ELSEVIER

Artificial Intelligence in Medicine 28 (2003) 1–25

Artificial Intelligence in Medicine

www.elsevier.com/locate/artmed

A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer

P.J.G. Lisboa[a,*], H. Wong[a], P. Harris[a], R. Swindell[b]

[a]*School of Computing and Mathematical Sciences, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, UK*

[b]*Medical Statistics Department, Christie Hospital, Wilmslow Road, Withington, Manchester M20 4BX, UK*

Received 29 July 2002; received in revised form 15 November 2002; accepted 10 December 2002

Abstract

A Bayesian framework is introduced to carry out Automatic Relevance Determination (ARD) in feedforward neural networks to model censored data. A procedure to identify and interpret the prognostic group allocation is also described.

These methodologies are applied to 1616 records routinely collected at Christie Hospital, in a monthly cohort study with 5-year follow-up. Two cohort studies are presented, for low- and high-risk patients allocated by standard clinical staging.

The results of contrasting the Partial Logistic Artificial Neural Network (PLANN)–ARD model with the proportional hazards model are that the two are consistent, but the neural network may be more specific in the allocation of patients into prognostic groups. With automatic model selection, the regularised neural network is more conservative than the default stepwise forward selection procedure implemented by SPSS with the Akaike Information Criterion.

Keywords: Artificial neural networks; Automatic Relevance Determination (ARD); Censorship; Cohort study; Proportional hazards; Nottingham Prognostic Index (NPI)

1. Introduction

Assigning patients into prognostic risk groups is of considerable importance in the management of breast cancer patients. This process requires a quantitative model of survival, usually focused on the first 5 years following surgery, but it is also crucially

[*] Corresponding author. Tel.: +55-51-33165571; fax: +55-51-33168010.
E-mail address: mwajner@ufrgs.br (P.J.G. Lisboa).

doi:10.1016/S0933-3657(03)00033-2

important to provide some form of interpretation of the prognostic groups in terms of clinically relevant variables. In practice, oncologists frequently use an algorithm commonly referred to as the Nottingham Prognostic Index [1] which was derived using the classical statistical method of proportional hazards [2], which is a linear in the parameters approach to the modelling of censored data.

In the context of developmental models for decision support systems [3], the purpose of this paper is to conduct a pre-clinical cohort study to establish a rationale for the design of artificial neural network systems for prognostic modelling using principles of good practice in the control of model complexity.

Neural networks are attractive for this application because they overcome two potential limitations of proportional hazards modelling, namely, the assumption that the time development of the hazards is proportional to that of a fixed baseline population and the assumption that the covariates influence the model through explicit linear terms. However, the main limitation of generic non-linear models such as neural networks is their propensity to overfit the data. This demands a robust methodology to limit the model complexity in two different ways. First, by ensuring that the iterative estimation of the model parameters is stopped at a point where the generality of the model is high and, second, by carrying model selection so only the relevant variables are included in the model.

This paper describes an extension of a neural network model for censored data, the Partial Logistic Artificial Neural Network (PLANN) [4], to include regularisation within a Bayesian framework that enables model selection by a method called Automatic Relevance Determination [5]. The second part of the paper presents a method to identify and characterize prognostic risk groups in a study of two cohorts of breast cancer patients, using records acquired during routine clinical practice.

2. Data description

The data used in this study consist of the records of 1616 female patients referred to the Manchester Christie Hospital between 1983 and 1989. The event of interest was defined to be death attributed to breast cancer, as determined from the coronary report or by a specialist consultant. Other causes of death and other losses to follow-up were regarded as censorship, and all surviving patients are censored after 5 years.

A total of 18 categorical variables were collected, which are summarised in (Table 1). The data set contains missing values, particularly in the following variables: *number of nodes involved* (968 missing), *oestrogen level* (537 missing) and *pathological size* (414 missing). This leaves 447 complete records, a proportion that is not atypical of historical hospital records, bearing in mind that they are usually not collected for the purpose of designing decision support systems, but rather to provide an audit trail for evidence-based clinical practice. However, this poses a significant difficulty for the statistical modelling of the data.

There are two main ways of dealing with substantial amounts of missing data. They can be filled-in by modelling using the available data, or they may be coded as a separate attribute. The former method is commonly used where the data are believed to be missing at random, so that missing attributes are not expected to carry information that substantially

Table 1
Variables recorded in the standard breast cancer database

Variable	Number of attributes
Menopausal status	3
Age group	3
Predominant site	5
Side	2
Maximum tumour diameter	3 + unknown
Clinical stage tumour	T0 (no tumour) + T1–4
Clinical stage nodes	N0–3
Clinical stage metastasis	M0–1
Clinical stage (TNM)	5
Radiotherapy	2
Histology	3 + unknown
Surgery	9
Number of nodes involved	4 + unknown
Adjuvant treatment	16
Number of nodes removed	4 + unknown
Nodes ratio	4 + unknown
Pathological size	3 + unknown
Oestrogen level	3 + unknown

affects the model outcome. However, in these data the missing values are not at all random, and often the attribute missing correlates with a poor outcome. This is to be expected since record entry is more likely to be cut-short when there is sufficient evidence to justify a particular prognostic outcome, which would normally arise where a few key attributes are sufficient to indicate that the disease is advanced. Survival plots of each variable confirmed that the missing values do not always approximate the mean survival. For this reasons, it was decided to code missing values for the purpose of this study as a separate attribute.

It was mentioned in the introduction that the Nottingham Prognostic Index for breast cancer is widely used in clinical practice. However, these data do not record triple lymph node biopsy, which is the sampling procedure required by that Index. In this case, clinicians frequently employ the internationally recognized TNM clinical stating method, which stands for tumour, lymph nodes and metastatic spread. Patients were partitioned using TNM staging into low risk and high cohorts, in a procedure that is customary both in the clinical management of these patients and in parametric statistical analysis. The low risk group was defined by *tumour stage* either 1 or 2 and *pathological size* either <2 or 2–5 cm, *node stage* either 0 or 1, and *metastasis stage* 0. This is a typical assignment of operable primary breast tumour patients, to whom the Index applies. The remaining patients form a high-risk cohort.

There were 917 patients in the low risk cohort, 633 patients in the high-risk cohort and 66 records with *tumour stage* 0 were discarded. Although low risk is assigned to patients with tumour less than 5 cm in the maximum dimension, with at most a few mobile affected lymph nodes and with no detectable metastatic spread, this group will be shown to contain risk groups that overlap entirely with groups identified in the high-risk cohort. This may be due, in part, to the assignment of records with missing values of *pathological size* to high-risk.

Censorship is an inherent feature of survival data that arises when follow-up stops before the end of the study period, censoring all further information regarding the occurrence of the event of interest. An example of censorship is an intercurrent death, that is to say a patient who dies within the follow-up period but from a cause not attributed to breast cancer. It is the occurrence of censorship that differentiates survival modelling from other forms of statistical analysis, such as binary classification.

The next section describes existing statistical and neural network models for censored data. Section 4 introduces the theoretical contribution of the paper in the application of a Bayesian regularisation framework to the PLANN model, followed in Section 5 by a practical contribution, which introduces a framework for using neural networks to make prognostic assignments of individual patients to risk groups and for the interpretation of these groups in terms of clinical variables.

3. Modelling survival

Survival models apply to censored data, therefore, they require an objective function that includes patients in the risk group only for as long as they are observed to be alive, but removes them from the study once they are censored. Our study is carried out monthly over 5-years, therefore, all patients are censored after 60 months. In both models used in this study the appropriate objective function data is the likelihood of hazard. In a discrete time model the hazard is the probability of the event of interest occurring in a specific time interval, in other words the posterior for a death attributed to breast cancer being observed in a particular time interval, conditional on survival to the start of the interval, namely,

$$h(t_k) = P(t \leq t_k | t > t_{k-1}). \tag{1}$$

The precise form of the expression for the likelihood takes different forms for the proportional hazards model and the neural network. In each case, since it does not take direct account of the actual event or censorship times, it is called a partial likelihood function [6].

The next two sections introduce the benchmarking statistical model and the neural network survival model followed in this paper.

3.1. The benchmark model: proportional hazards

The proportional hazards model, sometimes called Cox regression [1], is a multiple linear regression of the hazard function, under the assumption that all time dependence is specified by a reference group of patients called the baseline population. It is possible to introduce time dependence via interaction terms, but this adds considerable complexity to the model design and is generally a heuristic procedure reliant on the statistical expertise of the user.

For discrete time intervals, the proportional hazards model parameterises the odds of survival in proportion to a baseline, as follows [6],

$$\frac{h_p(t_k)}{1 - h_p(t_k)} = \frac{h_0(t_k)}{1 - h_0(t_k)} \exp(\beta^{\mathrm{T}} x_p), \tag{2}$$

where p denotes the individual patient record and x_p is a static covariate vector containing a set of explanatory variables extracted from the patient record. We are interested only in the variables available immediately after surgery, therefore, we exclude the treatment attributes from the study. A baseline population must be selected to establish the time dependence of the hazards. Its covariate vector will contain all zeros. There is no deterministic procedure to do select which attributes should form the baseline, but we have followed standard practice in consistently selecting the attributes that are expected to result in the higher survival. Note that the dependence on the covariates is aggregated into the scalar $\beta^{\mathrm{T}} x_p$ which represents a risk score or prognostic index (PI). Given our choice of baseline, this index is expected to be negative. Prognostic indexes have been used to allocate patients into prognostic groups ranked by mortality risk [7,8].

In the limit of infinitely short time intervals, the discrete time model in Eq. (2) converges to the familiar parameterisation of proportional hazards in continuous time given by

$$h_p(x_p, t) = \exp(\beta^{\mathrm{T}} x_p) h_0(t). \tag{3}$$

The survival function for the pth individual is calculated using [6]

$$\hat{S}_p(t) = [\hat{S}_0(t)]^{\exp(\hat{\beta}^{\mathrm{T}} x_p)}. \tag{4}$$

3.2. A neural network model for censored data: overview of the PLANN model

Artificial neural networks (ANN) are non-linear, semi-parametric models that have recently been considered as alternative methods for survival analysis in the presence of censorship [3]. Such models form a natural non-linear extension of the discrete time proportional hazards, since using the standard MLP network structure with time as an input the analytical expression for the network becomes

$$\frac{h_p(x_p, t_k)}{1 - h_p(x_p, t_k)} = \exp\left(\sum_{h=1}^{N_h} w_h g\left(\sum_{i=1}^{N_i} w_{ih} x_{pi} + w t_k + b_h\right) + b\right), \tag{5}$$

where the indices i and h denote the input and hidden node layers, respectively, and the non-linear function $g(\cdot)$ has the usual sigmoidal form

$$g(a) = \frac{1}{1 + \mathrm{e}^{-a}}. \tag{6}$$

In this model the dependence on the covariates and on time is combined into the non-linear term on the right-hand side of Eq. (5). If the sigmoid were replaced by a linear function, then the argument of the exponential would become $\beta^{\mathrm{T}} x_p + \theta_1 t_k + \theta_2$, returning to a factorisation of the dependence of the discrete time hazard on the explanatory variables and time, but with a sigmoidal parametric form assumed for the latter. Once the network weights w are estimated, the survivorship is calculated from the estimated discrete time hazard by multiplying the conditionals for survival over successive time intervals, treated as independent events, to give

$$S(t_k) = \prod_{l=1}^{k} P(t > t_l | t > t_{l-1}) = \prod_{l=1}^{k} (1 - h(t_l)). \tag{7}$$

The estimation of the weights requires a likelihood term that reflects the status of the patient p at time t_k. This is achieved with an indicator label, or target τ_{pk} which assumes the value of 0 if the patient is observed to be alive and 1 to indicate a death attributed to breast cancer in the time interval $t_{k-1} < t \leq t_k$. It is convenient to express the resulting expression as a negative log-likelihood,

$$G = -\sum_{p=1}^{\text{no. of patients}} \sum_{k=1}^{t_l} [\tau_{pk} \log(h_p(x_p, t_k)) + (1 - \tau_{pk})\log(1 - h_p(x_p, t_k))]. \tag{8}$$

where t_l is the month when the patient was last observed. The generic non-linear model in Eqs. (5)–(7) with parameters estimated by minimizing the cost function in (8) is called the Partial Likelihood Artificial Neural Network (PLANN) [4].

In a discrete time study of survival with PLANN, the network output represents the predicted hazard for a fixed time interval, whose midpoint is entered as a separate input variable. This model does not require proportionality of the hazards over time and it implicitly models interactions between the explanatory variables and with time. Moreover, it predicts a smooth hazard function that is independent of the baseline population.

However, generic non-linear models such as neural networks are prone to over-fitting the data unless careful regularisation is applied in order to control the complexity of the model. A Bayesian framework has proved to be robust in estimating the weight parameters [10] and it also lends itself to model selection [5]. This framework is described in detail in the next section.

4. Automatic Relevance Determination

Bayes' theorem has been proposed as a principled regularisation framework to ensure generality of the model predictions for new data. The implementation of this method involves three steps in sequence. First, a penalisation term is added to the objective function. Second, the regularisation parameters that control the penalty term are estimated. Third, the whole framework is interpreted as the evidence in favour of candidate network structures, enabling model selection to be carried out. In addition, when modelling heavily skewed data such as is typical in survival modelling, where the observed deaths are very much fewer then the number of months when patients are observed alive, the prevalence of the different binary values of the indicator label must be taken explicitly into account. Each of these aspects is now discussed in turn.

4.1. Bayesian regularisation framework for ARD

Let us denote the PLANN parameter set w, the data D, the penalty parameters α and the model hypothesis H.

In a Bayesian framework, the purpose of parameter estimation is to maximize the evidence for the weight set $\{w\}$, which is given by

$$P(w|D, \alpha, H) = \frac{P(D|w, \alpha, H)P(w|\alpha, H)}{P(D|\alpha, H)}. \tag{9}$$

The numerator consists of the likelihood that the model fits the data,

$$P(D|w,\alpha,H) = \mathrm{e}^{-G}, \tag{10}$$

multiplied by the prior distribution of the weights, which is normally assumed to be centred at zero with variance $1/\alpha$,

$$P(w|\alpha,H) = \frac{\mathrm{e}^{-E(w,\alpha)}}{Z_W(\alpha)}, \tag{11}$$

where $E(w,\alpha) = (1/2)\sum_{m=1}^{N_\alpha}\alpha_m\sum_{n=1}^{N_m}w_{mn}^2$. The index n indicates a group of weights w_{mn} sharing a common regularisation parameter α_m of which there are N_m. These weights correspond to attributes from a single field, or variable. As the training progresses, the α_m for variables with little predictive power increase in size, forcing the corresponding weights towards zero, hence, the term 'weight decay' commonly used for this regularisation method. This is the main mechanism for complexity control in the neural network, and it has a key rôle later during model selection. The normalisation constant is readily calculated from a product of univariate normal distributions, giving $Z_W(\alpha) = \prod_{m=1}^{N_\alpha}(2\pi/\alpha_m)^{N_m/2}$.

Finally, the weights are estimated by an iterative 'training' process that optimises

$$P(w|D,\alpha,H) \propto \mathrm{e}^{-G}\,\mathrm{e}^{-E(w,\alpha)} = \mathrm{e}^{-S(w,\alpha)}, \tag{12}$$

by minimizing the penalized objective function $S(w,\alpha) = G + E(w,\alpha)$. In our study this was carried out by scaled-conjugate gradients optimisation, implemented in the Matlab code Netlab [11].

4.2. Setting the regularisation parameters

In general it is possible to adjust the regularisation parameters, sometimes called 'hyper-parameters', with empirical measures of generality such as cross-validation. However, this approach is computationally very expensive and, when extended to model selection, may not be robust. An alternative is to use the Bayesian framework introduced earlier.

In order to adjust the hyper-parameters to suitable values we need to maximize the corresponding Bayesian expression

$$P(\alpha|D,H) = \frac{P(D|\alpha,H)P(\alpha|H)}{P(D|H)}. \tag{13}$$

The first-term in the numerator is the normalizing constant in Eq. (9) and therefore it can be written as

$$P(D|\alpha,H) = \int P(D|w,\alpha,H)P(w|\alpha,H)\,\mathrm{d}w = \int \frac{\mathrm{e}^{-S(w,\alpha)}}{Z_W(\alpha)}\,\mathrm{d}w. \tag{14}$$

This integral is not analytical but its value may be approximated using a Taylor expansion about the 'most probable' weights w^{MP} [10] estimated as described in the previous section, to yield

$$S^*(w,\alpha) \approx S(w^{\mathrm{MP}},\alpha) + \tfrac{1}{2}(w-w^{\mathrm{MP}})^{\mathrm{T}}A(w-w^{\mathrm{MP}}), \tag{15}$$

where the matrix A is the Hessian of S with respect to the weights. In effect we are specializing to a unimodal distribution around the estimated weights, assuming a multivariate normal form.

With this approximation, the evidence for the hyper-parameters becomes [10]

$$P(\alpha|D,H) \propto \frac{\exp(-S(w^{\mathrm{MP}},\alpha))}{Z_W(\alpha)}(2\pi)^{N_W/2}\det(A)^{-1/2}. \tag{16}$$

Maximising Eq. (15) results in a closed form solution for the hyper-parameters, giving

$$\gamma_m = N_m - \alpha_m \mathrm{Tr}_m(A^{-1}) = \frac{N_m \sum_{n=1}^{N_m}(W_{mn}^{\mathrm{MP}})^2}{\sum_{n=1}^{N_m}(W_{mn}^{\mathrm{MP}})^2 + \mathrm{Tr}_m(A^{-1})}, \tag{17}$$

$$\frac{1}{\alpha_m} = \frac{\sum_{n=1}^{N_m}(W_{mn}^{\mathrm{MP}})^2}{\gamma_m} = \frac{\sum_{n=1}^{N_m}(W_{mn}^{\mathrm{MP}})^2 + \mathrm{Tr}_m(A^{-1})}{N_m}, \tag{18}$$

where $\mathrm{Tr}_m(A^{-1})$ is the trace of the inverse Hessian taken over the weights sharing α_m. This trace term is a measure of the uncertainty in the estimation of the weights [5], therefore, it follows from Eq. (17) that the intermediate parameter γ_m is positive and reaches its upper limit only when all of the N_m weights associated with α_m have zero error bars. The interpretation of γ_m is that it represents the number of well determined parameters in the group m of weights. There are separate (α_m, γ_m) hyper-parameters for each input covariate, each shared among multiple attributes; for the time covariate; for the bias terms b_h in the hidden units; for the weights w_h to the single output unit; and for the output node bias b. Taken together, the values of γ add-up to the number of effective parameters in the model, which is usually a fraction of the number of estimated weights.

Eq. (18) reflects the 'empirical Bayes' approach used by this framework, whereby the variance of the prior distribution of the weight, $1/\alpha_m$ is adjusted to match the sample variance of the estimated weights about the assumed mean of zero, averaged over the number of well determined parameters or, equivalently, calculated from the sample variance by adding the predicted variance for the weights.

The application of this methodology resulted in a large value for α_m in each variable, indicating that one of the attributes was redundant. This lends itself naturally to the use of a baseline population with all-zero covariates and the same attributes were chosen as for the baseline in the proportional hazards model.

4.3. Model selection with the neural network

Having completed the estimation of the PLANN parameters with ARD regularisation to soft-prune irrelevant variables in the model, the Bayesian framework can be utilized in full to carry out model selection. In our experience, careful identification of the explanatory variables is the single major determinant of the accuracy and generality of predictive models. Model selection requires a third level in the ARD methodology, beyond estimation of the evidence for the weight parameters and regularisation hyper-parameters, to estimate the evidence in support of a particular model hypothesis H, using

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}. \tag{19}$$

Assuming a flat prior for the space of possible models, considered here as the available set of explanatory variables, the evidence for a particular model selection is given by

$$P(H|D) \propto P(D|H) = \int P(D|\alpha, H)P(\alpha|H)\,\mathrm{d}\alpha, \tag{20}$$

which is the analytical expression for the normalizing constant in the evidence for the hyper-parameters, Eq. (13), just as its numerator required an estimation of the denominator in the evidence calculation for the weights, Eq. (9).

Assuming a log-normal distribution for the hyper-parameter priors $P(\alpha|H)$, the variance the variance of the distribution of $\log(\alpha_m) \sim 2/\gamma_m$ [10]. Approximating the integral in Eq. (20) by its mode multiplied by the width of the prior [5,10] we obtain an analytical approximation to the evidence in support of candidate PLANN model, given by

$$\log(P(D|H)) \approx -S(w^{\mathrm{MP}}, \alpha) - \frac{1}{2}\log(\det(A)) + \frac{1}{2}\sum_{m=1}^{N_\alpha}(N_m \log(\alpha_m)) + \frac{1}{2}\sum_{m=1}^{N_\alpha}\log\left(\frac{2}{\gamma_m}\right) + \frac{N_\alpha}{2}\log(2\pi) + \log(N_h!N_h^2). \tag{21}$$

Notice that the evidence is calculated using the most probable values of the network outputs, and it includes a combinatorial term to allow for the inherent symmetries in the neural network, whose hidden nodes may be permuted at will, and the sign of the weights connected to each hidden node can be reversed but with a change in the output node bias will lead to a functionally identical model. This term is expected to legitimise the unimodal approximation to the evidence made by this framework, even though the error surface of the neural network is necessarily multimodal.

In practice, the universal approximation properties of neural networks are only valid for sufficiently large numbers of hidden nodes. While, in principle, this could exacerbate the difficulties with overfitting, within a regularisation framework it is quite feasible to use a large network and thus keep the number of hidden nodes constant during model selection. This also has the advantage of preventing the model from falling into local minima in the function space, which can result in accurate fitting of the data but with unnecessarily complex surfaces [12].

4.4. Marginalisation of the network predictions

In survival modelling, the distribution of the binary indicator label, or target, is extremely skewed due to the scarcity of events and the large number of time steps used in the analysis. Skewed target distributions are improperly regularised using the standard Netlab software because equal numbers of ‘zeros’ and ‘ones’ are assumed.

In particular, the posterior distribution for the network parameters in the Bayesian framework requires a modulation of the network outputs towards what is called the guessing line in the Receiver Operating Characteristic framework, which corresponds to assigning the output to the prevalence [3] when the weights have large error bars.

The predicted hazard is the mean calculated from the distribution of the activation $a(\cdot)$ which is the argument of the exponential in right-hand side of Eq. (5), $h(x,t)\equiv g(a)$, giving

$$h_g(x,t) = \int g(a)P(a|x_p,t,D)\,\mathrm{d}a. \tag{22}$$

In this expression, D represents the target distribution contained in the training data.

This integral is not analytical, but it can be evaluated when the activation is expanded as a linear function of the weights

$$a^*(x_p,t,w) \approx a^{\mathrm{MP}}(x_p,t,w) + g^{\mathrm{T}}(x_p,t,w^{\mathrm{MP}})\,(w-w^{\mathrm{MP}}). \tag{23}$$

The distribution of the activation is found by integrating over the posterior distribution for the weights, Eq. (12), using the Taylor expansion of the objective function, Eq. (14) variance s^2 [10], resulting in

$$P(a|x_p,t,D) \propto \exp\left(-\frac{(a-a^{\mathrm{MP}})^2}{2g^{\mathrm{T}}A^{-1}g}\right). \tag{24}$$

Having obtained the variance of the activation values, the predicted hazards is now well approximated [5,10] by

$$h_g(x,t) \approx g\left(\frac{a^{\mathrm{MP}}(x,t))}{\sqrt{1+(\pi/8)g^{\mathrm{T}}A^{-1}g}}\right). \tag{25}$$

Therefore, the inherent uncertainty in the network prediction may be described either by specifying a range of values for the integrand in Eq. (21) or by marginalising over the activation and thus moderating the network output towards the guessing line. However, $h_g(\cdot) \rightarrow_{s\rightarrow\infty} 1/2$, therefore, the prevalence of the binary targets assumed in Eq. (25) is 50%.

It is now straightforward to update the standard regularisation framework to take account of the prevalence $P_\tau = P(\tau_{pk}=1)$ by re-scaling the log-likelihood, together with the calculation of the gradient and the Hessian, as follows [13]

$$\tilde{G} = -\sum_{p=1}^{\text{no. of patients}}\sum_{k=1}^{t_1}\left[\frac{1}{2P_\tau}\tau_{pk}\log(h_p(x_p,t_k)) + \frac{1}{2(1-P_\tau)}(1-\tau_{pk})\log(1-h_p(x_p,t_k))\right], \tag{26}$$

and compensating the resulting marginalised network prediction $\tilde{h}_g(x,t)$

$$h_g(x_p,t) = \frac{\tilde{h}_g(x_p,t)P_\tau}{\tilde{h}_g(x_p,t)P_\tau + (1-\tilde{h}_g(x_p,t))(1-P_\tau)}. \tag{27}$$

5. Application to breast cancer prognosis

An important clinical application of survival models is the allocation of patients into prognostic risk groups, as this directly influences the choice of treatment. In this study the overall data are treated as two cohorts separated using standard TNM criteria, as described in Section 2.

The proposed method to identify and interpret prognostic risk groups consists of the following stages:

(a) Model selection to identify a predictive model for the hazard.
(b) Assignment of a prognostic group index, or risk score, to each individual record. The aggregation of the risk scores into distinct prognostic risk groups is then carried out in a pairwise manner using the log-rank test [14].
(c) The evaluation of the model starts by plotting the observed survival using Kaplan–Meier (KM) curves [15] for the grouped data, whose error bars [6] should show little overlap. The predictive accuracy of the survival models is then gauged by comparing the KM curves with the mean grouped survival curves calculated from the predicted hazards, for each month over the 5-year-period of follow-up.
(d) Having established the predictive accuracy of the survival models, it is necessary to interpret the composition of the prognostic groups, by profiling the histograms of selected variables.

In this study, model selection was carried out with a time step of 1 year, to reduce the computational overhead. Once a preferred model was established, the predictive modelling then used a time step of one month. All of the results presented in this section were obtained by cross-validation so that, for each individual, they represent an out-of-sample prediction. The low-risk cohort was predicted using five-folds and the high-risk cohort with three-folds.

5.1. Results with the proportional hazards model

The proportional hazards model was implemented in SPSS [9] with stepwise model selection using the Akaike Information Criterion (AIC) [6]. This procedure begins by searching for the most significant univariate model. At any stage in the selection process, each remaining variable is added to the model and the most significant multivariate model is selected. Each variable is then dropped from the model, in turn, to test whether there is evidence that any of the existing variables has become redundant. The procedure then continues by iteration, using a significance measure based on

$$\mathrm{AIC} = -2\log(\hat{L}) + \alpha N_{\beta}, \tag{28}$$

where $\hat{L}$ is the optimised log-likelihood function for the proportional hazards model, N_{β} the number of degrees of freedom which is the same as the number of attributes excluding the those for the baseline population and α is a predetermined constant that took the recommended value of 3 which is roughly equivalent to a 5% significance level to distinguish between nested models with few variables [6]. This parameter may be adjusted upwards if the resulting models are judged to generalise poorly.

The final model for the low-risk cohort comprised four variables with a total of eight degrees of freedom, namely, *histology*, *pathological size*, *clinical stage nodes* and *nodes ratio*.

Four risk groups were identified in from the prognostic index $\beta^{\mathrm{T}}x_p$, shown in Fig. 1a.

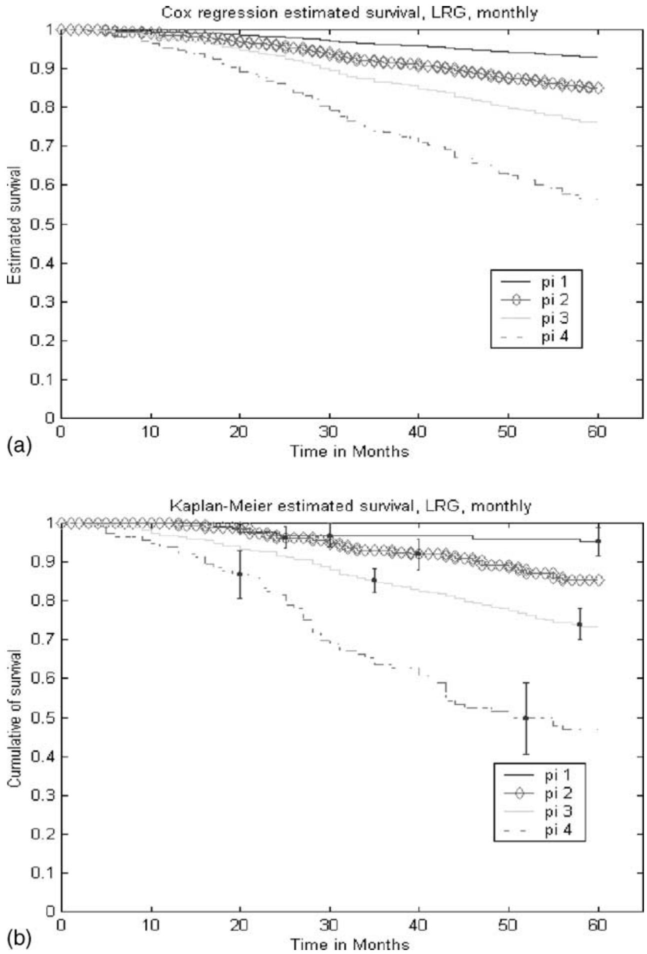

Fig. 1. (a) Partitioning of the low-risk group (LRG) into four prognostic groups with 127, 189, 487 and 114 patients, respectively; (b) mean survival predicted for each prognostic risk group in the low-risk cohort, by the proportional hazards model; (c) the corresponding grouped Kaplan–Meier curves.

There is a good match between predicted and observed survivorship shown in Fig. 1b and c. The resulting cluster profiles, in Fig. 2, show a clear clinical pattern. As the risk increases, *histology* moves away from attribute 3 (mixed medullary, etc.) to attributes 1 (infiltrating ductal carcinoma) and 2 (infiltrating lobular carcinoma) and *pathological size* gradually shifts from 1 (<2 cm) to 2 (2–5 cm). The group at highest risk is characterised by *clinical stage N* of 1 (ipsilateral and mobile axillary nodes) from 0 (no affected nodes found). The frequency histogram of *nodes ratio* is less differentiated across the groups.

By reference, a prospective study of the Nottingham Predictive Index [1], which has a prognostic index scale based on tumour size, lymph node stage assigned from a triple node biopsy, and three levels of histological grade based on tissue differentiation, identified three

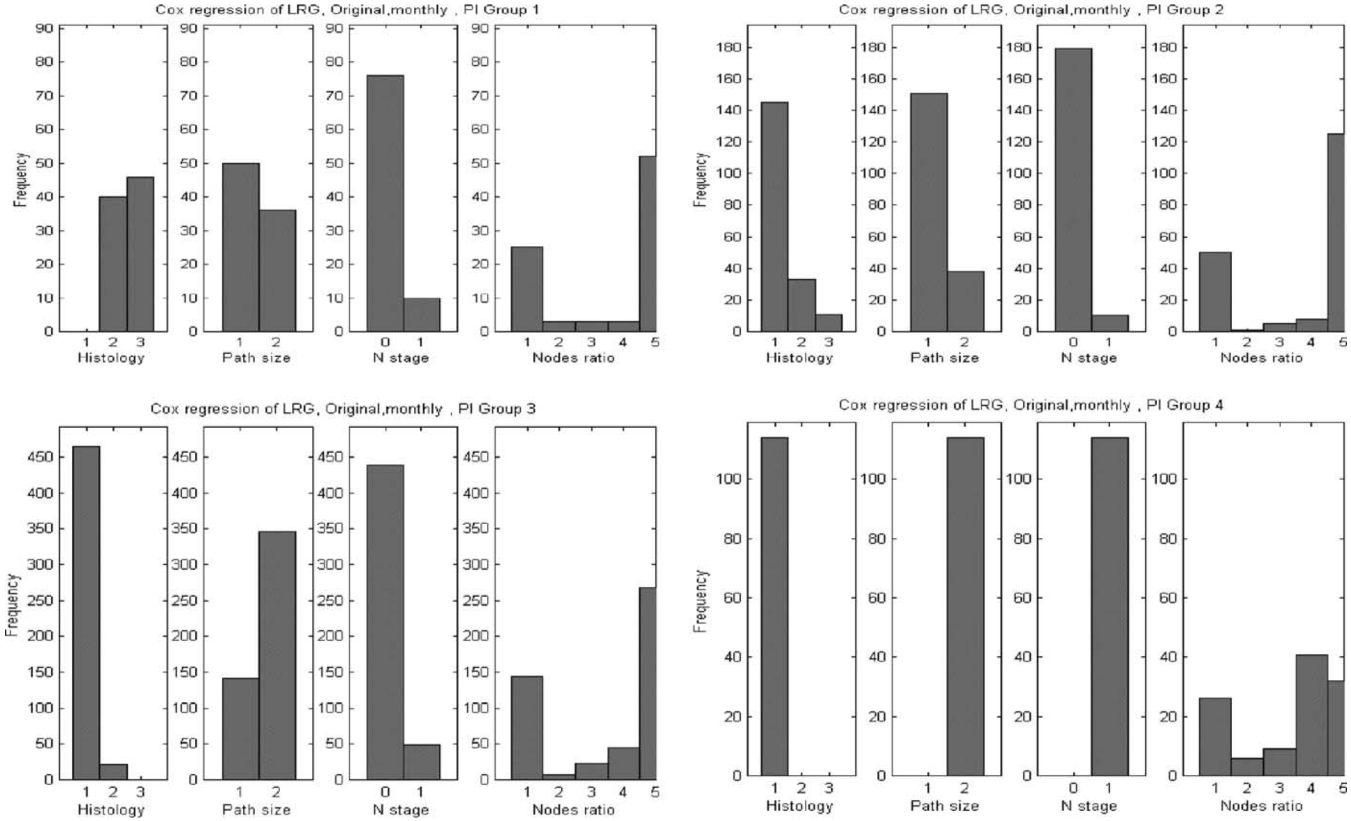

Fig. 2. Attributes profiles for the four risk groups identified in the low-risk cohort by the proportional hazards model.

prognostic groups, two of which overlap with the survival range in Fig. 1. An exact comparison between the two models is not possible because they use different clinical indicators but, nevertheless, these results suggest that *nodes ratio* may be redundant and there too many prognostic groups were identified by the default model selection procedure, even using AIC. A more detailed analysis of the proportional hazards model including time effects and covariate interactions is possible, but it is outside the scope of this paper since the neural network modelling will also be carried out using a default procedure.

The selected variables for the high-risk group are *menopausal status*, *node stage*, *pathological size*, *clinical stage* and *nodes ratio*. Three prognostic groups were identified, with substantial overlap with the survival range in the low-risk group, in Fig. 3a and b.

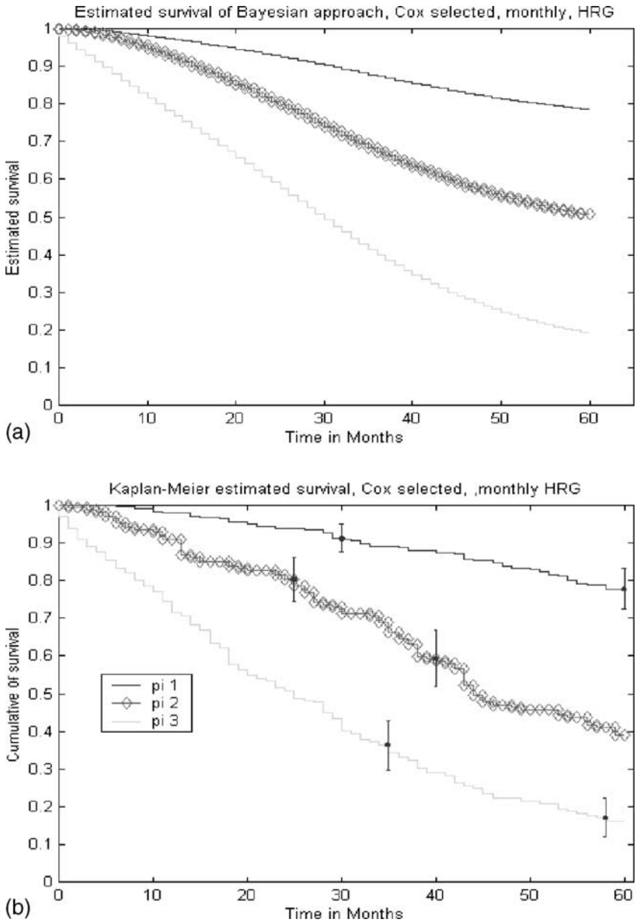

Fig. 3. (a) Mean survival predicted by the proportional hazards model for the high-risk group; (b) the corresponding grouped Kaplan–Meier curves.

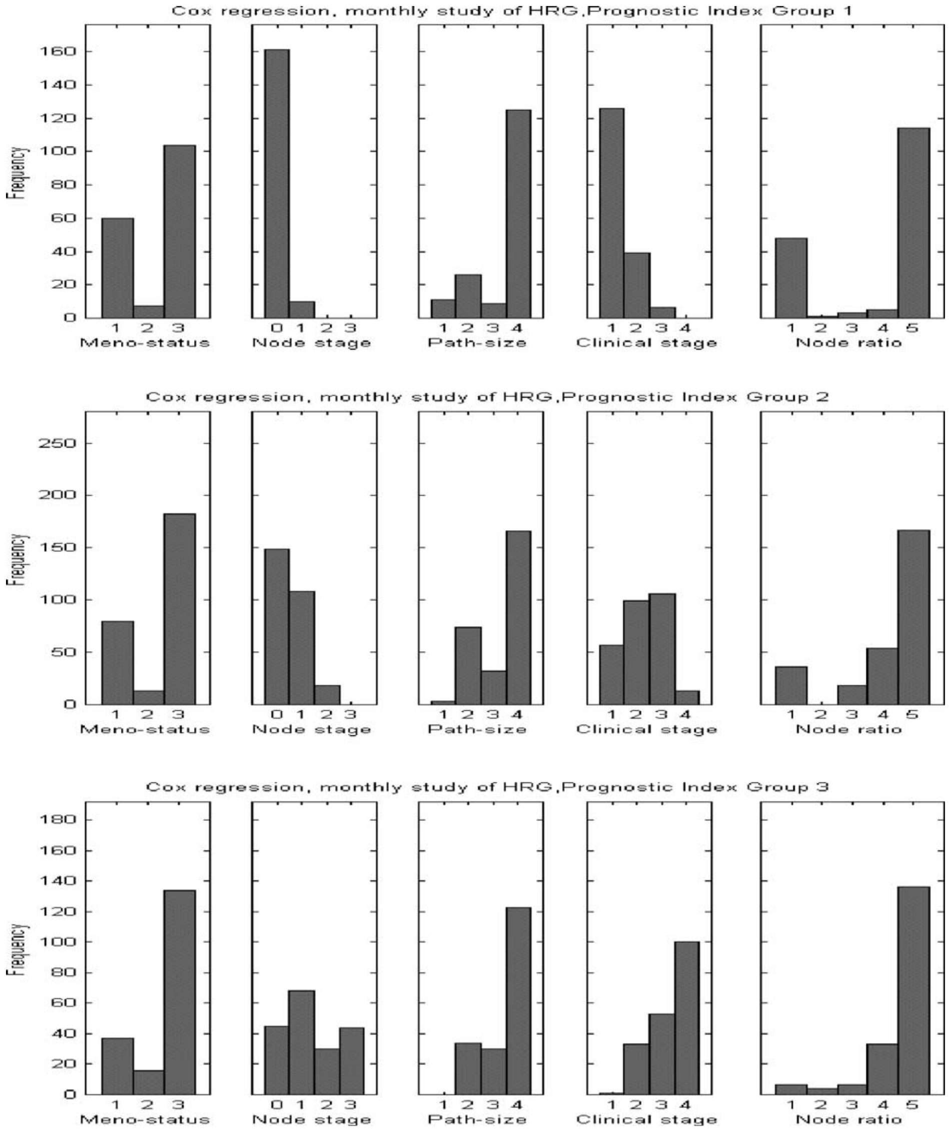

Fig. 4. Attribute histograms for the prognostic risk groups identified in the high-risk cohort by the proportional hazards model.

This is due, at least in part, to the automatic assignment of low-risk patients with missing values of *pathological size* to the high-risk group, indexed by clinical stage 1 (Fig. 4).

In the attribute profiles shown in Fig. 4, variables *node stage* and *clinical stage* show the clearer patterns with increasing risk.

5.2. Contrast with the neural network

A comparison of the defining equations for the discrete time proportional hazards mode, Eq. (2) and the PLANN model, Eq. (5), indicates that the equivalent of the linear prognostic index $\beta^{\mathrm{T}} x_p$ is the non-linear index comprising the argument of the logistic function, namely, the activation $a(x_p, t)$ in Eq. (6). This is readily calculated from the hazard by inverting the logistic sigmoid,

$$a(x_p, t_k) = \log\left(\frac{h_p(x_p, t_k)}{1 - h_p(x_p, t_k)}\right), \tag{29}$$

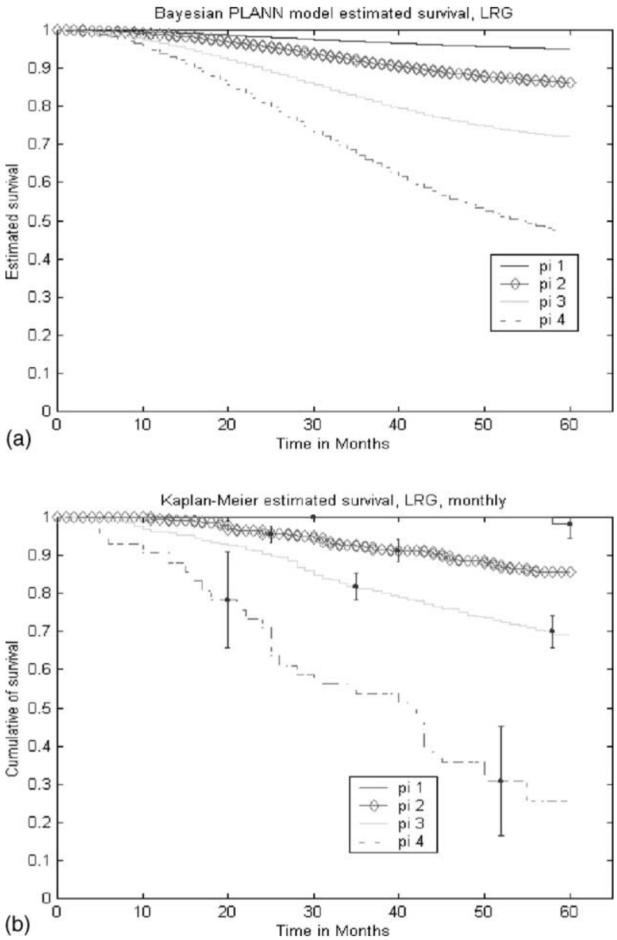

Fig. 5. (a) Partition of the prognostic index predicted by the PLANN–ARD model for the low-risk cohort, into risk groups with 56, 359, 460 and 42 patients, respectively; (b) mean survivorship predicted by the PLANN–ARD model with the same variables as used by Cox regression; (c) the corresponding grouped Kaplan–Meier curves.

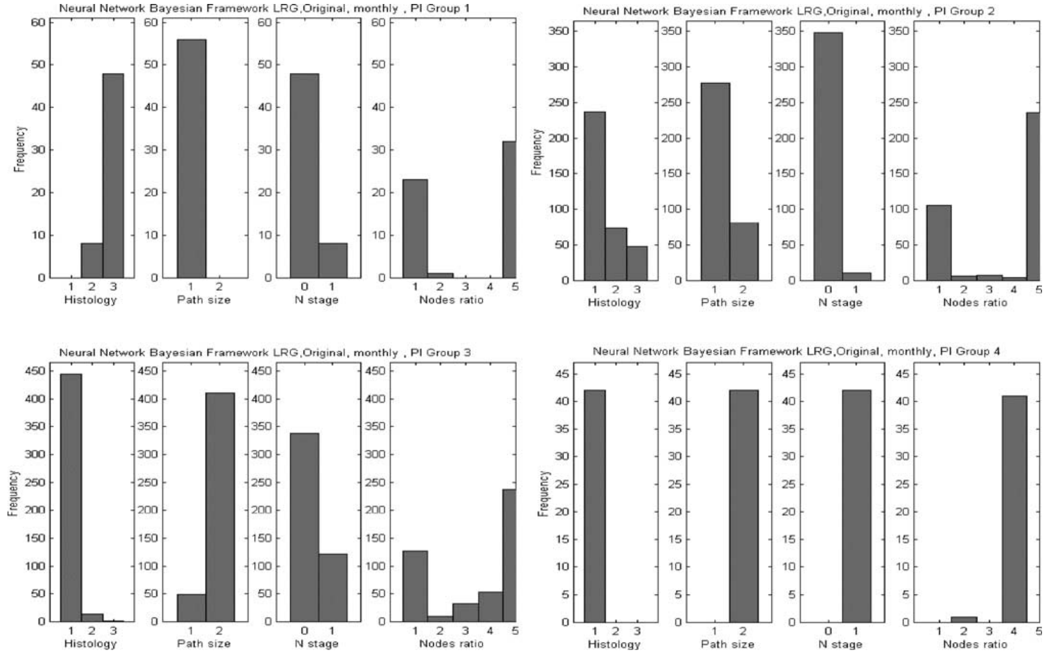

Fig. 6. Variable attributes of four prognostic groups identified by PLANN–ARD.

which is just the log-odds-ratio of the predicted hazard. However, the hazard varies both over time and with the covariates. Therefore, the simplest prognostic index for PLANN is obtained by averaging the predicted activation of the 60 months' duration of the study:

$$\mathrm{PI}(x_p) = \frac{\sum_{k=1}^{60} a(x_p, t_k)}{60}. \tag{30}$$

The PLANN–ARD model was first applied using the variables selected by Cox regression, to verify that it was capable of matching the predictions of a model that is linear-in-the-parameters, albeit actually a non-linear model of the attributes. In effect, we are testing implicitly for interactions between the covariates, or with time. Fig. 5a shows that four risk groups were again identified, with good agreement between the predicted and observed survival, in Fig. 5b and c. All PLANN networks had eight hidden nodes, as this had been deemed sufficient in cross-validation tests, and the regularisation framework does not require further pruning.

Note from Fig. 5c that in the highest surviving group there was only one event observed over the period of follow-up. Fig. 6 shows that *pathological size* <2 cm is a defining characteristic in this prognostic group. Prognostic group 4 appears to have been uniquely identified by a nodes ratio >60%.

These results indicate that PLANN–ARD matches the predictions made by Cox regression for the same explanatory variables, resulting in greater specificity in the assignment of patients into risk group. This effect indicates that some amount of interaction is present among the covariates, which the neural network is capable of utilising in its prognostic assignments.

In PLANN the hazard may be plotted directly as a function time, as shown in Fig. 7.

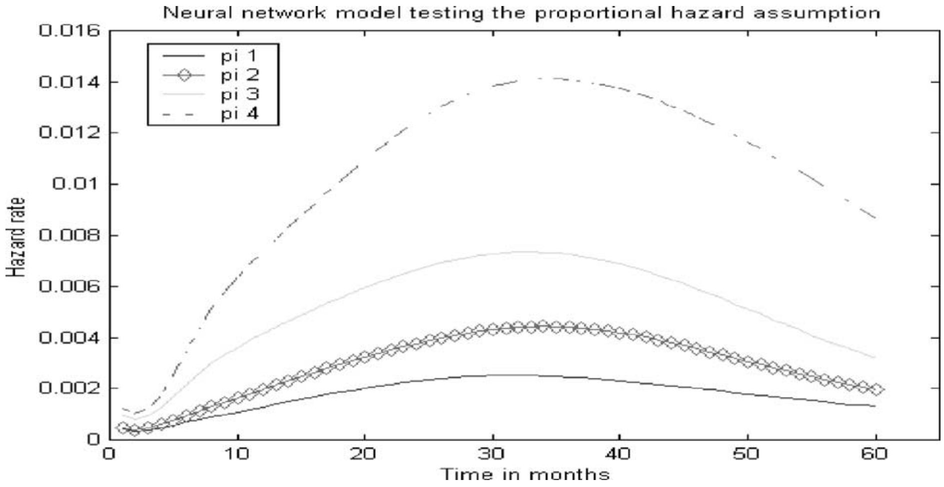

Fig. 7. Predicted hazard for the prognostic groups identified in the preceding figures. The proportionality of the hazards is largely satisfied, with only a slight delay in the peak hazard as the mortality risk increases.

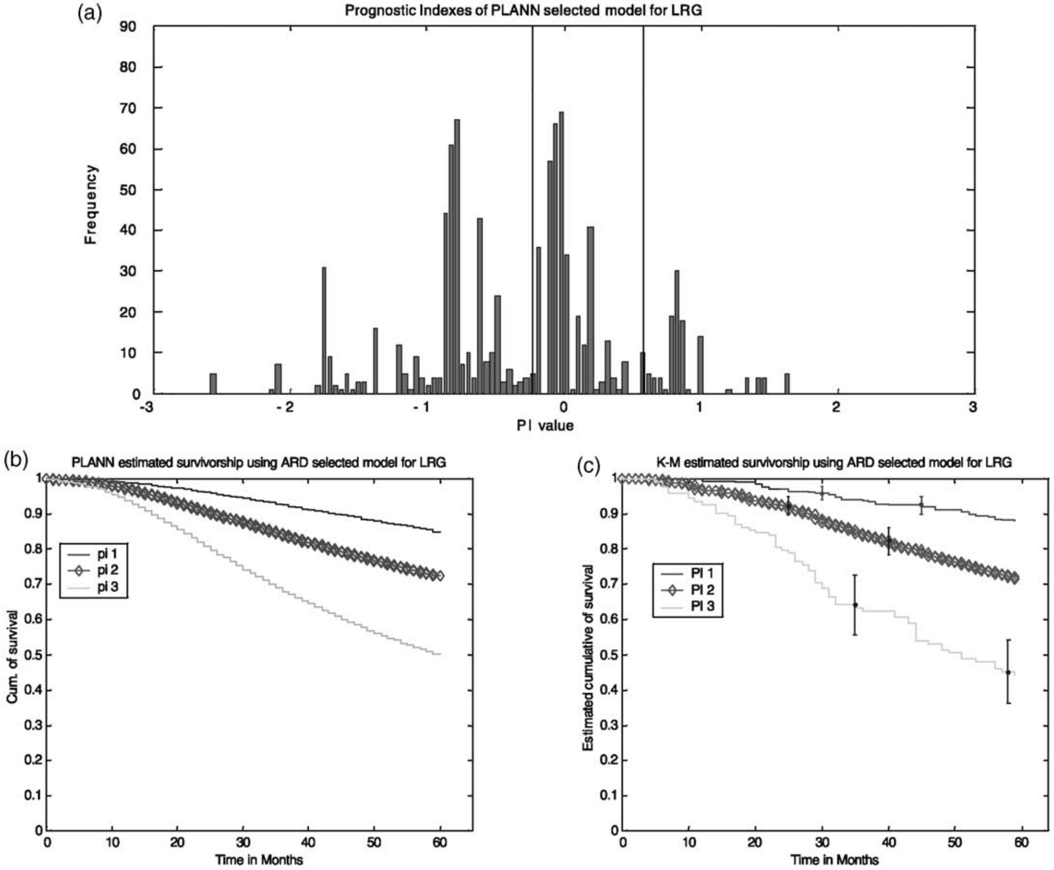

Fig. 8. (a) Allocation of prognostic groups using the log-rank test applied to the prognostic index of PLANN–ARD for the low-risk cohort; (b) mean survival curves for the risk groups identified in (a) with 423, 370 and 124 cases, respectively; (c) corresponding grouped Kaplan–Meier curves.

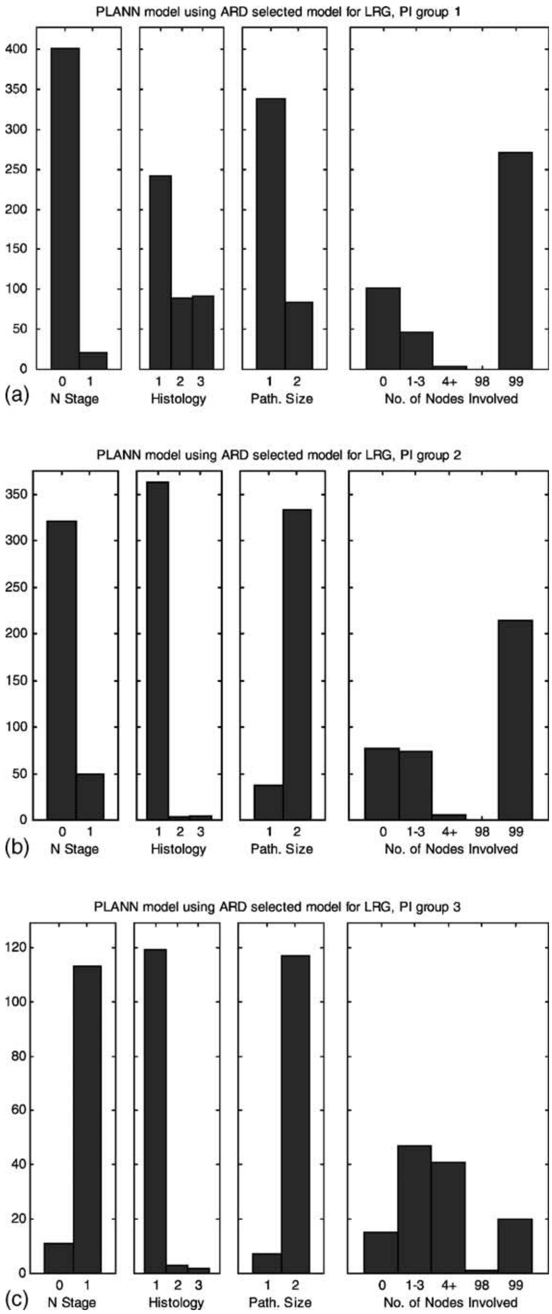

Fig. 9. Attribute histograms for the prognostic groups identified with PLANN–ARD in the low-risk group.

Table 2
Evidence calculation with PLANN–ARD for the low-risk cohort

Variables added to the model	Evidence estimated by ARD	Log-likelihood estimated by cross-validation
Clinical stage nodes	2529.1	1418.6
Histology	2501.2	1421.8
Pathological size	2484.6	1425.4
Number of nodes involved	2470.3	1498.9
Tumour diameter	2520.5	–

Table 3
Evidence calculation with PLANN–ARD for the high-risk cohort

Variables added to the model	Evidence estimated by ARD	Log-likelihood estimated by cross-validation
Clinical stage (TNM)	1347.6	1749.2
Clinical stage tumour	1379.4	1791.8
Clinical stage metastasis	1428.4	1793.7

5.3. Results with PLANN–ARD

Model selection for the neural network was carried out using the Bayesian regularisation framework outlined in Section 4. The model for the low-risk cohort selected four variables, as indicated in Tables 2 and 3. Notice that the evidence has reaches a minimum but this

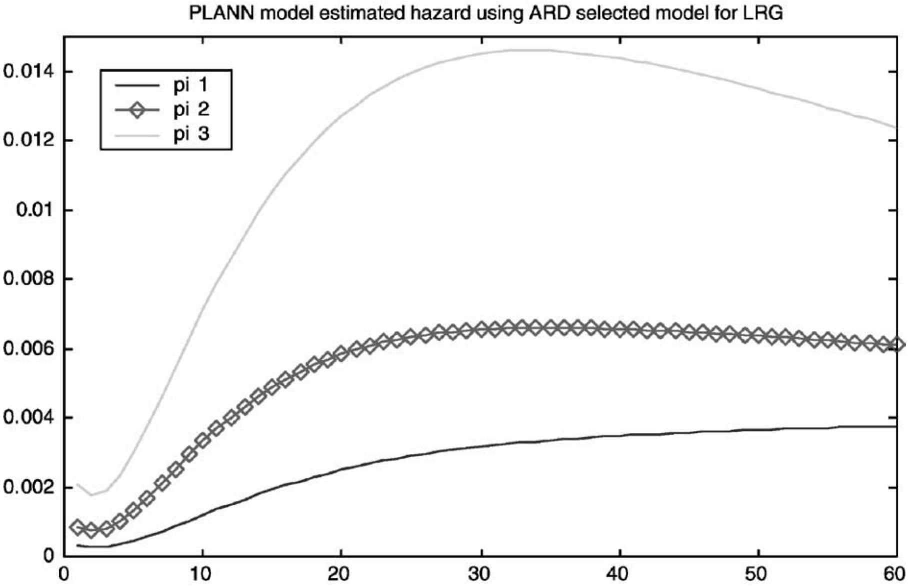

Fig. 10. Mean monthly hazard predicted by PLANN–ARD for the data in the preceding figures.

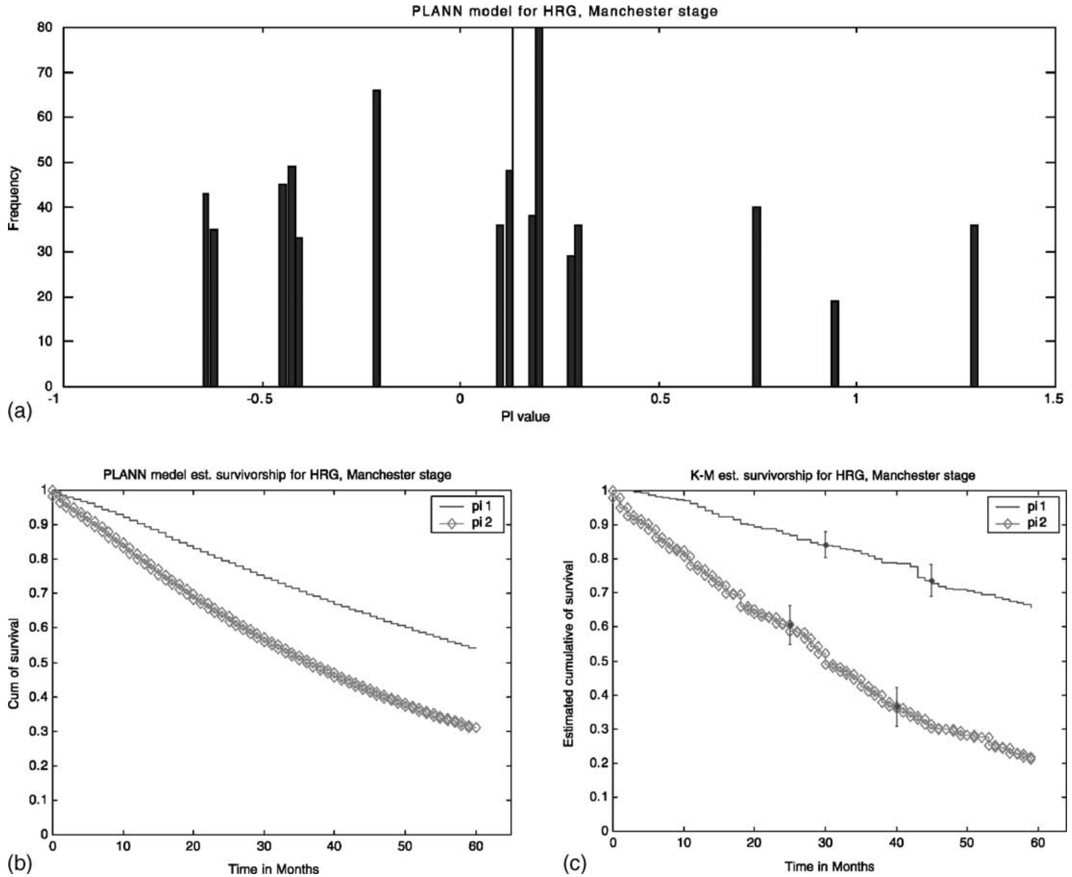

Fig. 11. (a) Allocation of prognostic groups using the log-rank test applied to the prognostic index of PLANN–ARD for the high-risk cohort. Only two groups are identified; (b) mean survival curves for the risk groups identified in (a) with 355 and 278 cases, respectively; (c) corresponding grouped Kaplan–Meier curves.

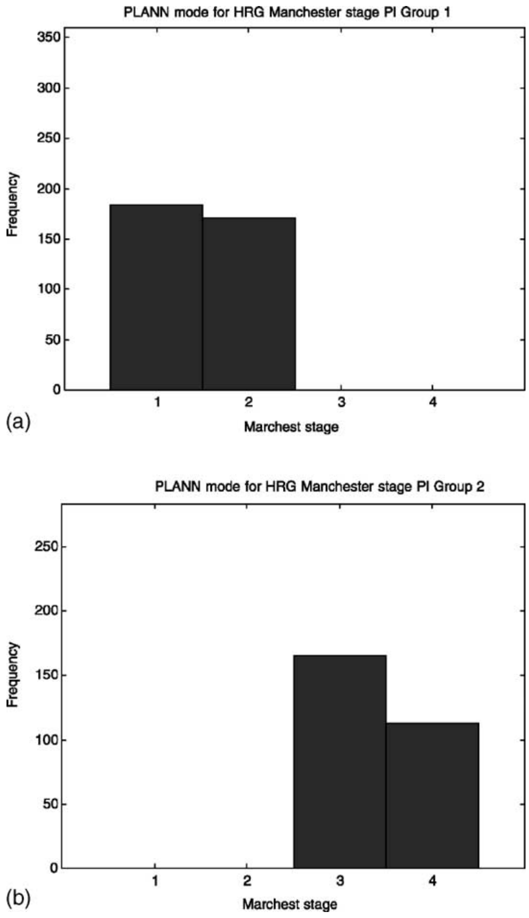

Fig. 12. Attribute histograms for the prognostic groups identified with PLANN–ARD as a univariate model for the high-risk group, defined by TNM clinical staging.

is not replicated in another measure of generality, which is the log-likelihood estimated out-of-sample.

The variables selected by the model are vary similar to those in the proportional hazards model, with the exception of *number of nodes involved* replacing *nodes ratio*. However, the prognostic group allocation is more conservative than found previously, with patients assigned to only three risk groups, shown in Fig. 8.

The attributes profiles in Fig. 9 clearly show that *pathological size* and *histology* are the key factors to differentiate the highest surviving-group from the rest, and *clinical stage nodes* specifies the lowest-surviving group. Note that nearly all cases with *lobular carcinoma*, labelled by histological attribute 2, have been assigned into a single group, making the risk assignment specific to this variable. The predicted hazard, in Fig. 10, is also different from the previous results, suggesting a later peak in the hazard for the lowest risk group.

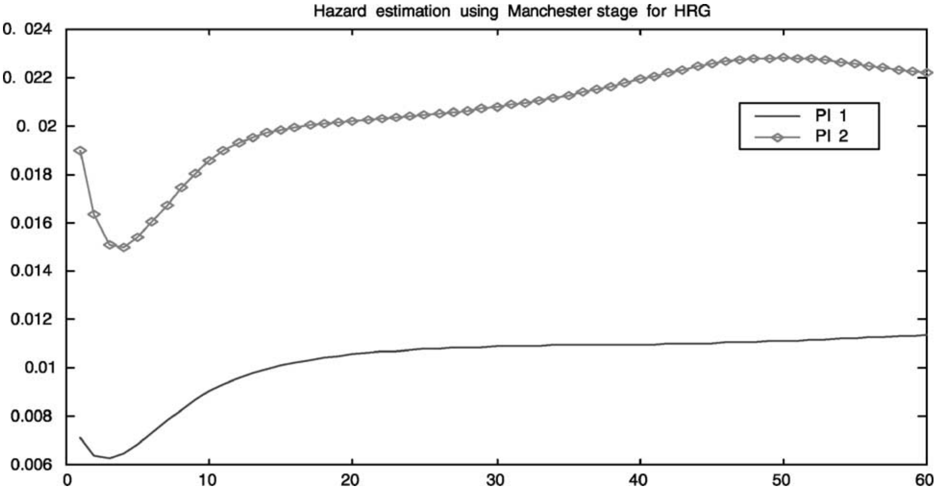

Fig. 13. Mean monthly hazard predicted by PLANN–ARD for the high-risk group.

Model selection for the high-risk cohort resulted in a univariate model, rediscovering current clinical practice for these patients, that is to rely entirely on TNM staging. This is clear from Table 3 where two indicators of model performance are listed for increasing numbers of variables. The next best variables are also clinically relevant, but would seem not to add significantly to the prognostic allocation, which then diverges into four prognostic groups as was the case with the proportional hazards model applied to the same data.

Figs. 11 and 12 show the prognostic index, risk group assignment and attribute profiles, respectively, while the continuous prediction of the hazard is plotted in Fig. 13.

The grouping of TNM staging is exactly as expected and the predicted monthly hazard is higher than for the low-risk cohort, but nearly flat which is also as expected from the nearly exponential structure of the observed survivorship of groups of high-risk patients. This effect is sufficiently strong to identify the time covariate as not always relevant in PLANN–ARD modelling of this group of patients. Nevertheless, the time covariate was forced into the model to explicitly show the dependence on time (Fig. 13).

6. Conclusion

A Bayesian framework with covariate-specific regularisation has been introduced as an extension of the PLANN neural network model for censored data, to carry model selection using Automatic Relevance Determination.

The results of contrasting the PLANN–ARD model with the clinically well accepted proportional hazards model were that the two are consistent, but the neural network may be more specific in the allocation of patients into prognostic groups using a default procedure that is also described in this paper.

With automatic model selection, however, we obtain the perhaps surprising result that the regularised neural network is more conservative than the default stepwise forward selection procedure implemented by SPSS with the Akaike Information Criterion. While this is not entirely unexpected due to the limited amount of experimentation with this model, in particular with regard to the choice of penalty parameter, it is nevertheless encouraging that the PLANN–ARD model immediately produced results that make clinical sense and appear to be more robust than those with the classical statistical model.

Future work will involve the application of these models to a second cohort of patients recruited after those modelled here.

References

[1] Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham Prognostic Index in primary breast cancer. Breast Cancer Res Treat 1992;22:207–19.

[2] Cox DR. Regression models and life tables. J R Stat Soc, B 1972;74:187–220.

[3] Lisboa PJG. A review of evidence of health benefit from artificial neural networks in medical intervention. Neural Networks 2002;15(1):9–37.

[4] Biganzoli E, Boracchi P, Mariani L, Marubini E. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. Stat Med 1998;17:1169–86.

[5] MacKay DJC. Probable networks and plausible predictions—a review of practical Bayesian methods for supervised neural networks. Network: Comput Neural Syst 1995;6:469–505.

[6] Collett D. Modelling survival data in medical research. London: Chapman & Hall; 1994.

[7] Ravdin PM, Clark FM, Hilsenbeck SG, Owens MA, Vendely P, Pandian MR, et al. A demonstration that breast cancer recurrence can be predicted by neural network analysis. Breast Cancer Res Treatment 1992;21:47–53.

[8] Christensen E. Multivariate survival analysis using Cox's regression model. Hepatology 1987;7(6): 1346–987.

[9] SPSS Base 9.0/user's guide. SPSS Inc.; 1999.

[10] Bishop CM. Neural network for pattern recognition. Oxford: Clarendon Press; 1995.

[11] Nabney I. NETLAB: algorithms for pattern recognition. London: Springer; 2001.

[12] Lisboa PJG, Etchells TA, Pountney DC. Minimal MLPs do not model the XOR logic. Neurocomputing 2002;48:1033–7.

[13] Lisboa PJG, Vellido A, Wong H. Bias reduction in skewed binary classification with Bayesian neural networks. Neural Networks 2000;13:407–10.

[14] Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep 1966;50:163–70.

[15] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457–81.