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

Kantidakis, G.

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.

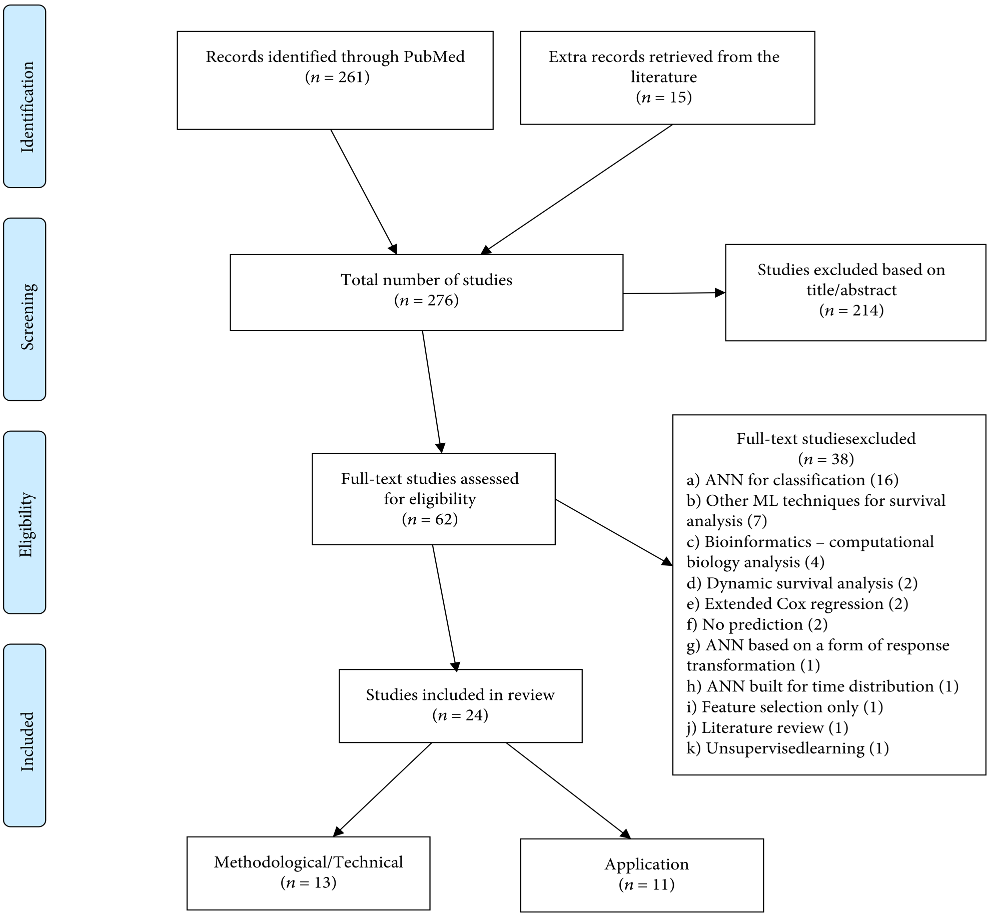
Identification
Screening
Eligibility
Included
Records identified through PubMed (n = 261)
Extra records retrieved from the literature (n = 15)
Total number of studies (n = 276)
Studies excluded based on title/abstract (n = 214)
Full-text studies assessed for eligibility (n = 62)
Full-text studiesexcluded (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.

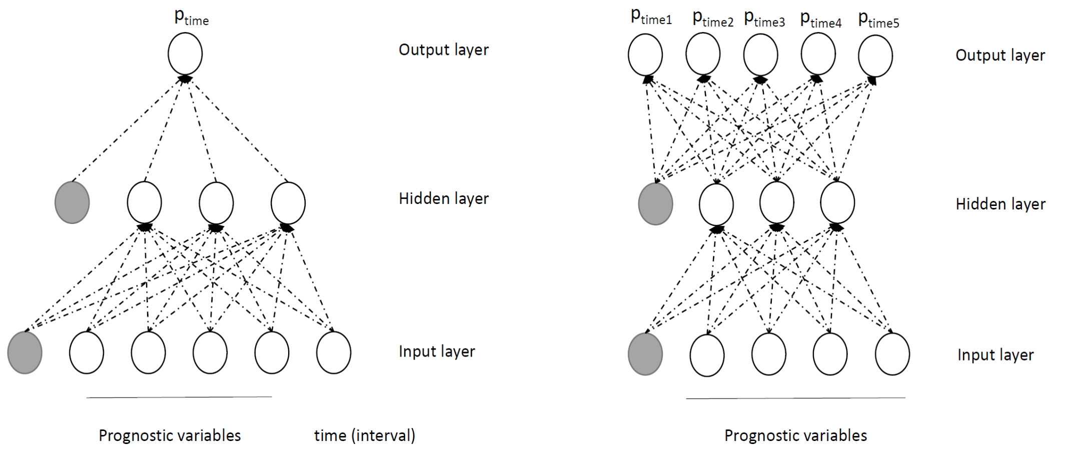

Figure 5.2: **Two basic architectures of survival neural networks.** Left panel: A network where time (interval) is coded as a prognostic variable (input feature). Data transformation into a long format is required for each patient. The output layer makes predictions in a given time interval. Right panel: 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].

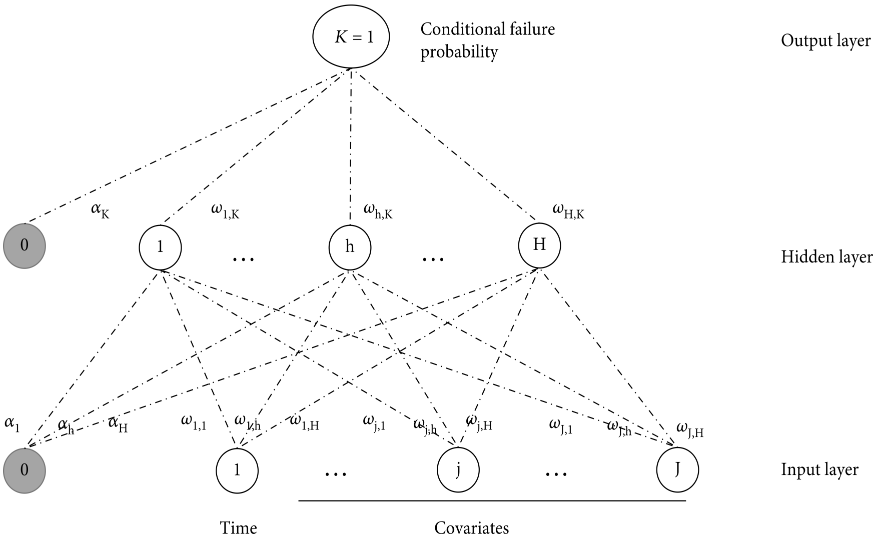

Figure 5.3: Visualization of the PLANN by Biganzoli *et al.* (1998) [20]

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

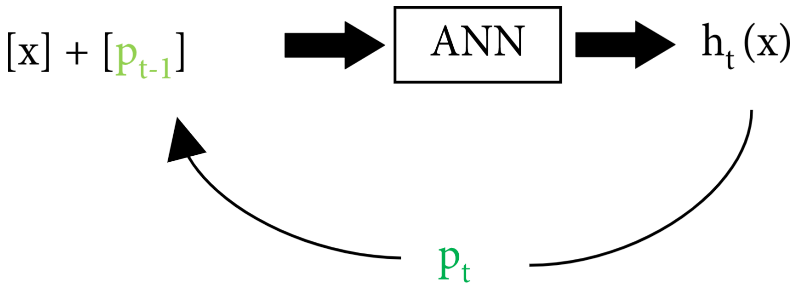

Figure 5.4: A schematic representation of the SNN by Han *et al.* in 2018 adapted from [33] built for 242 patients with synovial sarcoma. Here ANN means artificial neural network, $\boldsymbol{x}$ is the set of 9 clinical features, p_{t-1} is the survival probability of the previous year $t-1$ sequentially 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. Hyper-parameter 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.