STATISTICS IN MEDICINE
Statist. Med. **17**, 1169–1186 (1998)

FEED FORWARD NEURAL NETWORKS FOR THE ANALYSIS OF CENSORED SURVIVAL DATA: A PARTIAL LOGISTIC REGRESSION APPROACH

ELIA BIGANZOLI[1]*, PATRIZIA BORACCHI[2], LUIGI MARIANI[1] AND ETTORE MARUBINI[1,2]

[1]*Divsione di Statistica Medica e Biometria, Istituto Nazionale per lo Studio e la Cura dei Tumori, Via Venezian, 1, 20133 Milano, Italy*

[2]*Istituto di Statistica Medica e Biometria, Università degli Studi di Milano, Via Venezian 1, 20133 Milano, Italy*

SUMMARY

Flexible modelling in survival analysis can be useful both for exploratory and predictive purposes. Feed forward neural networks were recently considered for flexible non-linear modelling of censored survival data through the generalization of both discrete and continuous time models. We show that by treating the time interval as an input variable in a standard feed forward network with logistic activation and entropy error function, it is possible to estimate smoothed discrete hazards as conditional probabilities of failure. We considered an easily implementable approach with a fast selection criteria of the best configurations. Examples on data sets from two clinical trials are provided. The proposed artificial neural network (ANN) approach can be applied for the estimation of the functional relationships between covariates and time in survival data to improve model predictivity in the presence of complex prognostic relationships. © 1998 John Wiley & Sons, Ltd.

1. INTRODUCTION

One of the promising areas of modern statistics is the application of the research on learning processes for the analysis of complex problems.[1] Starting from the *perceptron*,[2] the first mathematical model of the parallel distributed learning process of neurons, multi-layer perceptron models, better known as *artificial neural networks* (ANNs) have been the subject of great interest since the second half of the 1980s, due to the development of the *back-propagation* algorithm.[3] In the statistical framework, modelling the underlying relationships of multivariate data implies previous definition of the correct functional relationship between the variables considered, which must be expressed by a finite number of parameters. In many applied problems this could be a hard task due to the lack of prior information on the studied phenomena. In these situations it

* Correspondence to: Elia Biganzoli; Divsione di Statistica Medica e Biometria, Istituto Nazionale per lo Studio e la Cura dei Tumori, Via Verezian 1, 20133 Milano, Italy. E-mail: biganzoli@istitutotumori.mi.it

Contract grant sponsor: Associazone Italiana per la Ricerca sol Cancro (AIRC)
Contract grant sponsor: Consiglio Nazionale delle Ricerche, Sottoprogetto 7
Contract grant number: N.96.00690.PF39

CCC 0277–6715/98/101169–18$17.50

Received December 1996
Revised August 1997

could be appropriate to consider flexible modelling approaches for both exploratory and predictive purposes. Since ANNs can be regarded as flexible models suitable for non-linear multivariate problems,[4] they have been applied to several classification and prediction tasks in the biomedical field, with the purpose of improving the discriminant power of diagnostic tools, or the prediction of outcome.[5,6] Recent papers have pointed out the extension of standard regression models for survival data as neural networks suitable for processing censored outcome time data. Approaches for grouped time data were proposed by Liestol *et al.*[7] For continuous time data Liestol *et al.*[7] proposed a piecewise constant hazard approach, while Faraggi and Simon[8] extended the linear proportional hazard Cox model with an ANN predictor. Other approaches have been published in the clinical literature, based on the use of standard ANN techniques with a particular organization of the data to deal with the censoring problem.[9] One important issue is the estimation of the conditional probability for the occurrence of a specific event as a function of time,[10,11] and of putative prognostic factors. Flexible modelling of covariate effects on the shape of this function, and its direct graphical exploration, may suggest new clinical and physiological hypotheses. Otherwise, the results may provide evidence that simpler modelling approaches, such as those based on proportional hazards and linear covariate effects, can be adopted without distortion of the true functional relationship.[12]

The aim of the present paper is the proposal of a flexible ANN approach, in a discrete survival time context, which provides smoothed hazard function estimation and allows for non-linear covariate effects. Our work starts from the definition of a general approach that can be implemented with standard ANN modelling tools. Criteria suitable for model selection are proposed, and, finally, direct graphical interpretation of model results is provided. In practice we propose an ANN model as a non-linear generalization of logistic regression, suitable for grouped failure time data. Since the approach is connected with the theory of partial likelihood we will call it PLANN from the acronym. The network model is represented in Figure 1; the input layer is composed of J units (nodes) plus one bias unit, one input node is for time while the others are for the covariates. The input nodes are fully connected with the H nodes of the hidden layer. A single output node estimates conditional failure probability values from the connections with the hidden and the bias units.

Section 2 gives the statistical framework for the application of ANN models for handling general regression problems. In Section 3 we examine the discrete model setting of survival analysis. Section 4 shows how non-linear generalization of logistic regression such as ANNs can be applied for modelling conditional probability of failure. In Section 5 we compare our ANN approach with the previous proposals appearing in the statistical literature. Two examples, with single and multiple covariate analyses, are reported in Section 6. In the first example, we applied the PLANN approach to data sets from a two-arm clinical trial conducted by the California Oncology Group on patients with head and neck cancer. In this study, radiation therapy alone (arm A) was compared with radiation plus chemotherapy (arm B). The endpoint of interest was disease recurrence. Arm A included 51 patients of whom 42 suffered recurrence, while arm B included 45 patients with 31 events. Efron[13] obtained smoothed hazard estimates from these data, separately for the two treatment arms, using a logistic regression approach for grouped survival times.

In the second example we considered the data set from the Veteran's Administration (VA) lung cancer study.[14] In this trial, male patients with advanced inoperable tumours were randomized to either standard (69 subjects) or test chemotherapy (68 subjects). The primary endpoint for efficacy assessment was survival time; only 9 of the 137 were censored. Information on performance status

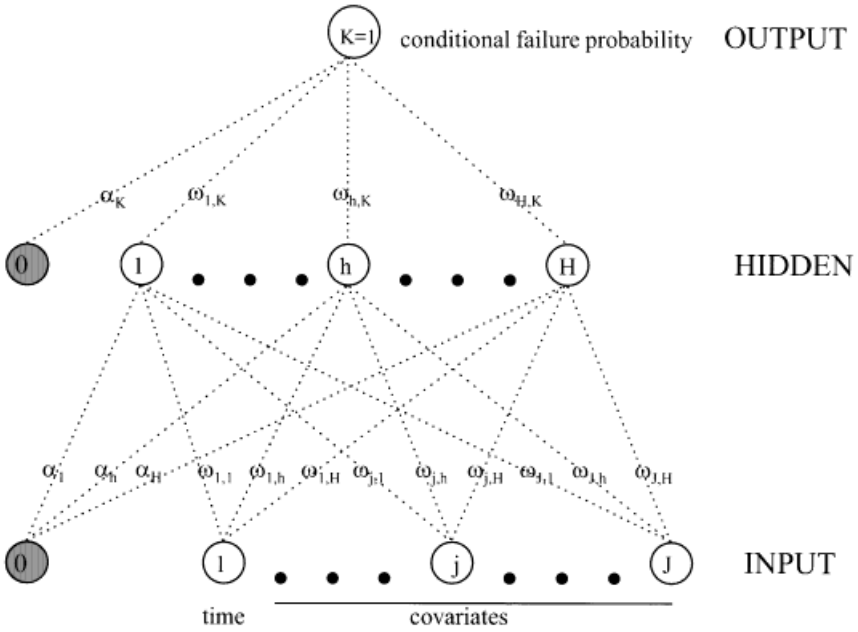

Figure 1. Feed forward neural network model for partial logistic regression (PLANN). The units (nodes) are represented by circles and the connections between units are represented by dashed lines. The input layer has J units, for time and the covariates, plus one bias unit (0). The hidden layer has H units plus the bias unit (0). A single output unit ($K = 1$) computes conditional failure probability. α_h and α_K are the weights for the connections of the bias unit with the hidden and output units. w_{jh} and w_{hK} are the weights for the connections between input and hidden units and hidden and output units, respectively

at baseline (Karnofsky rating – KPS), disease duration in months, age in years at randomization, prior therapy (yes, no), and cell type (large, squamous, small, adeno), was available.

2. ARTIFICIAL NEURAL NETWORKS AND GENERALIZED REGRESSION MODELS

The ANN term was introduced in the second half of the 1980s with the discovery of *back-propagation* as a method for jointly estimating the parameters of a multi-layer perceptron model.[3] *Feed forward* ANNs, are strictly equivalent to non-linear multivariate regression methods. Their topological interconnected structure is represented as a *directed graph of nodes without cycles.*[15] They are built (*trained*) with an initial set of observations (*patterns*) with the aim of generalizing the results to patterns not used for the generation of the model. The nodes (*neurons*) are the basic units of the model, organized hierarchically as layers, and linked by input–output relationships. A general ANN model has an input layer, one or more intermediate *hidden* layers, and the output layer. We will concentrate our attention on three layer networks with only one hidden layer with $h = 1, 2, \ldots, H$ nodes. Let $j = 1, 2, \ldots, J$ and $k = 1, 2, \ldots, K$ be the input and output nodes, respectively, x_{ij} will be the input values and y_{ik}^{o} the observed responses (*targets*) for each subject $i = 1, 2, \ldots, n$. The model will compute the outputs $\hat{y}_{ik}$ to approximate the y_{ik}^{o}. The input layer has only the role of distributing the inputs to the hidden layer. Each node in the hidden layer computes a weighted sum of the inputs x_{ij} with weights w_{jh}, adds a constant α_h (*bias*), and applies

a function (*activation*) ϕ_h to obtain its output. The outputs of the hidden layer become the inputs of the output layer nodes; their outputs are computed in the same way as the hidden layer with weights w_{hk} and activation ϕ_o. The presence of the hidden nodes provides a non-linear dependence of the outputs on the input variables. The mathematical representation of an ANN with a single hidden layer is given below:

$$\hat{y}_k(\mathbf{x}_i, w) = \phi_o\left(\alpha_k + \sum_{h=1}^{H} w_{hk}\phi_h\left(\alpha_h + \sum_{j=1}^{J} w_{jh}x_{ij}\right)\right). \tag{1}$$

The graphical representation of (1) for PLANN is in Figure 1.

The activation function ϕ_h used for the hidden nodes is generally the logistic one

$$\phi_h(u) = \frac{\exp(u)}{1 + \exp(u)}$$

while ϕ_o depends on the specific regression problem. The estimates of the weights w, parameters of the model, are obtained by minimizing an appropriate error function. Several error functions can be used based on the specific problem; the most frequent is the quadratic error

$$E = \sum_{k=1}^{K}\sum_{i=1}^{n}(\hat{y}_k(\mathbf{x}_i, w) - y_{ki}^o)^2 \tag{2}$$

while for binary classification problems the appropriate function is the cross-entropy error[4] given by

$$E = -\sum_{k=1}^{K}\sum_{i=1}^{n}\{y_{ik}^o \log \hat{y}_k(\mathbf{x}_i, w) + (1 - y_{ik}^o)\log[1 - \hat{y}_k(\mathbf{x}_i, w)]\}. \tag{3}$$

The absolute minimum of the error function (3) occurs when $y_{ik}^o = \hat{y}_{ik}$ for all the n subjects for the K outputs, and is expressed by

$$E_{\min} = -\sum_{k=1}^{K}\sum_{i=1}^{n}\{y_{ik}^o \log y_{ik}^o + (1 - y_{ik}^o)\log[1 - y_{ik}^o]\}. \tag{4}$$

Therefore, the use of binary target variables y_{ik}^o, necessarily implies that expression (4) vanishes at its minimum. Particular cases of feed forward ANNs with only input and output layers are therefore equivalent to generalized linear regression models (GLMs) with ϕ as link function, and the appropriate error term defined implicitly from the error function E. Several iterative algorithms can be used to search for the error function minimum, the most widely used is the back-propagation method (gradient descent). Other techniques are based on quasi-Newton methods.[16] Feed forward ANNs with logistic outputs can be regarded as flexible non-linear regression models for conditional probability estimation.[4] Their flexibility must be optimally tuned to achieve the best trade-off between fitting the training data and approximation of the underlying true functional dependence with the smallest bias.[17] Several techniques can be applied to modulate the degree of fitting of a neural network such as the choice of the number H of hidden nodes, the use of regularization techniques, such as the addition of a penalty term to the error function, or early stopping in the number of iterations of the optimization algorithm. All these techniques are directed to the control of the effective complexity of the model which is related to the number of parameters estimated. The choice of the best network configuration is finally based on the maximization of its predictive capability. Therefore, model selection is mainly done with validation techniques.[15]

3. DISCRETE TIME MODELS FOR SURVIVAL DATA

Different strategies can be applied to model censored survival data. For continuous-time survival data, the *hazard function* $h(t)$ is defined as:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(T < t + \Delta t | T \geqslant t)}{\Delta t}. \tag{5}$$

In the discrete context a set of L times $0 < t_1 < t_2 < \cdots < t_L$ is obtained which arises from the finite precision of time determinations. Analogously we can consider the grouping of continuous survival times into $l = 1, 2, \ldots, L$ disjoint intervals $A_l = (t_{l-1}, t_l]$ with $t_0 = 0$ and l_i the last observation interval for the ith subject, and apply the relationships derived for the discrete setting.[18] The pertinent functions are the survival function

$$S(t_l) = P(T > t_l) \tag{6}$$

the discrete probability function

$$f_l = P(T \in A_l) = S(t_{l-1}) - S(t_l) \tag{7}$$

and the discrete hazard rate h_l, defined as the conditional failure probability:

$$h_l = P(T \in A_l | T > t_{l-1}) = \frac{f_l}{S(t_{l-1})}. \tag{8}$$

The conditional failure probabilities h_l approximate the continuous hazard function $h(t)$ as the intervals A_l become infinitesimal. From (7) and (8) it is easy to verify that, having defined $S(t_0) = 1$

$$S(t) = \prod_{l:t_l \leqslant t} (1 - h_l). \tag{9}$$

The contribution to the likelihood function of the ith subject will be given by the product of conditional survival probabilities for the time intervals in which he/she is observed and the conditional failure probability in the interval A_{l_i} in which the event of interest occurs. Only the case of right censoring will be considered, so for the set U of uncensored subjects the contribution is

$$P(T_i \in A_{l_i}) = f_{l_i} = h_{l_i} \prod_{l=1}^{l_i - 1} (1 - h_{il}) \tag{10}$$

while for the set C of censored subjects it is

$$P(T_i > t_{l_i}) = S(t_{l_i}) = \prod_{l=1}^{l_i} (1 - h_{il}). \tag{11}$$

If one introduces the censoring indicator d_{il}, equal to 1 in the interval A_l containing the event of interest for the uncensored subjects, and equal to 0, otherwise, the total likelihood is

$$L = \prod_{i=1}^{n} \prod_{l=1}^{l_i} h_{il}^{d_{il}} (1 - h_{il})^{1 - d_{il}} = \prod_{l=1}^{L} \prod_{i \in R_l} h_{il}^{d_{il}} (1 - h_{il})^{1 - d_{il}} \tag{12}$$

where R_l is the set of individuals at risk in the lth interval of time. In this way a product of Bernoulli likelihoods is obtained, one for each individual i in the interval l in which he/she is

observed. If we consider a homogeneous population, (12) can be rewritten in the binomial form

$$L = \prod_{l=1}^{L} \binom{n_l}{s_l} h_l^{s_l}(1 - h_l)^{n_l - s_l} \tag{13}$$

where n_l and s_l are the number of subjects at risk and the number of failures in the time interval l, respectively. The discrete time model can be fitted by considering the observations in each time interval as independent across intervals, with the event indicator d_{li} as response variable with a Bernoulli distribution (12). Alternatively, the number of events s_l in each interval is modelled with a GLM model with binomial error, and n_l as binomial weights (13). Though approximate, the assumption of independence of the contribution to the likelihood for each individual across time intervals leads to reasonable results.[13] For covariates, Cox[19] proposed the proportional odds model for grouped survival times as follows:

$$\frac{h_l(\mathbf{x}_i)}{1 - h_l(\mathbf{x}_i)} = \frac{h_l(\mathbf{0})}{1 - h_l(\mathbf{0})} \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i) \tag{14}$$

where $\mathbf{x}_i$ is the covariate vector for subject i, $\boldsymbol{\beta}$ is the vector of regression coefficients and $h_l(\mathbf{0})$ is the baseline hazard rate for individuals with $\mathbf{x} = \mathbf{0}$. Defining $\theta_l = \log[\frac{h_l(\mathbf{0})}{1 - h_l(\mathbf{0})}]$, expression (14) is written as

$$h_l(\mathbf{x}_i) = \frac{\exp(\theta_l + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)}{1 + \exp(\theta_l + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)}. \tag{15}$$

Thus the discrete hazard rates are modelled by a logistic regression model having a predictor which is a linear combination of covariate values $\mathbf{x}_i$, with coefficients $\boldsymbol{\beta}$, and the values θ_l. This approach allows for the joint modelling of covariates and time interval effects by considering the interval l as a block factor in a GLM model, to obtain $\boldsymbol{\beta}$ estimates adjusted for the baseline log-odds.[18] In this case attention is focused on $\boldsymbol{\beta}$ rather than θ_l estimates, since this approach does not necessarily provide an interpretable shape for the discrete hazard function. A smoothed estimated of the discrete hazard function can be obtained by applying a vector of transforms $\mathbf{a}_l$ for each mid-point a_l of the time interval A_l as covariates vector. Several kinds of transformation can be considered, for example polynomials, or more flexible approaches. Efron[13] has specifically considered smoothing cubic splines in a *partial* logistic regression model as defined in connection with the theory of partial likelihood.[20] A crucial point for this approach is the choice of the number and location of the *spline* knots. Although the location of the knots could be estimated as adjunctive parameters, this approach could be very cumbersome and is rarely applied. A derivation based on logistic regression with polynomial smoothing of the hazards was introduced by Maul[21] for the joint modelling of covariate values and time interval effects. An approach based on dynamic modelling and penalized likelihood estimation was adopted in the same framework by Fahrmeir.[22]

4. PARTIAL LOGISTIC REGRESSION MODELS WITH ANN (PLANN)

In this section we propose the application of an ANN as an alternative approach to those described in the previous paragraph, for smoothed hazard rate estimates. This is achieved by flexible modelling of the joint dependence of hazards from time a_l and the covariate vector $\mathbf{x}_i$. Our work began with the GLM approach for modelling censored survival data for the discrete or

grouped situation.[13,18] We can take the negative logarithm of the likelihood (12) obtaining

$$E = -\sum_{i=1}^{n}\sum_{l=1}^{l_i}\{d_{il}\log h_l(\mathbf{x}_i, a_l) + (1-d_{il})\log[1-h_l(\mathbf{x}_i, a_l)]\} \tag{16}$$

that is equivalent to the cross-entropy error function (3). Total error (16) is summed both over the n subjects and over time intervals $l = 1, 2, \ldots, l_i$ in which the subject i is observed. It is easily shown that this can be calculated on a derived data set in which the vectors $\mathbf{x}_i$ for each subject are replicated for all the intervals in which the subject is observed, and coupled with the event indicator d_{il} defined above. By using the error function (16) in a neural network model with no hidden nodes and logistic activation function ϕ_o, a linear logistic regression model equivalent to (15) is obtained. Here the target variable is represented by the event indicator d_{il}. We propose the generalization of the *partial logistic regression* model to a feed forward ANN (PLANN) by the addition of a hidden layer of neurons. The PLANN model has one input node j assigned to each explanatory variable x_j and an additional input for the time interval a_l (Figure 1); the logistic function is used as activation for both the hidden nodes and the single output node. The output values $\hat{y}_1(\mathbf{x}_i, a_l, w)$ will provide smoothed estimates of the discrete hazard rates $h_l(\mathbf{x}_i, a_l)$. The PLANN model is not constrained to proportional odds assumptions since interaction of time and covariate effects will be modelled implicitly. The plot of the output of the PLANN model can be used to explore the shape of the hazard function depending on time and covariates. The PLANN model can be implemented by using data sets with subject vectors replicated for the time intervals as described above, and by using (16) as error function. An advantage of this kind of data structure is the possibility of straightforward use of time-dependent covariates since each subject is represented, for each observation interval, by one input vector which can change across intervals. If all the explanatory variables are categorical, the subjects can be grouped in the design cells $m = 1, 2, \ldots, M$ on the basis of covariate values, each with n_m subjects and s_m events; thus it is possible to model the empirical estimates $\hat{h}_m$ of discrete hazard rates obtained by

$$\hat{h}_m = \frac{s_m}{n_m} \tag{17}$$

which become the target values of the network. In this case the minimum of the error, which is expressed by (4) need not vanish, and its value depends on the particular data set; for this reason a suitable error function is obtained by subtracting expression (4) from (16) to obtain

$$E = -\sum_{m=1}^{M}\left[\hat{h}_m\log\frac{h_l(\mathbf{x}_i, a_l)}{\hat{h}_m} + (1-\hat{h}_m)\log\frac{1-h_l(\mathbf{x}_i, a_l)}{1-\hat{h}_m}\right]n_m \tag{18}$$

that is the Kullback–Leibler distance which always has its minimum at 0 irrespective of the type of data used.[4] This distance function has general application in the classification framework; it provides values which are equivalent to half the deviance of the logistic regression model for grouped and non-grouped cell data. In our applications we used (18) with $\hat{h}_m$ as target for grouped cell data and d_{il} for non-grouped cell data. Survival function estimates for the time intervals are calculated according to equation (9). The model can be easily implemented using software packages for ANNs based on back-propagation, or by using high-level programming languages containing specific routines for function optimization. We performed model optimization with variable-metric quasi-Newton algorithms, since this approach is generally more efficient than gradient descent techniques like back-propagation.[16] In particular, we applied the `nnet` S-plus

function provided by Venable and Ripley.[23] As explained in Section 2, network architectures with different degrees of complexity can be obtained by: (a) choice of the number of hidden nodes; (b) introduction of a penalty term in the loss function. For the latter, we adopted a commonly used approach in ANNs, called *weight decay*; it penalizes large weight values, by modifying the loss function as

$$E^* = E + \lambda \sum w^2.$$

Arguments based on Bayesian considerations[15] suggest $\lambda \approx 0{\cdot}01$–$0{\cdot}1$ depending on the degree of fit expected. The use of penalty has the advantage of both improving the convergence of optimization algorithms, and of avoiding overfitting. When weight decay is used, it is common practice to rescale covariates by multiplying them with appropriate factors so as to approximately span from 0 to 1, so as to be comparable with hidden unit outputs. Several criteria can be applied for the selection of the best model. In consideration of the computational drawbacks of cross-validation techniques, we adopted the NIC criterion suggested by Amari[24] and discussed in detail for its statistical applications by Ripley.[25] NIC is a generalization of Akaike's AIC criterion

$$\text{NIC} = \text{deviance} + 2p^*$$

where p^* is the effective number of parameters estimated. In our specific case this formula can be directly rewritten for the Kullback–Liebler distance E_{KL} as

$$\text{NIC} = 2E_{\text{KL}} + 2p^*.$$

When weight decay is used it has the effect of reducing p^* with respect to the number of connections of the chosen network model; so λ exerts an effective control on p^*. The formula for the calculation of p^*, reported by Amari,[24] is

$$p^* = \text{trace}(GQ^{-1}) \tag{19}$$

where Q is the expected Hessian matrix of the log-likelihood, and G is the expected value of the outer product of the score functions, evaluated at the fitted values of the model parameters w. Stone[26] proved that NIC is equivalent to leave one out cross-validation for large samples, it being computationally more advantageous. For the above reason, models with low NIC values have the best trade-off between the likelihood and the actual number of parameters estimated. A possible problem using this approach is that the NIC theory relies on a single minimum for the model loss function, and can be unreliable in the case of several local minima. Since the latter is common in non-linear models such as ANNs, we will consider NIC, in this context, as a criterion for exploring the performance of a large number of model configurations. On the other hand, general cross-validation techniques can also be cumbersome in the presence of multiple local minima of the loss-function.

5. PLANN AND PREVIOUSLY PROPOSED ANN APPROACHES FOR GROUPED SURVIVAL DATA

The PLANN approach is based on the discrete time context like the one proposed by Liestol *et al.*[7] Different from PLANN, the neural network model in their proposal has multiple outputs with one output node k for each interval of time l, for a total of L output nodes. The particular case of the proportional odds/hazards setting can be achieved by constraining all the weights of the connections out of each hidden node to have the same weights. A single layer network with

these constraints on the input nodes is equivalent to the linear model of (14). Data for the ith individual consists of a vector of $\mathbf{x}_i$ inputs and a target vector $\mathbf{d}_i = [d_{i1}, d_{i2}, \ldots, d_{l_i}]$. This situation implies that the network has a randomly varying number of target elements according to the time intervals l where an individual is at risk. Although this approach can be implemented by using a slight computational modification for the loss function (3), this cannot easily be done with standard ANN software. In the PLANN approach only one output node is used for the network; the loss function (18) which has general use is straightforwardly applied here, provided that vector $\mathbf{x}_i$ is replicated for all the intervals in which the ith subject is at risk, as previously described. Non-linear and non-proportional covariate effects are modelled with the approach by Liestol *et al.*, but smoothed estimates of the hazard function are not directly obtained from the network. Instead PLANN is proposed for the flexible modelling of the hazard function. In their paper Liestol *et al.* provided examples based on the subdivision of the time axis into four or five disjoint intervals; the effect of the covariates on the conditional event probability is, therefore, studied over large time intervals. Model validation is obtained by a *v-fold cross-validation*[27] procedure, with the subdivision of the original data set into five equally sized subsets used for testing five distinct models generated on the remaining part of the data; the *total error of prediction* is calculated as the sum of the log-likelihood of the test sets. The use of the NIC criterion in the PLANN approach allows for faster exploration of the results obtained from different network configurations. Ravdin and Clark[9] adopted an ANN with one output node, and the input covariate vector of each subject who presented the event is replicated for all the L time intervals, while censored subjects are replicated only for the observation intervals. The time interval index is included as an additional covariate. If a patient has been censored the target value is set to 0 for all the observation intervals. For a patient who developed the event, the target value is set to 0 for the interval before the occurrence of the event, and to 1 at the interval of the event occurrence and all subsequent intervals. Ravdin and Clark also proposed a correction for the bias introduced by the presence of data vectors of the uncensored subjects in the intervals after the occurrence of the event; randomly selected vectors of the uncensored subjects must be deleted so as to match the ratio between censored and uncensored patients estimated with the Kaplan–Meier method. It is not clear what type of error function was applied for the training of the network, but the authors stated that after training, the output of the network referred to as a prognostic index, is roughly proportional to the unconditional event probability estimated with the Kaplan–Meier method. PLANN is directed to model conditional event probabilities; when data are organized as we proposed in this paper, the subsequent estimation of survival probability is straightforward from relationship (9).

6. APPLICATION OF THE PLANN MODEL

6.1. Head and neck cancer trial

In the paper from Efron[13] the hazard estimates were based on the following cubic-linear spline $\mathbf{a}_l = (1, a_l, (a_l - 11)_-^2, (a_l - 11)_-^3)$ where $(t_l - 11)_- = \min\{(t_l - 11), 0\}$. A dynamic logit model with linear predictor $\theta_l + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i$ together with a random walk model for θ_l and $\boldsymbol{\beta}$ was adopted by Fahrmeir.[22] In our analysis, discretization of one-month intervals was applied for both arms of the study. Since only one explanatory binary variable for treatment was used, the data could be grouped into cells. We applied the PLANN model to obtain smoothed hazard estimates while jointly modelling the dependence of the discrete hazards from time and treatment. The

Table I. Search for the best model for the head and neck trial data: values of NIC

Number of hidden nodes (H)	Penalty factor (λ)			
	0·025	0·05	0·075	0·1
2	99·31	100·92	138·46	104·56
3	100·61	99·96	102·36	103·64
4	99·88	103·04	102·27	102·04
5	100·91	100·40	99·73	103·97
6	98·50	102·56	99·36	100·11
7	98·69	99·40	99·63	99·78
8	98·93	100·46	99·38	99·47
9	98·85	99·63	98·71	99·74
10	98·37	102·67	98·71	105·88
11	98·94	99·69	98·09	100·41
12	99·36	98·51	**97·81**	99·05
13	99·07	98·82	98·29	98·63
14	99·07	98·84	97·83	98·94
15	99·52	99·56	98·44	99·51

optimization process was repeated using multiple random starting points so as to verify the stability of model results in the presence of several local minima of the error surface. To select the network architecture we adopted a strategy suggested by Ripley,[28] namely, several neural network configurations were tested by modifying the number of hidden nodes and the value of the penalty factor, in a factorial array. We explored configurations of the number of hidden nodes ranging from 2 to 15, while for the penalty factor λ we tested four values: 0·025; 0·05; 0·075, and 0·1. Different seed values were used for the optimization of the same network configurations leading to substantially overlapping results. Model selection implies the choice of the appropriate complexity of the PLANN model to obtain the correct fit of the underlying true hazard function, thus preventing either underfitting or overfitting. Table I reports the results, in terms of NIC, of the search for the optimal PLANN model; although NIC values are rather irregular across the rows and columns of the table, it appears that an increasing number of hidden nodes will require increasing penalty factors to obtain lower NIC values. The configuration with the lowest NIC value has 12 hidden nodes and a penalty $\lambda = 0{\cdot}075$. One possible concern is the choice of a model with a high number of parameters to be estimated, but we found that the number of free parameters effectively estimated in the presence of the penalty term is around 6. In general the number of free parameters needed will depend upon the effective degree of complexity of the shape of the conditional hazard function. The spline approach considers the use of four linear coefficients for each arm of the study (excluding the position of the spline knot), therefore, in this case the ANN model does not seem to be overparameterized. It is noteworthy that the models with the lower NIC values provide almost equal graphical representations of the hazard function. This finding has already been discussed by Ripley for different data.[28] Figure 2(*a*) displays the results of this PLANN model; the continuous lines join the output values y_i of the neural network model. The estimates obtained with the cubic-linear spline proposed by Efron[13] are displayed in the same figure as dashed lines. The PLANN approach has generated smoothed hazard estimations by the joint modelling of the dependence of conditional failure probabilities from the time interval and the treatment covariate. The patterns for smoothed hazards obtained with the ANN model are quite similar to those obtained by Efron with the spline approach. Figure 2(*b*) shows

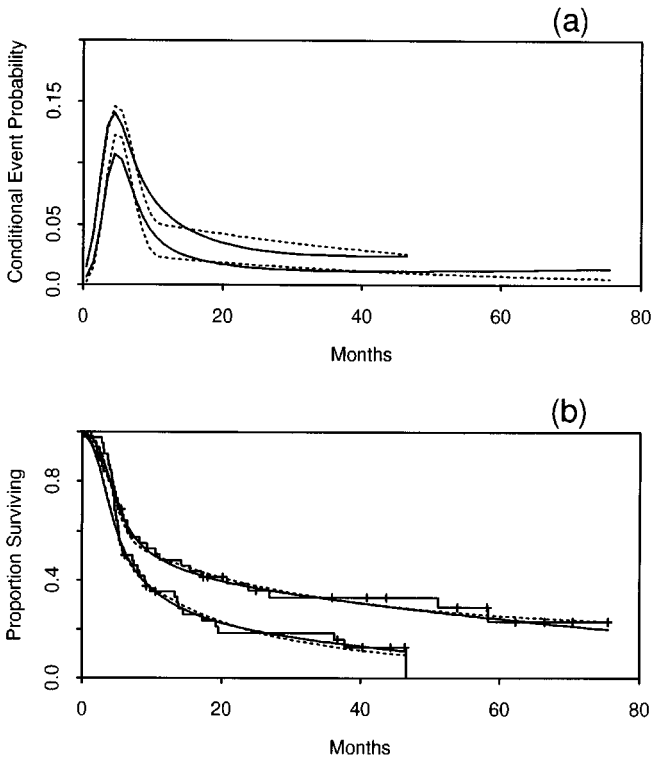

Figure 2. Head and neck cancer trial: (*a*) estimates of conditional failure probability obtained with the best PLANN configuration ($H = 12$, $\lambda = 0{\cdot}075$, solid line) and the cubic-linear spline proposed by Efron[13] (dashed line); (*b*) corresponding survival function and Kaplan–Meier estimates

the corresponding survival function calculated from equation (11), together with the Kaplan–Meier estimates. With the same number of nodes, but lowering the penalty from $\lambda = 0{\cdot}075$ to $\lambda = 0{\cdot}025$, we obtained a lower degree of smoothing and a tendency to overfit the empirical data as shown in Figures 3(*a*) and (*b*). The plot of the hazard function in Figure 3(*a*) shows a similar pattern of peaks as the corresponding one obtained by Fahrmeir[22] using dynamic discrete time models.

6.2. Veteran Administration lung cancer trial

To apply the PLANN approach, we discretized the survival times into weeks. The model was built on all the available explanatory covariates. Single indicator variables were used to distinguish treatment and prior therapy groups, three indicator variables were used for the cell type. An analysis of these data with parametric regression models and the Cox model by Kalbfliesch and Prentice,[14] showed a strong prognostic effect of KPS and of cell type, while there was no apparent dependence of survival time on age or disease duration, or the two treatments. A procedure for testing the adequacy of parametric models for these data provided evidence against both Weibull and log-normal models. Further, the shape of the failure time distribution was found to depend

Table II. Search for the best model for the VA lung cancer data: values of NIC

Number of hidden nodes (H)	Penalty factor (λ)			
	0·025	0·05	0·075	0·1
3	950·2	970·0	957·9	960·6
4	970·9	968·3	954·1	958·3
5	**935·1**	946·9	950·7	969·1
8	961·8	951·6	978·6	973·6

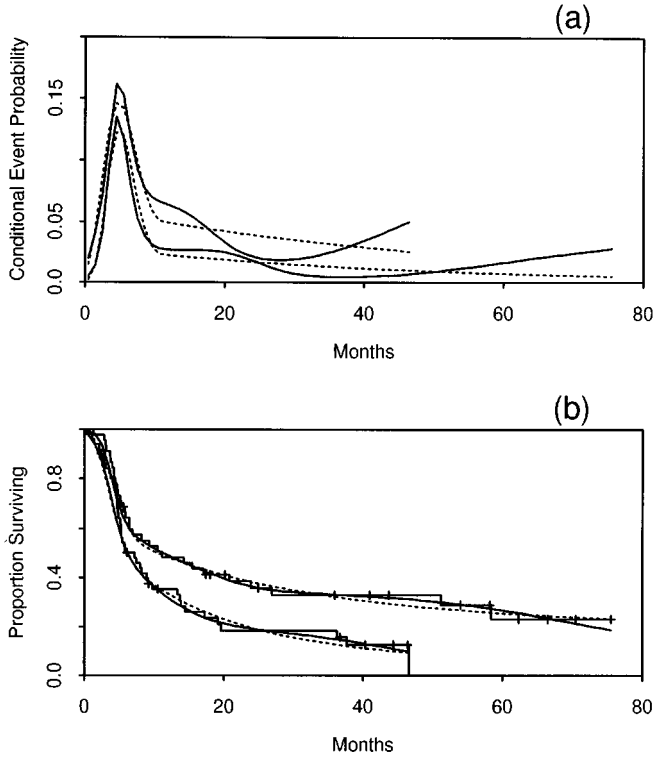

Figure 3. Head and neck cancer trial: (*a*) estimates of conditional failure probability obtained with a suboptimal PLANN model ($H = 12$, $\lambda = 0{\cdot}025$, solid line) and the cubic-linear spline proposed by Efron[13] (dashed lines); (*b*) corresponding survival function and Kaplan–Meier estimates.

on whether or not the patient had received prior therapy. For these reasons the task of modelling the underlying hazard function is certainly not easy. As in the first example we tested several network configurations with different numbers of hidden nodes and penalty coefficients in a factorial array; compared with the previous example, the analysis was conducted on a smaller number of combinations because of the higher computing times required. Table II shows the results from a series of distinct network configurations. The best performing PLANN model has five hidden nodes and $\lambda = 0{\cdot}025$ with an effective number of 42 free parameters estimated. For this model, we plotted the effects of performance status (Figure 4(*a*)) and age (Figure 4(*b*)) on the

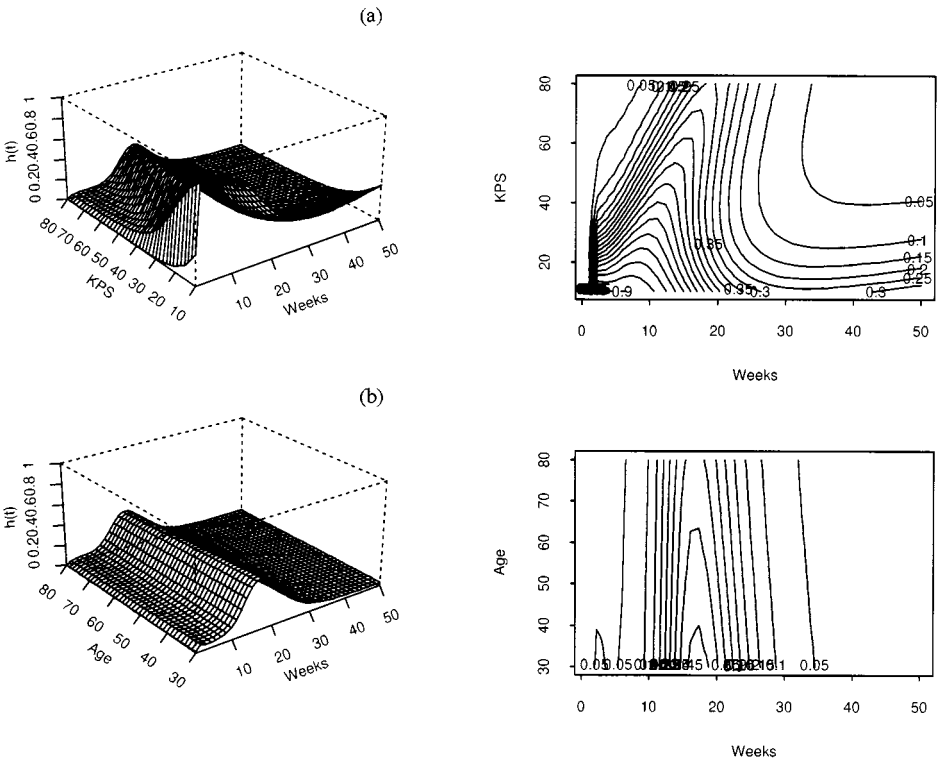

Figure 4. VA lung cancer trial: (*a*) PLANN estimates of conditional failure probability for the selected model, in dependence of time and KPS for the design cell without prior treatment, test therapy, small cell type – other continuous covariates are set to their median values; (*b*) effect of age in the same design cell

shape of the hazard function, fixing the values of the remaining variables at their medians for continuous covariates, and to no prior therapy, test treatment and small cell type for the categorical ones. The resulting three-dimensional plot shows that the effect of performance status on the shape of the conditional hazard estimates are both on the height of the peak value and its time location; this can be better examined with a contour plot. The effect of patient age seems small in this situation with the hazard peak values very slowly decreasing for older ages. The temporal shift observed for performance status is totally absent for age. The shift observed for performance status for this design cell seems concordant with the exploratory analysis carried out by Bennett.[29] In this paper smoothed empirical estimates of the hazard function were plotted for patients without prior treatment, classified by performance status as high (score over 50) or low (score 50 or below); a non-monotonic hazard shape was observed with a maximum occurring at nearly 20 days for the low KPS group while the high KPS group has a lower maximum shifted to 120 days. We did not observe the same shift in maximum hazard values for patients with prior therapy, in accordance with the findings of Kalbfleisch and Prentice,[14] of different failure time distributions depending on whether or not the patient had received prior therapy. Finally, Figure 5 shows how it is possible to explore graphically non-linear relationships and interaction

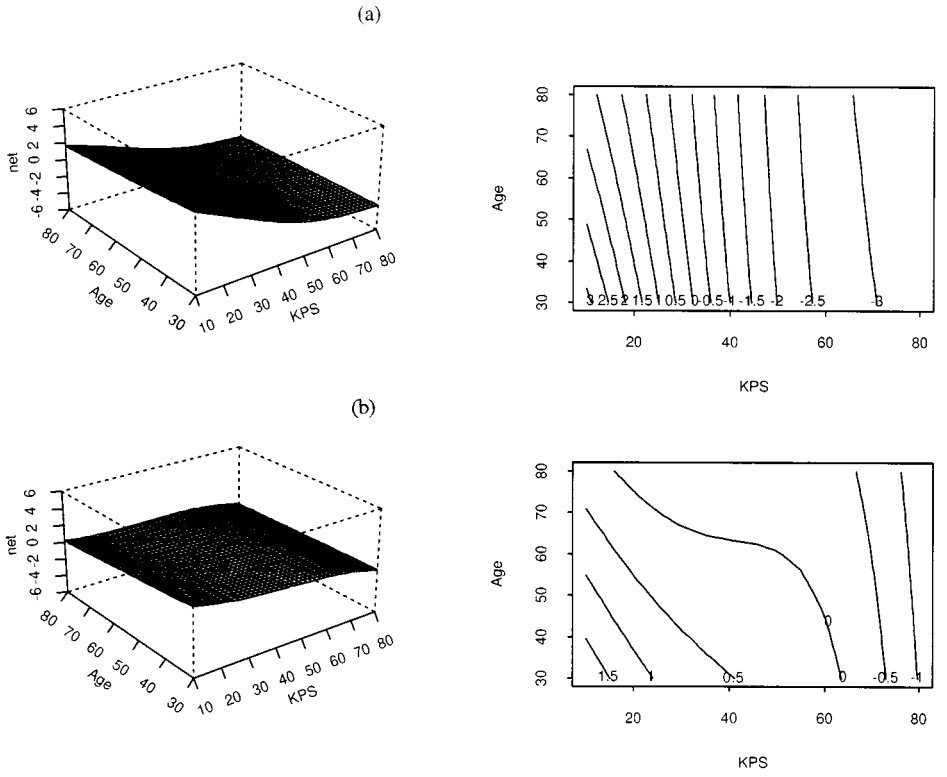

Figure 5. PLANN *net* values (before the application of the output activation function) as a function of KPS and age at fixed time intervals of (*a*) 5 weeks, and (*b*) 15 weeks

effects between explanatory variables with the PLANN approach. This figure reports *net* values, calculated by the output node before the application of the logistic activation, as a function of performance status and age. In the comparison with linear models they are the neural network equivalent of the linear predictor $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i$ of logistic regression before the application of the logistic transform, but for the PLANN model the predictor is given by the function

$$\alpha_1 + \sum_{h=1}^{H} w_{hk}\phi_h\left(\alpha_h + \sum_{j=1}^{J} w_{jh}x_{ij}\right).$$

In the absence of interaction between time and covariates it is possible to partition this function into the sum of two terms $g(a_l) + f(\mathbf{x}_i)$. Thus, in the proportional effects situation, and by fixing the time interval a_l, the predictor will be given by $f(\mathbf{x}_i) + I$ where I is a constant. For these reasons the plots of $f(\mathbf{x}_i)$ at different intervals must exhibit a parallel shift in a proportionality situation. To verify this last point, we calculated these plots for the chosen model predictor at 5 (Figure 5(*a*)) and 15 weeks (Figure 5(*b*)). Although the effect of performance status is maintained, there seems to be a change in its shape depending on time, thus suggesting the presence of non-proportional effects. With these plots the very low effect of age on

the hazards, and the slight non-linear effect of performance status, can be better appreciated for this cell design. Further, there seems to be no relevant interaction effect between the two variables considered.

7. DISCUSSION

Feed forward artificial neural networks represent a particular class of non-linear regression models so, in the statistical framework, they are mainly used for conditional probability estimation in pattern recognition.[30] These kinds of models can also be viewed as non-linear generalizations of GLMs because of the strict equivalence of the concept of link function for a GLM with the activation function in the output nodes for ANN, and the equivalence of the concept of loss function with the error terms of a GLM model. The modelling of censored survival times can be based on the estimation of the hazard, or the conditional probability of the event of interest. Previous proposals in the GLM framework jointly considered covariates and the time interval as block factors,[18] applied a flexible polynomial or spline smoothing,[13] or dynamic modelling and penalized likelihood estimation approaches.[22] The first approach allows for straightforwardly modelling linear interaction effects between covariates and/or between covariates and time, but it hardly provides interpretable shapes for the hazard function, and of the functional relationships between covariates and hazard. The second approach allows for smoothed modelling of both these functions but its correctness may be conditioned by the appropriate choice of the polynomial degree, locations of knots and constraints. This last point may be rather cumbersome in a multivariate context. The dynamic modelling approach is specifically directed towards the joint estimation of the shape of the hazard function and covariate effects, but it seems rather difficult to implement with standard algorithms. ANNs can be better suited for multivariate modelling of complex relationships between variables.[4] ANN approaches were proposed for modelling censored survival data as a generalization of the Cox model for both continuous and discrete (grouped) time data.[7,8] Our proposal starts from the linear approach for grouped survival times which introduces the time interval as a covariate in a GLM model. The PLANN approach allows for the joint modelling of time, and the continuous and categorical explanatory variables in a multi-layer perceptron model without proportionality constraints. As described above, the PLANN approach also allows for the straightforward modelling of time dependent explanatory variables. For each subject, output is the estimated conditional probability of the occurrence of an event as a function of the time interval and of covariate patterns. Flexible modelling can be applied both for predictive aims or as a powerful exploratory tool. Although common drawbacks in flexible modelling like ANNs are lacking in the direct interpretability of model coefficients and non-appropriateness of the standard statistical test, from an exploratory viewpoint ANN approaches can be very helpful for the improvement of modelling strategies. PLANN model outputs can be plotted to check whether model assumptions, such as proportional hazards and/or linear covariate effects, are tenable and to verify their impact on model predictive ability. On the contrary, potential disadvantages of PLANN are those of flexible modelling approaches: overfitting of the data used for the generation of the model itself; long computational times; presence of sub-optimal minima in the error function. A point that must be remarked upon regards optimal model selection when error penalty is applied during the optimization phase. NIC is specifically developed for this application but assumes the existence of a strong single local minimum. It is not easy to assess this condition when algorithms for local search are used. A possible approach is the repetition of the optimization process using multiple starting points to

track the minima of the error surface. By adopting this approach we verified that for the same number of hidden nodes and penalty terms, relatively stable values of p^* are obtained in the two examples. This could be an indirect indication of the usefulness of NIC, its values being mainly dependent on the model configuration, rather than on the specific local minimum. Nevertheless, we are aware that this approach can be used for initial exploratory purposes, while for predictive modelling a final validation procedure is needed.

In the first example the shape of the smoothed hazard function obtained by PLANN, jointly modelling time and treatment effects, largely overlapped the estimate provided by Efron.[13] It is also noteworthy to consider the model with higher complexity which, from NIC criterion, could be overfitted; it can provide information regarding the shape of a more complex underlying hazard function. In fact, the possibility of exploring more complex shapes with low bias could be very important from a biological point of view. By using the spline approach it is not so easy to obtain the same information without bias because of the need to specify a functional form which is less general than the ANN predictor. In Efron's example, the appropriate choice of degree of spline, of location, and number of knots is hardly feasible on the basis of the empirical hazards estimates only. In the second example the possibility of using PLANN in a multivariate context is shown, obtaining the estimates of the hazard function for different covariate combinations. The plots of network estimates showed different shapes of the hazard function for selected combinations of covariates, suggesting that a proportional hazards assumption was not tenable. This was also pointed out by other authors.[14,29] The functional relationships modelled by PLANN between logarithm of hazards and continuous covariates seem to agree with those adopted by Kalbfliesch and Prentice. Further work on simulated data sets will provide deeper knowledge of the statistical properties of the PLANN approach in comparison with traditional modelling strategies, so as to investigate the possible advantages in terms of predictive ability. Nevertheless, we think that our proposal shows how flexible modelling based on ANNs could be applied in survival analysis for exploratory purposes.

Recent papers pointed out a putative *competition*, existing between traditional linear models and their neural network extensions.[31,32] From our experience, and the above results, we think that a correct approach is the integration of traditional linear techniques with flexible ones, for optimal modelling of complex multivariate phenomena. If a statistician is faced with a problem that can be suitably modelled with linear approaches it would be disadvantageous to apply complex non-linear models. Nevertheless, in the presence of complex problems, flexible non-linear methods such as neural networks may provide additional advantages with respect to linear approaches.

ACKNOWLEDGEMENTS

We wish to thank Prof. Richard Simon from NCI Bethesda for helpful comments on the manuscript and Mrs Marina Sperni for helping with the English of the text. This work was partially supported by the Associazione Italiana per la Ricerca sul Cancro (AIRC) and by the Consiglio Nazionale delle Ricerche, Sottoprogetto 7, Grant N.96.00690.PF39

REFERENCES

1. Ripley, B. D. 'Statistical aspects of neural networks', *in* Barndorff-Nielsen, O. E., Jensen, J. L. and Kendall, W. S. (eds), *Networks and Chaos – Statistical and Probabilistic Aspects*, Chapman & Hall, London, 1993, pp. 40–123.

2. Rosenblatt, F. 'The perceptron – a perceiving and recognizing automaton', *Report 85-460-1*, Cornell Aeronautical Laboratory, 1957.
3. Rumelhart, D. E. and McClelland, J. L. (eds). *Parallel Distributed Processing*: *Explorations in the Microstructure of Cognition. Vol. 1 Foundations*, MIT Press, Cambridge, Massachusetts, 1986.
4. Bishop, C. M. *Neural Networks for Pattern Recognition*, Oxford University Press Inc., New York, 1995.
5. Baxt, W. G. 'Application of artificial neural networks to clinical medicine', *Lancet*, **346**, 1135–1138 (1995).
6. Dybowski, R. and Gant, V. 'Artificial neural networks in pathology and medical laboratories', *Lancet*, **346**, 1203–1207 (1995).
7. Liestol, K., Andersen, P. K. and Andersen, U. 'Survival analysis and neural nets', *Statistics in Medicine*, **13**, 1189–1200 (1994).
8. Faraggi, D. and Simon, R. 'A neural network model for survival data', *Statistics in Medicine*, **14**, 73–82 (1995).
9. Ravdin, P. M. and Clark, G. M. 'A practical application of neural network analysis for predicting outcome of individual breast cancer patients', *Breast Cancer Research and Treatment*, **22**, 285–293 (1992).
10. Veronesi, U., Marubini, E., Del Vecchio, M., Manzari, A., Andreola, S., Greco, M., Luini, A., Merson, M., Saccozzi, R., Rilke, F. and Salvadori, B. 'Local recurrences and distant metastases after conservative breast cancer treatments: partly independent events', *Journal of the National Cancer Institute*, **87**, 19–27 (1995).
11. De Micheli, R., Abbattista, A., Miceli, R., Valagussa, P. and Bonadonna, G. 'Time distribution of the recurrence risk for breast cancer patients undergoing mastectomy: further support about the concept of tumor dormancy', *Breast Cancer Research and Treatment*, **41**, 177–185 (1996).
12. Mariani, L., Coradini, D., Biganzoli, E., Boracchi, P., Marubini, E., Pilotti, S., Salvadori, B., Silvestrini, R., Veronesi U., Zucali, R. and Rilke, F. 'Prognostic factors for metachronous contralateral breast cancer: a comparison of the linear Cox regression and its artificial neural network extension', *Breast Cancer Research and Treatment*, **44**, 167–178 (1997).
13. Efron, B. 'Logistic regression, survival analysis, and the Kaplan–Meier curve', *Journal of the American Statistical Association*, **83**, 414–425 (1988).
14. Kalbfleisch, J. D. and Prentice, R. L. *The Statistical Analysis of Failure Time Data*, Wiley, New York, 1980.
15. Ripley, B. D. 'Neural networks and flexible regression and discrimination', *in* Mardia, K. V. (eds), *Statistics and Images 2. Advances in Applied Statistics*, **2**, Carfax, Abingdon, 1994, pp. 39–57.
16. Hertz, J., Krogh, A. and Palmer, R. G. *Introduction to the Theory of Neural Computation*, Addison-Wesley, Redwood City, CA, 1991, pp. 124–129.
17. Geman, S., Bienenstock, E. and Doursat, R. 'Neural networks and the bias/variance dilemma', *Neural Computation*, **4**, 1–58 (1992).
18. Aitkin, M., Anderson, D., Francis, B. and Hinde, J. *Statistical Modelling in GLIM*, Oxford University Press, New York, 1989, pp. 257–315.
19. Cox, D. R. 'Regression models and life-tables (with discussion)', *Journal of the Royal Statistical Society*, *Series B*, **34**, 187–220 (1972).
20. Cox, D. R. 'Partial likelihood', *Biometrika*, **62**, 269–278 (1975).
21. Maul, A. 'A discrete logistic regression model for analyzing censored survival data', *Environmetrics*, **5**, 145–157 (1994).
22. Fahrmeir, L. 'Dynamic modelling and penalized likelihood estimation for discrete time survival data', *Biometrika*, **81**, 317–330 (1994).
23. Venable, W. N. and Ripley, B. D. *Modern Applied Statistics with S-Plus*, Springer-Verlag, New York, 1994.
24. Amari, S. L. 'Statistical neurodynamics of various types of associative nets', *in* Barndorff-Nielsen, O. E., Jensen, J. L. and Kendall, W. S. (eds), *Networks and Chaos – Statistical and Probabilistic Aspects*, Chapman & Hall, London, 1993, pp. 1–39.
25. Ripley, B. D. *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, 1996.
26. Stone, M. 'An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion', *Journal of the Royal Statistical Society*, *Series B*, **39**, 44–47 (1977).

27. Stone, M. 'Cross-validatory choice and assessment of statistical predictions (with discussion)', *Journal of the Royal Statistical Society*, *Series B*, **36**, 111–147 (1974).
28. Ripley, B. D. 'Statistical ideas for selecting network architectures', *in* Kappen, B. and Gielen, S. (eds), *Neural Networks*: *Artificial Intelligence and Industrial Applications*, Springer, London, 1995, pp. 183–190.
29. Bennett, S. 'Log-logistic regression models for survival data', *Applied Statistics*, **32**, 165–171 (1983).
30. Ripley, B. D. 'Neural networks and related methods for classification', *Journal of the Royal Statistical Society*, *Series B*, **56**, 409–456 (1994).
31. Schumacher, M., Rossner, R. and Vach, W. 'Neural networks and logistic regression: Part I', *Computational Statistics and Data Analysis*, **21**, 661–682 (1996).
32. Vach, W., Rossner, R., and Schumacher, M. 'Neural networks and logistic regression: Part II', *Computational Statistics and Data Analysis*, **21**, 683–701 (1996).

