS in any treatment line with a more profound effect in first-line population, although not statistically significant.

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

In **Part** II (**Chapters** 5, 6, 7, 8), the potential of existing and new machine learning (ML) methodologies was explored for survival prediction, and compared with traditional statistical models (SM) for real-life clinical data (small/medium or large sample sizes, low- or high-dimensional settings).

9.2.1 Machine learning versus statistical modelling

Since the last decade, machine learning (ML) has received increased attention in the medical area. The aim of prediction has been of particular interest as part of a growing trend towards personalised medicine [36]. However, concerns have been raised that the employment of ML techniques and artificial intelligence in general is over-hyped in some contexts (e.g., over-fitting the training data, lack of attention towards validation, unsuitable performance measures). Overall, from thousands of publications applying ML to medical data, very few algorithms have meaningfully contributed to clinical care [37]. One of the main reasons is that if a model fails in healthcare the consequences are life-threatening, and thus robust evidence is required [38].

In this thesis, ML methods were explored for prediction of survival data. Due to the presence of censored observations, the extension of ML methods to survival data is not straightforward. Over the years, a variety of methods have been proposed which are adaptations of the ML classifiers (for example random survival forests from random forests, survival neural networks from neural networks [39]). We investigated the potential of these techniques in contrast with conventional statistical benchmarks for time-to-event data and proposed a new extension to the partial logistic artificial neural networks methodology.

A review and critical appraisal

In **Chapter** 5, we performed - to the best of our knowledge - the first ever large-scale review on survival neural networks (SNNs) using prognostic factors for clinical prediction. Our goal was to provide a broad understanding of the literature (1st January 1990 - 31st August 2021). A total of 24 articles were identified based on a global search in PubMed. Relevant manuscripts were classified as methodological/technical (novel methodology or new theoretical model; 13 studies) or applications (11 studies). We discussed how SNNs are employed for prediction in the medical field and how researchers have tried to adapt a classification method to right-censored survival data. There are two methodological trends: either time is added as part of the input features and a single output node is specified, or multiple output nodes are defined for each time interval.

This work was supplemented with a critical appraisal to pinpoint current limitations and identify future research directions. Regarding some general characteristics of the studies, the median total sample size was 920 patients and the median number of predictors was 7 (low-dimensional data). Medical applications were mainly in the field of oncology (73.5%, 25 datasets). The strategy used to address the missing data (if any) was unclear for 9/21 (42.9%) studies (without 3 simulation studies). Major findings included inaccurate model development/validation and poor reporting. In the majority of studies (15, 62.5%), the approach to tune hyperparameters was unclear. The performance criterion for model development was unclear for 6 studies (25.0%). Programming language used for the development of the SNN was unclear in 7 studies (29.2%). We noticed large variability and improper performance measures for survival data, as well as lack of confidence intervals for the predictive measures in 13 of the 24 studies (54.2%). Calibration plots were available for only 11 studies (45.9%). All in all, 19 studies reported comparisons between Cox models and SNNs from which 15 studies (78.9%) did not consider interaction terms between the predictors.

According to these findings, some general recommendations are provided:

- Complete and transparent reporting of modelling steps and analysis is necessary (e.g., more details on training and test data), to enable reproducibility, and to allow critical appraisal of the results by a wider audience [40, 41].
- Hyperparameter selection and training should be more extensive with the performance criterion for model development clearly reported. A suitable performance measure should take into account the censoring mechanism.
- More attention to model calibration is urgently needed. Calibration should be assessed preferably through calibration plots.
- Larger datasets and/or more predictors are needed for better model development/validation and improved generalisability.
- Comparisons of SNNs with conventional regression models should be made in a fair manner, with the conventional models fully developed and interactions and/or non-linear terms included when appropriate.
- Further methods and guidelines for obtaining confidence intervals are needed. Multiple resampling of all empirical data using bootstrapping can be an advantageous approach when sample size is limited, as it avoids the need to split the data for model development and provides confidence intervals [42].
- Increasing the complexity of a SNN (or a ML prediction model) does not necessarily translate to improved performance on new clinical data. For such survival data, sample size and number of predictors is likely to be insufficient for employing advanced techniques.

These aspects are of great value as suboptimal clinical prediction models with ML or statistical modelling are responsible for research waste [43, 44].

Comparison between methods in different settings

A main objective of this thesis was to compare machine learning (ML) techniques with statistical models (SM) for time-to-event prediction models in the presence of right-censored medical data (liver transplantation, osteosarcoma, and STS).

Chapter 6 [45] provided the first ever study where ML techniques were tested on complex post-transplant liver data (large sample size, high-dimensional setting) from the United States and compared with traditional Cox models. Random survival forests (RSF) [46] and two novel extensions of the partial logistic artificial neural network (PLANN) [47] (neural networks with one hidden or two hidden layers) were applied to retrospective data (n = 62294, p = 97) provided by the Scientific Registry for Transplant Recipients to predict survival. These methods were compared versus three Cox models (with all variables, backward selection and LASSO) [48, 49]. Clinical endpoint was overall graft-survival (time between transplantation and the date of graft-failure or death). To assess the final predictive performance of the models, the concordance index [50], Brier score [51], and Integrated Brier Score (IBS) [52] were used. The strongest prognostic factors were identified for each model. RSF performed better than the Cox models in terms of C-index (0.622 for RSF versus $\leq$ 0.62 for Cox models). This shows the ability of RSF to discriminate better between low and high risk groups of patients. The Brier score was measured at each year for all methods. RSF showed results similar to the Cox models having slightly smaller total prediction error on the test data (IBS 0.182 versus 0.183). The neural networks (IBS 0.180) performed in general better than the Cox models or the RSF and had very similar performance over time. From the three ML techniques, PLANN extended with one hidden layer predicted survival probabilities most accurately. Its calibration was very similar to the Cox model with all variables. The RSF and the PLANN extended with two hidden layers were less calibrated on test data. Special emphasis was given on the interpretation of the models. An indirect comparison was performed to examine which are the most prognostic variables for a Cox model with all variables, a RSF and the two PLANNs extended. The Cox model with all variables and the PLANNs identified *re-transplantation* as the strongest predictor and *donor age*, *diabetes*, *life support* and *race* as relatively strong predictors. According to RSF, the most prognostic variables were *donor age*, followed by *re-transplantation*, *life support* and *serology status of Chronic hepatitis C virus*.

In **Chapter** 7 [53], SM were compared with ML for non-complex clinical data (small/medium sample size, low-dimensional setting) to investigate a different real-life setting. The dataset originated 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 [54]. A Monte-Carlo simulation study was performed to compare PLANN original [47] or extended (with one hidden layer) [45] with Cox models [48]. Real-life clinical data was mimicked to simulate synthetic data (5 predictors, 250 or 1000 observations) and to address different scenarios (20, 40, 61, or 80% censoring) in the absence of complex functional dependence relationships involving time and covariates. The endpoint of interest was overall survival (time to death from any cause since surgery). Models were evaluated in terms of C-index [50], Brier score at 0-5 years [51], IBS at 5 years [52], and miscalibration at 2 and 5 years in the simulated test datasets. It was shown that SNNs may reach a comparable performance in terms of the C-index, Brier score, or IBS. The standard deviations (over 1000 repetitions) overlapped to a large extent for all scenarios. Predictive performance improved (smaller Brier scores, higher C-indexes) when the sample size increased for all methods. Predictive ability was adequately robust to predefined adverse scenarios on training data (removing patients censored before 2 years, administrative censoring at 5 years). However, the Cox models were usually better calibrated (predicted survival probabilities were closer to the observed). This highlights in particular the relevance of reporting calibration of ML techniques to obtain a neutral comparison with SM. Miscalibration was rather strong for a larger percentage of censoring (less events).

In **Chapter** 8, the comparison between ML techniques and traditional SM was extended to competing risks (CRs) framework in another simple clinical setting (small/medium sample size, low dimensional data). Our aim was to develop and validate clinical prediction models for CRs with the first ever study of this kind in STS. A dataset with 3826 retrospectively collected patients with extremity STS and nine predictors from the PERsonalised SARcoma Care (PERSARC) Study Group was used. Three ML techniques a) PLANN for CRs (PLANNCR) original [55], b) PLANNCR extended (a new method developed by the authors), and c) RSF for CRs (RSFCR) [56] as well as two statistical models i) cause-specific Cox [48], and ii) Fine-Gray model [57] were compared. The endpoint of interest was time in years between surgery and disease progression (event of interest) or death (competing event). Predictive performance of the methods was assessed for the event of interest and the competing event in 100 validation datasets based on the Area Under the Curve (AUC) and the Brier score at 2, 5, or 10 years (t-year predicted risks evaluation) [58]. Miscalibration (absolute predictive accuracy) was estimated at the same time points [59]. Results showed that ML models have similar performance with SM in terms of Brier score and AUC at 2, 5, and 10 years for disease progression and death (95% confidence intervals overlapped). From the three ML models, predictive ability of PLANNCR extended was usually better than RSFCR and PLANNCR original especially in terms of AUC. This means that PLANNCR extended was able to better discriminate between low and high risk groups of patients. Nevertheless, the SM were frequently better calibrated than the three ML methods. Miscalibration of PLANNCR original and extended was more pronounced for the competing event (death). These findings indicate that more attention to model calibration is urgently needed for ML methods.

Advantages and disadvantages

Pros and the cons of SM and ML techniques for survival analysis are presented below in terms of interpretability, flexibility, and practical utility.

(a) **Interpretability**: Cox model - the well-established statistical benchmark for survival data - offers a straightforward interpretation via hazard ratios, which is very useful for clinicians to make informed decisions. On the contrary, ML techniques usually have limited interpretability. For example, in **Chapter** 6 we used the relative importance and the variable importance methods for neural networks and RSF to extract model interpretation, respectively [45]. However, these cannot indicate whether the effect of a covariate is protective or not.

(b) **Flexibility**: ML techniques make minimal assumptions and are very flexible as they can model automatically non-linear relationships between variables. On the other hand, SM are less flexible as they make some usually strong assumptions for the modelling process. For instance, the Cox model assumes proportionality of hazards over time and additivity of effects (as any regression model). Hence, any interactions between variables need to be manually pre-specified, which can be problematic in the presence of many variables and multiway interactions.

(c) **Practical utility**: SM such as the Cox model have a fast implementation in popular open-source programming languages such as `R` or `Python`. Then again, ML techniques require a nontrivial implementation time for data pre-processing, tuning of hyperparameters (the larger the number of hyperparameters the more the time and effort needed for model training), and are computationally more intensive to run. Typically, calculation complexity is based on sample size and predictors multiplicity. Moreover, model optimisation of neural networks is a delicate task which requires robust numerical methods and skillful use, else the network might converge in suboptimal minima in the error function [60, 61].

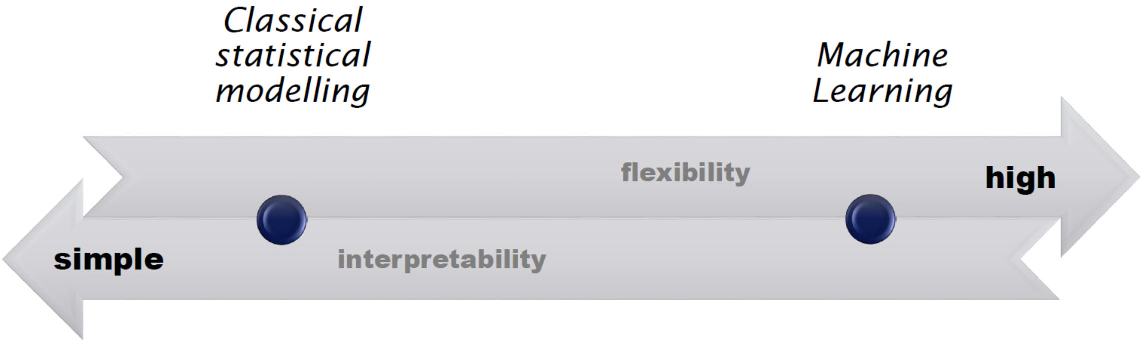


Figure 9.1: The trade-off between model interpretability and flexibility with SM and ML techniques (Wallisch *et al.* 2019 [62]).

Figure 9.1 presents the trade-off between simple interpretability and high flexibility with SM and ML techniques. Classical SM are typically placed on the left side of the arrows (simple interpretability, low flexibility). By adding interactions and non-linear terms, the flexibility of a model can be increased at a cost of interpretability. ML models are usually positioned towards the right side of the arrows (complicated interpretability, high flexibility). However, there are some ML techniques such as survival trees which can be placed near the left side. In general, an increase in model complexity will lead to higher flexibility, but a more complicated interpretability. SNNs developed in recent years usually have more complicated structures and make use of multiple hidden layers (deep learners). It should be noted, however, that increasing the complexity of an ML prediction model does not necessarily translate to improved performance on new clinical data. An increase in the complexity, and by extent flexibility of a network, may produce a model that is too attuned to the training data with poorer generalization to new data (overfitting), with the extra cost of a more limited interpretability.

Key considerations

Table 9.2 presents some important considerations for choosing between SM and ML. A much larger sample size might be necessary when using ML approaches to develop risk prediction models [42]. A main cause for this problem is that the number of predictor (also called ”feature”) parameters considered by ML will usually be substantially larger than that for regression (even with the same set of predictors) as these techniques model automatically second and higher order interaction terms. Therefore, they might actually need ”big data” to ensure small overfitting of the developed models [63]. On the contrary, the sample size of most medical research datasets is more appropriate for regression models such as the Cox proportional hazards model for time-to-event data. Furthermore, regression models provide a transparent model equation and allow for a straightforward interpretation which is likely to be crucial for clinicians to implement a model going forward in routine care practice.

Statistical models	**Machine Learning**
Focus on relationships between independent variables	Focus on prediction
Favor additivity of predictor effects	No special emphasis to additivity of effects
Low uncertainty tolerance (confidence intervals, hypothesis testing, assumptions)	High uncertainty tolerance (adaptability, no assumptions)
Small - medium sample size	Large - huge sample size
Small number of predictors	Large number of predictors
Semi-parametric, parametric models	Non-parametric models
Simple interpretability	Complicated interpretability
Low flexibility (small number of interactions pre-specified)	High flexibility (higher order interactions presence is expected)
Low signal to noise ratio (human outcomes, weather forecasting)	High signal to noise ratio (image recognition, playing games)

Table 9.2: **Key considerations for choosing between SM and ML** [64, 65]. The first columns shows when SM might be the better choice for analysis, whereas the second column indicates when ML can be the better choice.

9.2.2 Future perspectives

In this subsection, we discuss about the future of ML for time-to-event data.

Challenges of machine learning

While throughout the years a substantial increase in ML articles in medical research can be observed, there are few algorithms actually implemented in clinical practice [37, 66]. Below, we report some major challenges which need to be addressed prior to the establishment of ML methods.

- **Access to medical data.** Clinical data are (rightfully) challenging to access. An informed consent and ethics committee approval, as well as appropriate handling (anonymisation) are required before data sharing [67]. Limited public data availability also hinders the validation of ML algorithms and their results by other authors. Findings obtained with a set of methodologies are not easily comparable with different datasets (as they reflect other real-life settings).
- **Model complexity in relation to the amount of available data.** Clinical data typically include tens to thousands of patients. This number is usually not sufficiently large for the employment of modern flexible ML techniques such as neural networks which are data hungry [63, 68]. Hence, there is a substantial risk of overfitting (excessive tailoring of algorithms on training data), thus limiting the ability of these models to generalise (accurately perform) on new data. The use of shallower neural networks (one hidden layer) could reduce the risk of overfitting. ML techniques should preferably be employed for large datasets.
- **Clinical validation.** It is the most critical component of performance assessment. The presence of missing data, regional variations in practice, and logistical/infrastructural limitations may complicate model validation [69]. If possible, an external validation (e.g., using population from different centers) should follow a meticulous internal validation. Transparent reporting of ML prediction models is required [40, 44]. Both aspects of model discrimination and calibration should be taken into account for performance evaluation

[41, 70]. Advantages and shortcomings in comparison with traditional benchmarks (SM) should be documented. Recently, researchers are increasingly aware of the scientific rigor needed to be demonstrated prior to the employment of a ML model in clinical practice.

- **Model interpretation.** This is a key obstacle for the integration of ML methods in medicine. For example, (deep) neural networks are called ”black boxes” since they develop complex internal functions that are hard to interpret. The extraction of a meaningful model interpretation is critical for clinicians to take informed decisions and to trust these models in routine care. A common metric is needed to directly compare SM with ML. Research in this area will likely be on the spotlight for many years.

Research directions

In this thesis, we investigated the predictive ability of different ML techniques: RSF [46], PLANN original [47], PLANN extended [45, 53] for right-censored clinical data. Models were compared with Cox proportional hazards models [48] in different real-life settings. The work was also extended for CRs examining RSFCR [56], PLANNCR original [55], and PLANNCR extended versus the cause-specific Cox model and the Fine-Gray model [57]. In the future, research for ML models should also focus on left- and interval-censored data.

Several ML models have been developed and applied to deal with right-censored medical data. Wang *et al.* provide a comprehensive review of commonly used methods in survival analysis [39]. Such ML techniques include survival trees [71], SNNs some of which have been included in our review (**Chapter** 5, e.g. [72, 73]), support vector machines [74], or other ensemble methods such as boosting [75]. It would be interesting to compare the predictive ability of these techniques with traditional statistical benchmarks in a variety of settings (small or large sample size, low or high dimensional data) with actual or synthetic data (simulation studies) to establish their potential role in clinical practice.

Both traditional Cox models and PLANNs allow for the inclusion of time-dependent covariates. Cox models can incorporate these variables in standard software. PLANNs can naturally incorporate time-dependent covariates due to the essential data transformation into a long format for each patient. In the future, it might be useful to compare the predictive ability of these models for time-dependent variables (also between the cause-specific Cox model and PLANNCRs for competing events). Moreover, RSF (for a single event), and RSFCR, Fine-Gray model (for competing events) can be extended to provide dynamic predictions with time-dependent covariates by creating a landmark dataset at a set of landmark time points t_{LM} [52, 76].

Last but not least, a subsequent step after a rigorous evaluation of a new prediction model with ML is its implementation in open source software. Popular programming languages such as `R` or `Python` can be used to integrate new methodologies. Research should also focus on the uniformisation of packages between different software. This will make a larger number of methods widely accessible, and will lead more academics and other interested individuals to engage with them. As a consequence, this competition will promote good standards of practice, and will help with the development of more methodological extensions for survival data.

Fuelling the debate

ML and artificial intelligence will be on the spotlight of medical research for many years. Hence, it is of paramount importance to develop methods and assess their predictive performance against conventional statistical benchmarks. Appropriate performance measures should be selected to evaluate model discrimination / calibration. For survival data the censoring mechanism should be taken into account [52]. A complete and transparent reporting of all modelling steps is required to enable critical appraisal and allow reproducible analysis. Researchers should follow the best practice guidelines and recommendations on clinical prediction models [40, 41, 44, 77, 78].

Another necessary aspect to consider is the intended purpose of a ML prediction model in healthcare. Existing or new methods should address unmet needs such as research, benchmarking, or bedside application. Perhaps a model has some novelty (addressing diseases or outcomes where information is not available), or it shows clinical usefulness improving discrimination (separation of low- from high-risk patients) compared with current practice [78]. Or perhaps the application of ML techniques is motivated by exploration of the collected medical data to assess linear and additive model assumptions. Being aware of the advantages / disadvantages and key aspects of ML techniques and SM can help the reader to take meaningful decisions regarding the choice of methodology, and provide her with enough stimuli to seek for new developments.

9.2.3 In conclusion

In the second part of this thesis, we performed a review and critical appraisal that shed light on the current state of art of SNNs in medicine with prognostic factors for clinical prediction (January 1990 to August 2021). Furthermore, we extended existing methods in ML and compared ML techniques with SM for prediction in different real-life settings. For complex liver transplantation data (large sample size, many predictors), it was shown that ML techniques can be a useful tool for both prediction (discrimination / calibration) and interpretation in the survival context. In a simulation study performed for synthetic osteosarcoma data in a simple setting (small / medium sample size, small number of predictors), SNNs reached a comparable predictive performance with Cox models but were generally less well calibrated. In another study with extremity STS data in a simple clinical setting with competing events, ML methods were able to reach a comparable performance with traditional regression models but the latter were frequently better calibrated. Therefore, more attention to model calibration is urgently needed for ML.

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] 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.

[3] N. I. Cherny, U. Dafni, J. Bogaerts, N. J. Latino, G. Pentheroudakis, J. Y. Douillard, J. Tabernero, C. Zielinski, M. J. Piccart, and E. G. E. de Vries. ESMO-Magnitude of Clinical Benefit Scale version 1.1. *Annals of Oncology*, 28(10):2340–2366, 2017. doi: 10.1093/annonc/mdx310.

[4] L. M. Ellis, D. S. Bernstein, E. E. Voest, J. D. Berlin, D. Sargent, P. Cortazar, E. Garrett-Mayer, R. S. Herbst, R. C. Lilenbaum, C. Sima, A. P. Venook, M. Gonen, R. L. Schilsky, N. J. Meropol, and L. E. Schnipper. American society of clinical oncology perspective: Raising the bar for clinical trials by defining clinically meaningful outcomes. *Journal of Clinical Oncology*, 32(12):1277–1280, 4 2014. ISSN 15277755. doi: 10.1200/JCO.2013.53.8009.

[5] P. G. Casali, N. Abecassis, H. T. Aro, S. Bauer, R. Biagini, S. Bielack, S. Bonvalot, I. Boukovinas, J. V. M. G. Bovee, T. Brodowicz, J. M. Broto, A. Buonadonna, E. De Álava, A. P. Dei Tos, X. G. Del Muro, P. Dileo, M. Eriksson, A. Fedenko, V. Ferraresi, A. Ferrari, S. Ferrari, A. M. Frezza, S. Gasperoni, H. Gelderblom, T. Gil, G. Grignani, A. Gronchi, R. L. Haas, B. Hassan, P. Hohenberger, R. Issels, H. Joensuu, R. L. Jones, I. Judson, P. Jutte, S. Kaal, B. Kasper, K. Kopeckova, D. A. Krákorová, A. Le Cesne, I. Lugowska, O. Merimsky, M. Montemurro, M. A. Pantaleo, R. Piana, P. Picci, S. Piperno-Neumann, A. L. Pousa, P. Reichardt, M. H. Robinson, P. Rutkowski, A. A. Safwat, P. Schöffski, S. Sleijfer, S. Stacchiotti, K. Sundby Hall, M. Unk, F. Van Coevorden, W. T. A. Van Der Graaf, J. Whelan, E. Wardelmann, O. Zaikova, and J. Y. Blay. Soft tissue and visceral sarcomas: ESMO-EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 29(Supplement_4):iv51–iv67, 2018. ISSN 15698041. doi: 10.1093/annonc/mdy321.

[6] A. Gronchi, A. B. Miah, A. P. Dei Tos, N. Abecassis, J. Bajpai, S. Bauer, R. Biagini, S. Bielack, J. Y. Blay, S. Bolle, S. Bonvalot, I. Boukovinas, J. V. M. G. Bovee, K. Boye, B. Brennan, T. Brodowicz, A. Buonadonna, E. De Álava, X. G. Del Muro, A. Dufresne, M. Eriksson, F. Fagioli, A. Fedenko, V. Ferraresi, A. Ferrari, A. M. Frezza, S. Gasperoni, H. Gelderblom, F. Gouin, G. Grignani, R. Haas, A. B. Hassan, S. Hecker-Nolting, N. Hindi, P. Hohenberger, H. Joensuu, R. L. Jones, C. Jungels, P. Jutte, L. Kager, B. Kasper, A. Kawai, K. Kopeckova, D. A. Krákorová, A. Le Cesne, F. Le Grange, E. Legius, A. Leithner, A. Lopez-Pousa, J. Martin-Broto, O. Merimsky, C. Messiou, O. Mir, M. Montemurro, B. Morland, C. Morosi, E. Palmerini, M. A. Pantaleo, R. Piana, S. Piperno-Neumann, P. Reichardt, P. Rutkowski, A. A. Safwat, C. Sangalli, M. Sbaraglia, S. Scheipl, P. Schöffski, S. Sleijfer, D. Strauss, S. Strauss, K. Sundby Hall, A. Trama, M. Unk, M. A. J. van de Sande, W. T. A. van der Graaf, W. J. van Houdt, T. Frebourg, P. G. Casali, and S. Stacchiotti. Soft tissue and visceral sarcomas: ESMO–EURACAN–GENTURIS Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 32(11):1348–1365, 2021. ISSN 15698041. doi: 10.1016/j.annonc.2021.07.006. URL https://doi.org/10.1016/j.annonc.2021.07.006.

[7] 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.

[8] 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.

[9] 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.

[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] E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer*, 45(2):228–247, 2009. ISSN 09598049. doi: 10.1016/j.ejca.2008.10.026. URL http://dx.doi.org/10.1016/j.ejca.2008.10.026.

[12] L. Seymour, S. P. Ivy, D. Sargent, D. Spriggs, L. Baker, M. J. Ratain, M. L. Blanc, D. Stewart, J. Crowley, J. S. Humphrey, P. West, and D. Berry. The Design of Phase II Clinical Trials Testing Cancer Therapeutics: Consensus Recommendations from the Clinical Trial Design Task Force of the National Cancer Institute Investigational Drug Steering Committee. *Clinical Cancer Research*, 16(6):1764–1769, 2010. doi: 10.1158/1078-0432.CCR-09-3287.The.

[13] M. Savina, S. Litière, A. Italiano, T. Burzykowski, F. Bonnetain, S. Gourgou, V. Rondeau, J. Y. Blay, S. Cousin, F. Duffaud, H. Gelderblom, A. Gronchi, I. Judson, A. Le Cesne, P. Lorigan, J. Maurel, W. van der Graaf, J. Verweij, S. Mathoulin-Pélissier, and C. Bellera. Surrogate endpoints in advanced sarcoma trials: A meta-analysis. *Oncotarget*, 9(77):34617–34627, 2018. ISSN 19492553. doi: 10.18632/oncotarget.26166.

[14] K. Tanaka, M. Kawano, T. Iwasaki, I. Itonaga, and H. Tsumura. Surrogacy of intermediate endpoints for overall survival in randomized controlled trials of first-line treatment for advanced soft tissue sarcoma in the pre- and post-pazopanib era: A meta-analytic evaluation. *BMC Cancer*, 19(1):1–9, 2019. ISSN 14712407. doi: 10.1186/s12885-019-5268-2.

[15] W. D. Tap, R. L. Jones, B. A. Van Tine, B. Chmielowski, A. D. Elias, D. Adkins, M. Agulnik, M. M. Cooney, M. B. Livingston, G. Pennock, M. R. Hameed, G. D. Shah, A. Qin, A. Shahir, D. M. Cronier, R. Ilaria, I. Conti, J. Cosaert, and G. K. Schwartz. Olaratumab and doxorubicin versus doxorubicin alone for treatment of soft-tissue sarcoma: an open-label phase 1b and randomised phase 2 trial. *The Lancet*, 388(10043):488–497, 7 2016. ISSN 1474-547X. doi: 10.1016/S0140-6736(16)30587-6. URL https://pubmed.ncbi.nlm.nih.gov/27291997/.

[16] W. D. Tap, A. J. Wagner, P. Schöffski, J. Martin-Broto, A. Krarup-Hansen, K. N. Ganjoo, C. C. Yen, A. R. Abdul Razak, A. Spira, A. Kawai, A. Le Cesne, B. A. Van Tine, Y. Naito, S. H. Park, A. Fedenko, Z. Pápai, V. Soldatenkova, A. Shahir, G. Mo, J. Wright, and R. L. Jones. Effect of Doxorubicin Plus Olaratumab vs Doxorubicin Plus Placebo on Survival in Patients With Advanced Soft Tissue Sarcomas: The ANNOUNCE Randomized Clinical Trial. *JAMA*, 323(13):1266–1276, 4 2020. ISSN 1538-3598. doi: 10.1001/JAMA.2020.1707. URL https://pubmed.ncbi.nlm.nih.gov/32259228/.

[17] P. G. Casali, P. Bruzzi, J. Bogaerts, J. Y. Blay, M. Aapro, A. Adamous, A. Berruti, J. Bressington, B. Bruzzi, R. Capocaccia, F. Cardoso, J. E. Celis, A. Cervantes, F. Ciardiello, C. Claussen, M. Coleman, S. Comis, S. Craine, D. De Boltz, F. De Lorenzo, A. P. Dei Tos, G. Gatta, J. Geissler, R. Giuliani, E. Grande, A. Gronchi, S. Jezdic, B. Jonsson, L. Jost, H. Keulen, D. Lacombe, G. Lamory, Y. Le Cam, S. Leto di Priolo, L. Licitra, F. Macchia, A. Margulies, S. Marreaud, G. McVie, S. Narbutas, K. Oliver, N. Pavlidis, J. Pelouchova, G. Pentheroudakis, M. Piccart, M. A. Pierotti, G. Pravettoni, K. Redmond, P. Riegman, M. P. Ruffilli, D. Ryner, S. Sandrucci, M. Seymour, V. Torri, A. Trama, S. Van Belle, G. Vassal, M. Wartenberg, C. Watts, A. Wilson, and W. Yared. Rare Cancers Europe (RCE) methodological recommendations for clinical studies

in rare cancers: A European consensus position paper. *Annals of Oncology*, 26(2):300–306, 2015. ISSN 15698041. doi: 10.1093/annonc/mdu459.

[18] Jan Bogaerts, Matthew R. Sydes, Nicola Keat, Andrea McConnell, Al Benson, Alan Ho, Arnaud Roth, Catherine Fortpied, Cathy Eng, Clare Peckitt, Corneel Coens, Curtis Pettaway, Dirk Arnold, Emma Hall, Ernie Marshall, Francesco Sclafani, Helen Hatcher, Helena Earl, Isabelle Ray-Coquard, James Paul, Jean Yves Blay, Jeremy Whelan, Kathy Panageas, Keith Wheatley, Kevin Harrington, Lisa Licitra, Lucinda Billingham, Martee Hensley, Martin McCabe, Poulam M. Patel, Richard Carvajal, Richard Wilson, Rob Glynne-Jones, Rob McWilliams, Serge Leyvraz, Sheela Rao, Steve Nicholson, Virginia Filiaci, Anastassia Negrouk, Denis Lacombe, Elisabeth Dupont, Iris Pauporté, John J. Welch, Kate Law, Ted Trimble, and Matthew Seymour. Clinical trial designs for rare diseases: Studies developed and discussed by the International Rare Cancers Initiative. *European Journal of Cancer*, 51(3):271–281, 2015. ISSN 18790852. doi: 10.1016/j.ejca.2014.10.027.

[19] 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/.

[20] K. S. Panageas. Clinical trial design for rare cancers - why a less conventional route may be required. *Expert review of clinical pharmacology*, 8(6):661–663, 11 2015. ISSN 17512441. doi: 10.1586/17512433.2015.1088382. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4724195/.

[21] 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.

[22] 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.

[23] 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.

[24] 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/.

[25] S. Sleijfer, I. Ray-Coquard, Z. Papai, A Le Cesne, M. Scurr, P. Schöffski, F. Collin, L. Pandite, S. Marreaud, A. De Brauwer, M. Van Glabbeke, J. Verweij, and J. Y. Blay. Pazopanib, a multikinase angiogenesis inhibitor, in patients with relapsed or refractory advanced soft tissue sarcoma: A phase II study from the European organisation for research and treatment of cancer-soft tissue and bone sarcoma group (EORTC Study 620. *Journal of Clinical Oncology*, 27(19):3126–3132, 2009. ISSN 0732183X. doi: 10.1200/JCO.2008.21.3223.

[26] P. Schöffski, I. L. Ray-Coquard, A. Cioffi, N. B. Bui, S. Bauer, J. T. Hartmann, A. Krarup-Hansen, V. Grünwald, R. Sciot, H. Dumez, J. Y. Blay, A. Le Cesne, J. Wanders, C. Hayward, S. Marreaud, M. Ouali, and P. Hohenberger. Activity of eribulin mesylate in patients with soft-tissue sarcoma: A phase 2 study in four

independent histological subtypes. *The Lancet Oncology*, 12(11):1045–1052, 2011. ISSN 14702045. doi: 10.1016/S1470-2045(11)70230-3.

[27] W. T. A. van der Graaf, J. Y. Blay, S. P. Chawla, D. W. Kim, B. Bui-Nguyen, P. G. Casali, P. Schöffski, M. Aglietta, A. P. Staddon, Y. Beppu, A. Le Cesne, H. Gelderblom, I. R. Judson, N. Araki, M. Ouali, S. Marreaud, R. Hodge, M. R. Dewji, C. Coens, G. D. Demetri, C. D. Fletcher, A. P. Dei Tos, and P. Hohenberger. Pazopanib for metastatic soft-tissue sarcoma (PALETTE): A randomised, double-blind, placebo-controlled phase 3 trial. *The Lancet*, 379(9829):1879–1886, 2012. ISSN 1474547X. doi: 10.1016/S0140-6736(12)60651-5.

[28] I. Judson, J. Verweij, H. Gelderblom, J. T. Hartmann, P. Schöffski, J. Y. Blay, J. M. Kerst, J. Sufliarsky, J. Whelan, P. Hohenberger, A. Krarup-Hansen, T. Alcindor, S. Marreaud, S. Litière, C. Hermans, C. Fisher, P. C. W. Hogendoorn, A. P. Dei Tos, and W. T. A. van der Graaf. Doxorubicin alone versus intensified doxorubicin plus ifosfamide for first-line treatment of advanced or metastatic soft-tissue sarcoma: A randomised controlled phase 3 trial. *The Lancet Oncology*, 15(4):415–423, 2014. ISSN 14745488. doi: 10.1016/S1470-2045(14)70063-4.

[29] B. Bui-Nguyen, J. E. Butrynski, N. Penel, J. Y. Blay, N. Isambert, M. Milhem, J. M. Kerst, A. K. L. Reyners, S. Litière, S. Marréaud, F. Collin, and W. T. A. van der Graaf. A phase IIb multicentre study comparing the efficacy of trabectedin to doxorubicin in patients with advanced or metastatic untreated soft tissue sarcoma: The TRUSTS trial. *European Journal of Cancer*, 51(10):1312–1320, 2015. ISSN 18790852. doi: 10.1016/j.ejca.2015.03.023.

[30] B. Yücel, M. G. Celasun, B. Öztoprak, Z. Hasbek, S. Bahar, T. Kaçan, A. Bahçeci, and M. M. Şeker. The negative prognostic impact of bone metastasis with a tumor mass. *Clinics*, 70(8):535–540, 2015. ISSN 18075932. doi: 10.6061/clinics/2015(08)01.

[31] X. H. Du, H. Wei, P. Zhang, W. T. Yao, and Q. Q. Cai. Heterogeneity of Soft Tissue Sarcomas and Its Implications in Targeted Therapy. *Frontiers in Oncology*, 10:1–3, 2020. ISSN 2234943X. doi: 10.3389/fonc.2020.564852. URL https://www.frontiersin.org/article/10.3389/fonc.2020.564852.

[32] D. B. Fogel. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. *Contemporary Clinical Trials Communications*, 11(August):156–164, 2018. ISSN 24518654. doi: 10.1016/j.conctc.2018.08.001. URL https://doi.org/10.1016/j.conctc.2018.08.001.

[33] P. Pautier, A. Floquet, C. Chevreau, N. Penel, C. Guillemet, C. Delcambre, D. Cupissol, F. Selle, N. Isambert, S. Piperno-Neumann, A. Thyss, F. Bertucci, E. Bompas, J. Alexandre, O. Collard, S. Lavau-Denes, P. Soulié, M. Toulmonde, A. Le Cesne, B. Lacas, and F. Duffaud. Trabectedin in combination with doxorubicin for first-line treatment of advanced uterine or soft-tissue leiomyosarcoma (LMS-02): A non-randomised, multicentre, phase 2 trial. *The Lancet Oncology*, 16(4):457–464, 2015. ISSN 14745488. doi: 10.1016/S1470-2045(15)70070-7.

[34] P. Pautier, A. Italiano, S. Piperno-Neumann, C.M. Chevreau, N. Penel, D. Cupissol, P. Boudou Rouquette, F. Bertucci, C. Balleyguier, V. Lebrun-Ly, J. Blay, E. Kalbacher, C. Delcambre, E. Bompas, O. Collard, N. Isambert, C. Guillemet, M. Rios, M. Sundqvist, and F. Duffaud. LBA59 - LMS-04 study: A randomised, multicenter, phase III study comparing doxorubicin alone versus doxorubicin with trabectedin followed by trabectedin in non-progressive patients as first-line therapy, in patients with metastatic or unresectable leiomyo. *Annals of Oncology*, 32(suppl_5):S1283–S1346, 2021. URL https://oncologypro.esmo.org/meeting-resources/esmo-congress/lms-04-study-a-randomised-multicenter-phase-iii-study-comparing-doxorubicin-alone-versus-doxorubicin-with-trabectedin-followed-by-trabectedin-in.

[35] EORTC 1809 STRASS II Trial In Retroperitoneal Sarcoma Receives Support By The Anticancer Fund 2022 - EORTC : EORTC. URL https://www.eortc.org/blog/2020/04/28/eortc-1809-strass-ii-trial-in-retroperitoneal-sarcoma-receives-support-by-the-anticancer-fund/.

[36] 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.

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

[38] Ira S. Hofer, Michael Burns, Samir Kendale, and Jonathan P. Wanderer. Realistically integrating machine learning into clinical practice: A road map of opportunities, challenges, and a potential future. *Anesthesia and Analgesia*, 130(5):1115–1118, 2020. ISSN 15267598. doi: 10.1213/ANE.0000000000004575.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] E. W. Steyerberg. *Clinical prediction models: A Practical Approach to Development, Validation, and Updating*. Springer, 2nd edition, 2019. doi: https://doi.org/10.1007/978-3-030-16399-0. URL https://www.springer.com/gp/book/9783030163983.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] R. Tibshirani. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395, 1997. ISSN 02776715. doi: 10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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/.

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

[61] 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.

[62] C. Wallisch, A. Agibetov, G. Dorffner, D. Dunkler, and G. Heinze. *Statistical modelling or machine learning: Interpretability vs. flexibility?* 40th Annual Conference of the International Society for Clinical Biostatistics, 2019. URL https://kuleuvencongres.be/iscb40/images/iscb40-2019-e-versie.pdf.

[63] 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.

[64] F. E. Harrell Jr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer, 2nd edition, 2015. ISBN 978-3-319-19425-7. doi: https://doi.org/10.1007/978-3-319-19425-7. URL http://www.springer.com/series/692.

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

[66] N. H. Shah, A. Milstein, and Steven C. Bagley. Making Machine Learning Models Clinically Useful. *JAMA*, 322(14):1351–1352, 10 2019. ISSN 1538-3598. doi: 10.1001/JAMA.2019.10306. URL https://pubmed.ncbi.nlm.nih.gov/31393527/.

[67] 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.

[68] A. L. Beam and I. S. Kohane. Big data and machine learning in health care. *JAMA - Journal of the American Medical Association*, 319(13):1317–1318, 2018. ISSN 15383598. doi: 10.1001/jama.2017.18391.

[69] M. Nagy, N. Radakovich, and A. Nazha. Machine Learning in Oncology: What Should Clinicians Know? *JCO Clinical Cancer Informatics*, (4):799–810, 2020. ISSN 24734276. doi: 10.1200/cci.20.00049. URL https://doi.org/10.1200/CCI.20.00049.

[70] E. Christodoulou, J. Ma, G. S Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110:12–22, 2019. ISSN 18785921. doi: 10.1016/j.jclinepi.2019.02.004.

[71] I. Bou-Hamad, D. Larocque, and H. Ben-Ameur. A review of survival trees. *Statistics Surveys*, 5:44–71, 1 2011. ISSN 1935-7516. doi: 10.1214/09-SS047. URL https://projecteuclid.org/journals/statistics-surveys/volume-5/issue-none/A-review-of-survival-trees/10.1214/09-SS047.full.

[72] K. Liestol, P. K. Andersen, and U. Andersen. Survival analysis and neural nets. *Statistics in Medicine*, 13 (12):1189–1200, 1994. doi: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780131202.

[73] 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.

[74] F. M. Khan and V. B. Zubek. Support Vector Regression for Censored Data (SVRc): A Novel Tool for Survival Analysis. *2008 Eighth IEEE International Conference on Data Mining*, pages 863–868, 2008. doi: 10.1109/ICDM.2008.50.

[75] T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006. ISSN 14654644. doi: 10.1093/biostatistics/kxj011.

[76] 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.

[77] R. F. Wolff, K. G. M. Moons, R. D. Riley, P. F. Whiting, M. Westwood, G. S. Collins, J. B. Reitsma, J. Kleijnen, and S. Mallett. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, 170(1):51–58, 2019. ISSN 15393704. doi: 10.7326/M18-1376.

[78] 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.

Summary

This thesis sprang 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. Research was split into two parts. In **Part** I (**Chapters** 2, 3, 4), statistical analyses were performed for the EORTC - Soft Tissue and Bone Sarcoma Group (STBSG), whereas in **Part** II (**Chapters** 5, 6, 7, 8) the potential of survival prediction models with machine learning (ML) techniques was compared with traditional statistical models (SM) for sarcoma and non-sarcoma clinical data.

Part I: Clinical trials in soft-tissue sarcomas

This part provided modern efficacy thresholds to design new phase II clinical trials for common histotypes of locally advanced or metastatic soft-tissue sarcoma (STS) patients. The prognostic significance of bone metastasis in STS was investigated to identify high-risk patient populations.

In 2002, Van Glabbeke *et al.* published on behalf of the EORTC - STBSG a pooled analysis to estimate progression-free rates at 3 and 6 months for various groups of STS patients who participated in phase II trials of the EORTC. These historical values have been widely used (> 420 citations) to design new studies for all STS or for specific histology subgroups (in first-line treatment). 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 the papers of our literature review. In **Chapter** 2, a random-effects meta-analysis was performed to provide new benchmarks for designing phase II studies of advanced or metastatic LMS patients separately for first-line or pre-treated population. The primary endpoints of interest were progression-free survival rates (PFSRs) at 3 and 6 months, which are nowadays preferred and more frequently reported than progression-free rates (censoring non-disease-related death). When estimates could not be derived from publications, first authors and/or sponsors were contacted. The ESMO Magnitude of Clinical Benefit Scale (MCBS) was used to guide the treatment effect to target in future trials. Information was obtained on 7 first-line and 16 pre-treated trials for 1500 LMS patients. Under the alternative that the true benefit amounts to a hazard ratio of 0.65, a 6-month PFSR $\geq$ 70% can be considered to suggest drug activity in first-line. For pre-treated population, a 3-month PFSR $\geq$ 62% or 6-month PFSR $\geq$ 44% would suggest drug activity. Specific results were also provided for uterine LMS.

In **Chapter** 3, a second meta-analysis was performed for advanced or metastatic liposarcoma (LPS) or synovial sarcoma (SS) - the second and third most common histotypes in our literature review. Study endpoints were PFSRs at 3 and 6 months. The choice of the therapeutic benefit to target in future trials was guided again by the ESMO MCBS. Information was acquired for 1030 LPS patients (25 trials; 7 first-line, 17 pre-treated, 1 both) and 348 SS patients (13 trials; 3 first-line, 10 pre-treated). New benchmarks were proposed to design future histology-specific phase II trials. Minimum values to target in first-line at 3 and 6 months were 79% and 69% for LPS, 82% and 69% for SS. For pre-treated patients, recommended PFSRs at 3 and 6 months suggesting drug activity were 63%

and 44% for LPS, 60% and 41% for SS. Our findings here and in the previous chapter indicate that there is a need to raise the bar of thresholds for the commonest STS types in future histology-tailored phase II trials in order to achieve higher success rates in new prospective confirmatory phase III trials.

In **Chapter** 4, we investigated whether, and if so, to what extent, skeletal metastases at presentation affect the outcome of patients with advanced or metastatic STS. Selected patients participated in five clinical trials of EORTC - STBSG. Individuals were included if they started treatment with an active drug and had advanced/metastatic STS. The endpoints of interest were overall survival (OS) and progression-free survival (PFS). Univariate and multivariate pooled analyses (after correcting for 12 covariates) were employed with Kaplan–Meier and Cox regression to model the impact of bone metastasis at presentation per treatment line stratified by study. For the subset of patients with bone metastasis, the impact of another metastatic organ site (among liver, lymph node, lung, soft-tissue, or other) at diagnosis was explored with multivariate Cox regression models. 565 out of 1034 (54.6%) patients received first-line systemic treatment for locally advanced or metastatic disease. Bone metastases were present in 140 patients (77 first-line, 63 second-line or higher). The unadjusted difference in OS/PFS with or without bone metastasis was statistically significant only for first-line patients. For OS, the adjusted hazard ratios for bone metastasis presence were 1.33 (95%-CI: 0.99–1.78) and 1.11 (95%-CI: 0.81–1.52) for first-line/second-line or higher treated patients, respectively. Likewise, the adjusted hazard ratios for PFS were 1.31 (95%-CI: 1.00–1.73) and 1.07 (95%-CI: 0.80–1.43). Hence, the adjusted effects were not statistically significant, despite a trend for first-line patients. Subgroup analyses indicated bone and lymph node metastasis as the most detrimental combination for OS and bone and lung metastasis for PFS. Since skeletal metastases at study entry cannot be ascertained as a significant risk factor (per line of treatment), stratification is not justified in randomised studies with these patients.

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

In this part of the thesis, the predictive performance of existing and novel ML methods was compared with traditional SM for real-life clinical data (small/medium or large sample sizes, low- or high-dimensional settings) with time-to-event endpoints.

Nowadays, a growing interest can be observed in applying ML for clinical prediction by the medical community. Over the years, several algorithms have been developed and adapted to right censored data. Neural networks have been repeatedly employed to build clinical prediction models in healthcare. Even so, despite their non-negligible use, a comprehensive review on survival neural networks (SNNs) using prognostic factors is missing. In **Chapter** 5, we presented the first ever attempt at a structured overview of SNNs with prognostic factors for clinical prediction. Our aim was to provide a broad understanding of the literature (1st January 1990 - 31st August 2021, global search in PubMed). Relevant manuscripts were classified as methodological/technical (novel methodology or new theoretical model; 13 studies) or applications (11 studies). We discussed how SNNs are employed in the medical field for prediction and detailed how researchers have tried to adapt a classification method to right censored survival data. There are two methodological trends: either time is added as part of the input features and a single output node is specified, or multiple output nodes are defined for each time interval. This work was supplemented with a critical appraisal of model aspects that should be designed and reported more carefully. We identified key characteristics of prediction models (i.e., number of patients/predictors, evaluation measures, calibration), and compared SNNs' predictive performance to the Cox proportional hazards model. The median sample size was 920 patients, and the median number of predictors was 7. Major findings included poor reporting (e.g., regarding missing data, hyperparameters), as well as inaccurate model development/validation. Calibration was neglected in more than half of the studies. Cox models were not developed to their full potential, and claims for the performance of SNNs were exaggerated. Light was shed on the current state of art of SNNs in medicine with prognostic factors. Limitations were discussed, and future directions were proposed for researchers who seek to develop existing methodology.

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 (large sample size, high-dimensional setting). In **Chapter** 6, three ML techniques: a) random survival forests (RSF), and b-c) two methodological extensions of the partial logistic artificial neural network (PLANN) with one and two hidden layers were 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 were selected, over more than 600, to predict survival since liver transplantation on clinical/statistical grounds. A comparison was performed between these ML techniques and three different Cox models (all variables, backward, LASSO). Emphasis was given on the advantages and pitfalls of each method and on extracting interpretability from the ML methods. Well-established predictive measures were employed from the survival field (C-index, Brier score and Integrated Brier Score) and the strongest prognostic factors were identified for each model. Clinical endpoint was overall graft-survival defined as the time between transplantation and the date of graft-failure or death. The RSF showed slightly better predictive performance than Cox models based on the C-index. Neural networks showed better performance than both Cox models and RSF based on the Integrated Brier Score at 10 years. From the three ML techniques, PLANN extended with one hidden layer predicted survival probabilities the most accurately being as calibrated as the Cox model with all variables. The RSF and the PLANN extended with two hidden layers were less calibrated on test data. Regarding interpretability, the Cox model with all variables and the PLANNs identified *re-transplantation* as the strongest predictor and *donor age*, *diabetes*, and *life support* as relatively strong predictors. According to RSF, the most prognostic variable was *donor age*, followed by *re-transplantation*, *life support* and *serology status of Chronic hepatitis C virus*. All in all, it was shown that ML techniques can be a useful tool for both prediction and interpretation in this survival context.

In the previous study, our group provided new methodological extensions of the PLANN model. PLANN extended was developed and validated for complex liver transplantation data. However, it is not uncommon 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. Therefore, in **Chapter** 7, the focus was on the comparison between such models for non-complex clinical data (small / medium sample size, low dimensional) with a Monte Carlo simulation study to investigate a different real-life setting. Synthetic data (250 or 1000 patients) were generated that closely resembled five prognostic factors preselected based on a European Osteosarcoma Intergroup study (MRC BO06/EORTC 80931) that investigated the effect of dose-intense chemotherapy in patients with localised extremity osteosarcoma. The predictive performance of PLANN original and PLANN extended (with one hidden layer) was compared with Cox models for 20, 40, 61, and 80% censoring. Survival times were generated from a log-normal distribution. The endpoint of interest was overall survival defined as the time to death from any cause since the date of surgery. Models were evaluated in terms of the C-index, Brier score at 0-5 years, integrated Brier score (IBS) at 5 years, and miscalibration at 2 and 5 years (usually neglected). The ML models were able to reach a similar predictive performance on simulated data for most scenarios with respect to the C-index, Brier score, or IBS. However, the SM were frequently better calibrated. Performance was robust in scenarios where censored patients were removed before the 2^{nd} year or administrative censoring at 5 years was performed (on training data). Researchers should be aware of burdensome aspects of ML techniques such as data preprocessing, tuning of hyperparameters, and computational intensity that render them disadvantageous against conventional regression models in a simple clinical setting.

In health research, several chronic diseases are susceptible to competing risks (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, we aimed to develop and validate prognostic clinical prediction models for CRs with SM and ML techniques. Two SM a) cause-specific Cox, b) Fine-Gray model and three ML models i) PLANN original for CRs (PLANNCR original), ii) a methodological extension called PLANNCR extended, and iii) RSF for CRs (RSFCR) were employed. The predictive performance of all methods was assessed in terms of discrimination and calibration in another simple clinical setting (small / medium sample size, small number of predictors). The dataset at hand contained 3826 retrospectively collected patients with extremity STS (eSTS) and nine predictors from the PERsonalised SARcoma Care (PERSARC) Study Group. To the best of our knowledge, this was the first ever study of this kind for eSTS. The clinical endpoint was the time in years between surgery and disease progression (event of interest) or death (competing event). The Brier score, the area under the curve (AUC) and the model's miscalibration were used to evaluate predictive performance at 2, 5, and 10 years, respectively. Results showed that the ML models are able to reach a comparable performance with the SM based on the Brier score and AUC for disease progression and death (95% confidence intervals at 2, 5, and 10 years overlapped). Nevertheless, the SM were frequently better calibrated. Overall, ML techniques are less practical as they require substantial implementation time (data preprocessing, hyperparameter tuning, computational intensity). As such, for non-complex real life 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.

Nederlandse samenvatting

Dit proefschrift is voortgekomen uit een interdisciplinaire samenwerking tussen de European Organization for Research and Treatment of Cancer (EORTC), het Mathematisch Instituut van de Universiteit Leiden en de afdeling Medische Oncologie van het Leids Universitair Medisch Centrum (LUMC). Het onderzoek werd opgesplitst in twee delen. **Deel** I (**Hoofdstukken** 2, 3, 4) beschrijven statistische analyses uitgevoerd voor de EORTC - Soft Tissue and Bone Sarcoma Group (STBSG), in **Deel** II (**Hoofdstukken** 5, 6, 7, 8) werd het potentieel van overlevingsvoorspellingsmodellen met machine learning (ML) technieken vergeleken met traditionele statistische modellen (SM) voor sarcoom en niet-sarcoom klinische data.

Deel I: Klinische proeven bij wekedelensarcomen

Dit deel verschafte moderne drempelwaarden om nieuwe klinische fase II studies te ontwerpen voor de werkzaamheid van nieuwe behandelingen voor veelvoorkomende histotypes van lokaal gevorderde of gemetastaseerde wekedelensarcoom (STS)-patiënten. De prognostische betekenis van botmetastasen bij STS werd onderzocht om patiëntenpopulaties met een hoog risico te identificeren.

In 2002 publiceerde Van Glabbeke *et al.* namens de EORTC - STBSG een gepoolde analyse om de progressievrije percentages patiënten op 3 en 6 maanden te schatten voor verschillende groepen STS-patiënten die deelnamen aan fase II-onderzoeken van de EORTC. Deze historische waarden zijn op grote schaal gebruikt (> 420 citaties) om nieuwe studies te ontwerpen voor alle STS of voor specifieke histologische subgroepen (in eerstelijnsbehandeling). We hebben een uitgebreid literatuuronderzoek uitgevoerd om alle fase II of daaropvolgende klinische onderzoeken van geavanceerde of gemetastaseerde STS (2003 tot 2018) te identificeren, en zo het huidige landschap te beschrijven. Vanwege de aanzienlijke heterogeniteit tussen klinische onderzoeken werd besloten om eerst te focussen op leiomyosarcoom (LMS) - het meest voorkomende STS-subtype in de artikelen van ons literatuuronderzoek. In **Hoofdstuk** 2 werd een meta-analyse met een random-effecten model uitgevoerd om nieuwe drempelwaarden te bepalen voor het opzetten van fase II-onderzoeken van patiënten met gevorderd of gemetastaseerd LMS, afzonderlijk voor eerstelijns- of voorbehandelde populatie. De primaire eindpunten van belang waren progressievrije overlevingspercentages (PFSR's) na 3 en 6 maanden, die tegenwoordig de voorkeur hebben en vaker worden gerapporteerd dan progressievrije percentages (waarbij niet-ziektegerelateerde sterfte gecensureerd wordt). Wanneer schattingen niet konden worden afgeleid uit publicaties, is contact opgenomen met eerste auteurs en/of sponsors. De ESMO Magnitude of Clinical Benefit Scale (MCBS) werd gebruikt om een indicatie te krijgen van het behandelingseffect dat voorzien moet worden in toekomstige onderzoeken. Er werd informatie verkregen over 7 eerstelijns en 16 voorbehandelde onderzoeken bij 1500 LMS-patiënten. Onder het alternatief dat het werkelijke voordeel een hazard ratio van 0.65 bedraagt, kan een PFSR op 6 maanden van $\geq$ 70% worden overwogen om te zoeken naar een actieve behandeling in de eerste lijn. Voor een voorbehandelde populatie zou een PFSR van 3 maanden $\geq$ 62% of een PFSR van 6 maanden $\geq$ 44% wijzen op een werkzame behandeling. Specifieke resultaten werden ook verstrekt voor LMS van de baarmoeder.

In **Hoofdstuk** 3 werd een tweede meta-analyse uitgevoerd voor gevorderd of gemetastaseerd liposarcoom (LPS) of synoviosarcoom (SS) - het tweede en derde meest voorkomende histotype in ons literatuuronderzoek. Onderzoekseindpunten waren PFSR's na 3 en 6 maanden. De keuze van het therapeutische voordeel dat in toekomstige onderzoeken zou moeten worden nagestreefd, werd opnieuw geleid door de ESMO MCBS. Informatie werd verkregen voor 1030 LPS-patiënten (25 onderzoeken; 7 eerstelijns, 17 voorbehandeld, 1 beide) en 348 SS-patiënten (13 onderzoeken; 3 eerstelijns, 10 voorbehandeld). Er werden opnieuw drempelwaarden voorgesteld voor toekomstige histologie-specifieke fase II-studies. Streefwaarden in de eerste lijn na 3 en 6 maanden waren 79% en 69% voor LPS, 82% en 69% voor SS. Voor voorbehandelde patiënten waren de streefwaarden voor PFSR's na 3 en 6 maanden 63% en 44% voor LPS, 60% en 41% voor SS. Onze bevindingen hier en in het vorige hoofdstuk geven aan dat er behoefte is aan een hogere drempel voor de meest voorkomende STS-types in toekomstige, op histologie afgestemde fase II-studies om hogere succespercentages te bereiken in nieuwe prospectieve bevestigende fase III-studies.

In **Hoofdstuk** 4 hebben we onderzocht of, en zo ja, in welke mate botmetastasen bij start van de behandeling de prognose van patiënten met gevorderde of gemetastaseerde STS beïnvloeden. Geselecteerde patiënten namen deel aan vijf klinische studies van EORTC - STBSG. Individuen werden geïncludeerd als ze begonnen met de behandeling met een actief medicijn en gevorderd/gemetastaseerd STS hadden. De eindpunten voor dit onderzoek waren algehele overleving (OS) en progressievrije overleving (PFS). Univariate en multivariate gepoolde analyses (na correctie voor 12 covariaten) werden gebruikt met Kaplan-Meier en Cox-regressie om de impact van botmetastase bij presentatie per behandelingslijn (eerstelijns of later) te modelleren, gestratificeerd per studie. Voor de groep van patiënten met botmetastase werd de impact van de aanwezigheid van andere uitzaaiingen (onder andere in de lever, lymfeklieren, long, zacht weefsel of andere) op het moment van de diagnose onderzocht aan de hand van multivariate Cox-regressiemodellen. 565 van de 1034 (54.6%) patiënten kregen eerstelijns systemische behandeling voor lokaal gevorderde of gemetastaseerde ziekte. Botmetastasen waren aanwezig bij 140 patiënten (77 eerstelijns, 63 tweedelijns of later). Het niet-gecorrigeerde verschil in OS/PFS met of zonder botmetastase was alleen statistisch significant voor eerstelijnspatiënten. Voor OS waren de aangepaste hazard ratios voor de aanwezigheid van botmetastasen 1.33 (95%-BI: 0.99-1.78) en 1.11 (95%-BI: 0.81-1.52) voor eerstelijns/tweedelijns of later behandelde patiënten, respectievelijk. De gecorrigeerde hazard ratios voor PFS waren 1.31 (95%-BI: 1.00–1.73) en 1.07 (95%-BI: 0.80–1.43). De gecorrigeerde effecten waren dus niet statistisch significant, ondanks een trend voor eerstelijnspatiënten. Subgroepanalyses wezen op bot- en lymfekliermetastase als de meest schadelijke combinatie voor OS en bot- en longmetastase voor PFS. Aangezien skeletmetastasen bij aanvang van het onderzoek niet als significante risicofactor (per behandelingslijn) kunnen worden geïdentificeerd, is stratificatie in gerandomiseerde onderzoeken met deze patiënten niet aangewezen.

Deel II: Statistische modellen versus machine learning om overleving te voorspellen aan de hand van sarcoom en niet-sarcoom klinische gegevens

In dit deel van het proefschrift werden de voorspellende prestaties van bestaande en nieuwe ML-methoden vergeleken met traditionele SM voor de analyse van real-world time-to-event data (van kleine/middelgrote of grote steekproefomvang, met laag- of hoogdimensionale gegevens).

Tegenwoordig is er een groeiende interesse van de medische gemeenschap in toepassingen van ML voor klinische voorspelling. In de loop der jaren zijn er verschillende algoritmen ontwikkeld en aangepast aan rechtsgecensureerde data. Neurale netwerken zijn herhaaldelijk gebruikt om klinische voorspellingsmodellen in de gezondheidszorg te bouwen. Ondanks hun niet te verwaarlozen gebruik, ontbreekt er een uitgebreide beoordeling van overlevingsneurale netwerken (SNN's) op basis van prognostische factoren. In **Hoofdstuk** 5 presenteerden we de allereerste poging tot een gestructureerd overzicht van SNN's met prognostische factoren voor klinische voorspelling. Ons doel was om een breed overzicht van de literatuur te bieden (1 januari 1990 - 31 augustus 2021, global search in PubMed). Relevante manuscripten werden geclassificeerd als methodologisch/technisch (nieuwe methodologie of nieuw theoretisch model; 13 studies) of toepassingen (11 studies). We bespraken hoe SNN's in de medische wereld worden gebruikt voor voorspellingen en beschreven hoe onderzoekers hebben geprobeerd een classificatiemethode aan te passen aan rechtsgecensureerde overlevingsgegevens. Er zijn twee methodologische trends: ofwel wordt tijd toegevoegd als onderdeel van de invoerfuncties en wordt een enkel uitvoerknooppunt gespecificeerd, of worden meerdere uitvoerknooppunten gedefinieerd voor elk tijdsinterval. Dit werk werd aangevuld met een kritische beoordeling van modelaspecten die zorgvuldiger zouden moeten worden ontworpen en gerapporteerd. We identificeerden de belangrijkste kenmerken van voorspellingsmodellen (d.w.z. aantal patiënten/voorspellers, evaluatiemaatregelen, kalibratie) en vergeleken de voorspellende prestaties van SNN's met het Cox-model voor 'proportional hazards'. De mediaan van de steekproefomvangen was 920 patiënten en de mediaan van het aantal voorspellende factoren was 7. De belangrijkste bevindingen waren onder meer slechte rapportering (bijvoorbeeld met betrekking tot ontbrekende gegevens, hyperparameters), evenals onnauwkeurige modelontwikkeling/-validatie. Kalibratie werd in meer dan de helft van de onderzoeken verwaarloosd. Cox-modellen werden niet tot hun volle potentieel ontwikkeld en claims voor de prestaties van SNN's waren overdreven. Er werd licht geworpen op de huidige stand van de techniek van SNN's in de geneeskunde met prognostische factoren. Beperkingen werden besproken en toekomstige richtingen werden voorgesteld voor onderzoekers die bestaande methodologie verder willen ontwikkelen.

Er is een open discussie over de waarde van ML versus SM binnen de klinische en zorgpraktijk. ML-technieken kunnen een aantrekkelijke keuze zijn voor het modelleren van complexe gegevens (grote steekproefomvang, hoogdimensionale setting). In **Hoofdstuk** 6 werden drie ML-technieken: a) random survival forests (RSF), en b-c) twee methodologische uitbreidingen van het partiële logistieke kunstmatige neurale netwerk (PLANN) met één en twee verborgen lagen getest op een grote retrospectieve gegevensset van 62294 patiënten uit de Verenigde Staten, verstrekt door de Scientific Registry for Transplant Recipients. In totaal werden 97 variabelen geselecteerd, uit een total van meer dan 600, om overleving sinds levertransplantatie te voorspellen op klinische/statistische gronden. Er is een vergelijking gemaakt tussen deze ML-technieken en drie verschillende Cox-modellen (volledig model met alle variabelen, achterwaartse selectie, LASSO). De nadruk werd gelegd op de voordelen en valkuilen van elke methode en op de interpreteerbaarheid van de ML-methoden. Er werden goed gefundeerde parameters gebruikt (C-index, Brier-score en Integrated Brier-score) en de sterkste prognostische factoren werden voor elk model geïdentificeerd. Het klinische eindpunt was de transplantaat-overleving, gedefinieerd als de tijd tussen de transplantatie en de datum van het falen van het transplantaat of overlijden van de patiënt. De RSF vertoonde iets betere voorspellende prestaties dan Cox-modellen op basis van de C-index. Neurale netwerken vertoonden betere prestaties dan zowel Cox-modellen als RSF op basis van de Integrated Brier Score na 10 jaar. Van de drie ML-technieken waren de voorspelde overlevingskansen van de PLANN met één verborgen laag het nauwkeurigst, en net zo goed gekalibreerd als het Cox-model met alle variabelen. De RSF en de PLANN uitgebreid met twee verborgen lagen waren minder goed gekalibreerd op de testgegevens. Wat betreft de interpreteerbaarheid, identificeerden het Cox-model met alle variabelen en de PLANNs *hertransplantatie* als de sterkste voorspellende factor en *donorleeftijd*, *diabetes*, en *levensondersteuning* als relatief sterke voorspellers. Volgens RSF was *donorleeftijd* de meest voorspellende variabele, gevolgd door *hertransplantatie*, *levensondersteuning* en *serologiestatus van het chronische hepatitis C-virus*. Al met al werd aangetoond dat ML-technieken een nuttig hulpmiddel kunnen zijn voor zowel voorspelling als interpretatie in deze overlevingscontext.

In de vorige studie leverde onze groep nieuwe methodologische uitbreidingen van het PLANN-model. De PLANN uitbreiding werd ontwikkeld en gevalideerd voor complexe levertransplantatiegegevens. Het is echter niet ongebruikelijk dat een klein aantal patiënten wordt gerekruteerd in klinische studies en een beperkt aantal voorspellende kenmerken verzameld worden, bijvoorbeeld in sarcoomonderzoeken. Toch verwachten clinici dat ML-modellen mogelijk beter presteren dan SM. Daarom lag de focus in **Hoofdstuk** 7 op de vergelijking tussen dergelijke modellen voor niet-complexe klinische gegevensbanken (kleine / middelgrote steekproefomvang, laagdimensionaal) gecomplementeerd met een Monte Carlo-simulatiestudie om een verschillende real-life settingen te kunnen bestuderen. Er werden synthetische gegevens (250 of 1000 patiënten) gegenereerd die sterk leken op vijf prognostische factoren die vooraf waren geselecteerd op basis van een Europese Osteosarcoom Intergroup-studie (MRC BO06/EORTC 80931) waarin het effect van dosis-intensieve chemotherapie werd onderzocht bij patiënten met gelokaliseerd osteosarcoom in de extremiteiten. De voorspellende prestaties van PLANN original en PLANN extended (met één verborgen laag) werden vergeleken met Cox-modellen voor 20, 40, 61 en 80% gecensureerde gegevens. Overlevingstijden werden gegenereerd op basis van een log-normale verdeling. Het eindpunt van belang was de totale overleving, gedefinieerd als de tijd tot overlijden door welke oorzaak dan ook sinds de datum van de operatie. Modellen werden geëvalueerd op basis van de C-index, Brier-score op 0-5 jaar, geïntegreerde Brier-score (IBS) op 5 jaar en miskalibratie op 2 en 5 jaar (een parameter die meestal verwaarloosd wordt). De ML-modellen waren in staat om een vergelijkbare voorspellende prestatie te bereiken op gesimuleerde gegevens voor de meeste scenario's met betrekking tot de C-index, Brier-score of IBS. De SM waren echter vaak beter gekalibreerd. De prestaties waren robuust in scenario's waarin gecensureerde patiënten werden verwijderd voor het 2^{de} jaar of administratieve censurering na 5 jaar werd uitgevoerd (op trainingsgegevens). Onderzoekers moeten zich bewust zijn van de tijdsintensieve aspecten van het werken met ML-technieken, zoals data preparatie, afstemming van hyperparameters en rekentijd, waardoor ze nadelig zijn ten opzichte van conventionele regressiemodellen in een eenvoudige klinische setting.

In gezondheidsonderzoek zijn verschillende chronische ziekten vatbaar voor concurrerende risico's (CR's). Aanvankelijk werden SM ontwikkeld om de cumulatieve incidentie van een interessante gebeurtenis te schatten in de aanwezigheid van CR's. Dankzij de groeiende interesse in het toepassen van ML voor klinische voorspelling, zijn deze technieken ook uitgebreid naar CR's, maar de literatuur is nog beperkt. In **Hoofdstuk** 8 wilden we prognostische klinische voorspellingsmodellen voor CR's ontwikkelen en valideren met SM- en ML-technieken. Twee SM a) oorzaak-specifieke Cox, b) Fine-Gray-model en drie ML-modellen i) PLANN origineel voor CR's (PLANNCR origineel), ii) een methodologische uitbreiding genaamd PLANNCR uitgebreid, en iii) RSF voor CR's (RSFCR) werden gebruikt. De voorspellende prestaties van alle methoden werden beoordeeld in termen van discriminatie en kalibratie in een andere eenvoudige klinische setting (kleine / middelgrote steekproefomvang, klein aantal factoren). De beschikbare dataset bevat 3826 retrospectief verzamelde gegevens van patiënten met extremiteit STS (eSTS) en negen variabelen van de gepersonaliseerde SARcoma Care (PERSARC) Study Group. Voor zover wij weten, was dit de allereerste studie van deze soort voor eSTS. Het klinische eindpunt was de tijd in jaren tussen operatie en ziekteprogressie (interessante gebeurtenis) of overlijden (concurrerende gebeurtenis). De Brier-score, het gebied onder de curve (AUC) en de miskalibratie van het model werden gebruikt om de voorspellende prestaties na respectievelijk 2, 5 en 10 jaar te evalueren. De resultaten toonden aan dat de ML-modellen een vergelijkbare prestatie kunnen bereiken met de SM op basis van de Brier-score en AUC voor ziekteprogressie en overlijden (95% betrouwbaarheidsintervallen bij 2, 5 en 10 jaar overlapten). Niettemin waren de SM vaak beter gekalibreerd. Over het algemeen zijn ML-technieken minder praktisch omdat ze een aanzienlijke implementatietijd vergen (voorbereiding van de data, afstemming van hyperparameters, rekenintensiteit). Als zodanig moeten deze technieken voor niet-complexe real-life problemen alleen worden toegepast als aanvulling op SM als verkennende instrumenten voor de prestaties van modellen. Meer aandacht voor modelkalibratie is dringend nodig.

Acknowledgements

I would like to devote this section in all those who were there during the realisation of this research.
More specifically:

Professor Fiocco, dear Marta, all this started because of you, thank you very much for the opportunity to work with you at the Mathematical Institute during my master dissertation and your recommendation for an external collaboration with the European Organisation for Research and Treatment of Cancer (EORTC) that led me to Brussels to pursue a PhD. Your guidance, vision, and support were really essential to keep track of my projects' timelines and to think of new ideas.

Doctor Litière, dear Saskia, thank you very much for your help to integrate in such a special organisation as the EORTC, for your patience, valuable feedback, and pragmatism which have made my daily activities much more stimulating and meaningful. I would like to express my gratitude for your frequent psychological support during a very challenging period for me, and for allowing me to homework from Athens for an extended period of time to support my family.

Professor Gelderblom, dear Hans, I appreciate the opportunity you gave me to be part of the Leiden University Medical Center (LUMC) department of medical oncology, and I would like to thank you for being willing to think along quickly and provide me your precious clinical input during our monthly meetings. Your excellent collaboration with Marta and Saskia has made my PhD life easier.

The European Organisation for Research and Treatment of Cancer - Soft Tissue and Bone Sarcoma Group (EORTC - STBSG), the EORTC Cancer Research Fund (ECRF), and Leiden University Medical Center (LUMC) department of medical oncology, thank you for the financial support of my fellowship at EORTC headquarters.

The Scientific Registry of Transplant Recipients (SRTR), the PERsonalised SARcoma Care (PERSARC) Study Group, and the Medical Research Council (MRC), thank you for sharing the datasets used to perform the analyses of this thesis.

The EORTC statisticians (or non-statisticians), dear all, you are such a unique group. Thank you for all the lessons you have taught me, the great conversations, and for showing me the really healthy atmosphere of a non-profit organisation. Special thanks to Zeina, Aleksandra, Catherine, Jammbe, Anouk, and Stefan for their availability and encouragement. I would also like to thank Saïda for her daily support as the fellowship coordinator.

The EORTC fellows, dear guys, thank you for being a part of this exciting journey. I hope you enjoyed your time in Brussels as much as I did, and that this period has helped you grow both professionally and personally. Special thanks to my office-mates during the first year Nicolas, Blaise, and Lien.

My good friends (from EORTC or not), dear Facundo, Felix, Lambert, Maria, Cynthia, Daniel, Andrea, Dea, Sokratis, Thodoris, Dimos, thank you very much for making my life so much better and interesting. Thank you for the great time we have spent together, and the very fruitful conversations for work and life. Your support has been fundamental especially during this very challenging family situation. I cherish all the moments we spent and we will spend together.

My family members, dear father Ioannis, mother Ourania, sister Angelikoula, the last 1.5 year was very demanding for all of us. However, this situation also strengthened our bonds more than ever. Thank you for your daily unconditional love and support. Dear father, rest in peace.

Until we meet again.

Curriculum vitae

Georgios Kantidakis was born on November 7^{th}, 1993, in Athens, Greece. After graduating from the 4^{th} high school of Alimos, Athens (2011), Georgios obtained a bachelor's degree in Mathematics at the National and Kapodistrian University of Athens (2011 - 2016). During his bachelor, he worked as intern in the National Bank of Greece.

In September 2016, Georgios moved to Leiden, the Netherlands, to pursue a master's degree in Statistical Science for the Life and Behavioral Sciences at Leiden University (2016 - 2018). During his master studies, he carried out an internship investigating the effect of dose reduction and delays in duration of chemotherapy in osteosarcoma patients under the supervision of prof. dr. Marta Fiocco at the Mathematical Institute of Leiden University. This project was a collaboration with prof. dr. Hans Gelderblom at the department of medical oncology of Leiden University Medical Center (LUMC). His master thesis focused on prediction models since liver transplantation with an emphasis on the comparison between traditional statistical models and machine learning techniques under the supervision of prof. dr. Marta Fiocco.

His internship and master thesis projects sparked an external collaboration with the European Organisation for Research and Treatment of Cancer (EORTC). From November 2018, Georgios moved to Brussels, Belgium, as a fellow bio-statistician at the department of statistics in EORTC headquarters, and started working on his PhD projects as a combined function under the supervision of prof. dr. Marta Fiocco, dr. Saskia Litière at EORTC headquarters, and prof. dr. Hans Gelderblom (2018 - 2022). Georgios has been teaching assistant at the Survival analysis (Advanced Biostatistics) course of LUMC in 2021 and 2022. During his PhD time, he has presented his research at conferences in Belgium, the Netherlands, and France and in several virtual meetings (ISCB 2020, ESMO 2020, ISCB 2021) after the COVID-19 pandemic outbreak. He has been working on research projects for the EORTC – Soft Tissue and Bone Sarcoma Group (STBSG), and on investigating the potential of existing and novel machine learning models compared to statistical methods for sarcoma and non-sarcoma clinical data focusing on prediction of time-to-event outcomes.

List of publications

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, 2020. ISSN 14712288. doi: 10.1186/s12874-020-01153-1.

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.

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*, 2021:2160322, 2021. ISSN 1748-670X. doi: 10.1155/2021/2160322.

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*, 2022:5815875, 2022. ISSN 1369-1643. doi: 10.1155/2022/5815875.

G. Kantidakis, S. Litière, A. Neven, M. Vinches, I. Judson, J. Y. Blay, E. Wardelmann, S. Stacchiotti, L. D'Ambrosio, S. Marréaud, W. T. A. van der Graaf, B. Kasper, M. Fiocco, and H. Gelderblom. New benchmarks to design clinical trials with advanced or metastatic liposarcoma or synovial sarcoma patients: A EORTC Soft Tissue and Bone Sarcoma Group (STBSG) meta-analysis based on a literature review for soft-tissue sarcomas. *European Journal of Cancer*, 174:261-276, 2022. doi: 10.1016/j.ejca.2022.07.010.

G. Kantidakis, A. D. Hazewinkel, and M. Fiocco. Neural networks for survival prediction in medicine using prognostic factors: a review and critical appraisal. *Computational and Mathematical Methods in Medicine*, 2022:1176060, 2022. doi: 10.1155/2022/1176060.

G. Kantidakis, H. Putter, S. Litière, and M. Fiocco. Statistical models versus machine learning for competing risks: development and validation of prognostic models. *Submitted*.

R. Saesen, **G. Kantidakis**, A. Marinus, D. Lacombe, and I. Huys. How do cancer clinicians perceive real-world data and the evidence derived therefrom? Findings from an international survey of the European Organisation for Research and Treatment of Cancer. *Frontiers in Pharmacology*, 13:969778, 2022. doi: 10.3389/fphar.2022.969778.