Available online at www.sciencedirect.com
ScienceDirect
Neural Networks 21 (2008) 414–426
Neural Networks
www.elsevier.com/locate/neunet

2008 Special Issue

Time-to-event analysis with artificial neural networks: An integrated analytical and rule-based study for breast cancer☆

Paulo J.G. Lisboa[a], Terence A. Etchells[a,*], Ian H. Jarman[a], M.S. Hane Aung[a], Sylvie Chabaud[b], Thomas Bachelot[b], David Perol[b], Thérèse Gargi[b], Valérie Bourdès[c], Stéphane Bonnevay[d], Sylvie Négrier[b]

[a] *School of Computing and Mathematical Sciences, Liverpool John Moores University, UK*
[b] *Centre Léon Bérard, 28 rue Laennec, 69 373 Lyons Cedex 08, France*
[c] *THEMIS-ICTA Group, 60 avenue Rockefeller, 69008 Lyons, France*
[d] *Claude Bernard University Lyon 1, 43 Bd du 11 novembre 1918, 69622 Villeurbanne Cedex, France*

Received 8 August 2007; received in revised form 29 November 2007; accepted 13 December 2007

Abstract

This paper presents an analysis of censored survival data for breast cancer specific mortality and disease-free survival. There are three stages to the process, namely time-to-event modelling, risk stratification by predicted outcome and model interpretation using rule extraction. Model selection was carried out using the benchmark linear model, Cox regression but risk staging was derived with Cox regression and with Partial Logistic Regression Artificial Neural Networks regularised with Automatic Relevance Determination (PLANN-ARD). This analysis compares the two approaches showing the benefit of using the neural network framework especially for patients at high risk. The neural network model also has results in a smooth model of the hazard without the need for limiting assumptions of proportionality. The model predictions were verified using out-of-sample testing with the mortality model also compared with two other prognostic models called TNG and the NPI rule model. Further verification was carried out by comparing marginal estimates of the predicted and actual cumulative hazards. It was also observed that doctors seem to treat mortality and disease-free models as equivalent, so a further analysis was performed to observe if this was the case. The analysis was extended with automatic rule generation using Orthogonal Search Rule Extraction (OSRE). This methodology translates analytical risk scores into the language of the clinical domain, enabling direct validation of the operation of the Cox or neural network model. This paper extends the existing OSRE methodology to data sets that include continuous-valued variables.

Keywords: Time-to-event analysis; Rule extraction; TNG; NPI; PLANNARD

1. Introduction

This paper presents a longitudinal cohort study of time-to-event for 2535 consecutive patients with primary operable breast cancer, recruited prospectively at Centre Léon Berard (Lyons-France), between 1996 and 2004, with 10 years of follow-up. It is a baseline study because all of the covariates are measured only once, clinical variables recorded at the point of diagnosis, and histological values ascertained immediately following surgery. The purpose of the study is to compare neural network modelling with a widely used statistical methodology that is known to apply to breast cancer studies on this timescale, namely Cox regression, also known as proportional hazards modelling. There are two main events of interest, namely mortality and treatment failure. The follow-up period is discretised by month. The study demonstrates the application of a fully regularised generic non-linear model of covariate effects and time, the PLANN-ARD methodology, to competing risks. This is compared with Cox regression within a framework of risk stratification, similar to that used to derived the well-known severity of illness score for breast cancer known as the Nottingham Prognostic Index (NPI) (Galea, Blamey, Elston, & Ellis, 1992; Haybittle et al., 1982). A further

☆ An abbreviated version of some portions of this article appeared in Lisboa, Etechells, Jarman, Aung, and Perol (2007) as part of the IJCNN 2007 Conference Proceedings, published under IEE copyright.

* Corresponding author. Tel.: +44 1204 811423.
E-mail addresses: p.j.lisboa@ljmu.ac.uk (P.J.G. Lisboa), t.a.etchells@ljmu.ac.uk (T.A. Etchells).

0893-6080/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2007.12.034

comparison is made between the newly derived mortality model and two rule-based models obtained using Orthogonal Search Rule Extraction (Etchells & Lisboa, 2006), one called TNG staging (Jarman, Etchells, Ellis, Green, & Lisboa, 2007) derived from a non-linear model analogous to NPI and the other the NPI rule-based model (Jarman et al., 2007) derived from the NPI. With these risk models it was observed that doctors seem not to distinguish between disease-free and mortality models, therefore this was tested by investigating whether there was any difference in the observed survival when the target for the mortality model was changed to disease-free survival.

While analytical modelling has the capability to show differential mortality between patient groups, which can be evaluated by out-of-sample predictions, the scores derived from the neural network are not conveniently linear, as in the case of Cox regression. This has given rise to concerns about the transparency of the non-linear approach, which is central to clinical verification of the model using accepted clinical expertise. This was done by expressing the risk allocation in terms of low-order Boolean rules that permit a direct interpretation of the composition of each risk group. Moreover, replacing the neural network by the rule set for each of the three prognostic groups retains much of the discriminatory power of the original non-linear model, but now pertaining to an entirely white-box decision support system. The methodology used to extract the rules, Orthogonal Search Rule Extraction (OSRE) (Etchells & Lisboa, 2006), originally applied to data sets with binary, categorical or ordinal-valued variables. In this paper the OSRE methodology is extended to data sets that have continuous-valued variables.

2. Data description

The cohort comprises a prospectively collected case series of consecutive patients with primary operable breast cancer, defined as clinical stages T0-4, N0-1 and M0. Inclusion criteria for this study specify first diagnosis only, thus removing recurrences, and rejects occurrences of bilateral carcinoma. For the purpose of time-to-event modelling, the date of recruitment was that of diagnosis.

All patients were diagnosed with an infiltrating tumour, confirmed by histological analysis. Initial surgical excision was carried out at Centre Léon Berard and was either localised, i.e. a lumpectomy, or radical, i.e. a mastectomy. In all cases, surgery included axillary clearance as part of the study protocol. Patients received standard treatment which could include adjuvant therapy, whether endocrine, chemo or both, and radiotherapy when appropriate.

The patient cohort was divided longitudinally into a modelling data set with patients recruited between 1996 and 2000 ($N = 1156$) and an out-of-sample test data set with patients recruited during 2001–2004 ($N = 1379$). While the longitudinal time best reflects the potential clinical use of the model for risk staging of future patients, it necessarily curtails the extent of follow-up in the out-of-sample data, to five years. Therefore cross-validation was used internally to validate the results obtained on the modelling data set.

The events of interest were breast cancer specific mortality, with intercurrent death was treated as censorship, and Disease-Free Survival (DFS), which is often taken as a better indicator of the impact on the patient's quality of life arising from the disease and the effects of therapy. Neither measure of outcome is entirely accurate, since death attributed to breast cancer may incorrectly include, or exclude, related deaths such as heart attacks that may have been induced by the load on the body due to chemotherapy, for instance, or even by carcinoid heart disease linked to metastatic spread to the liver. The time of recurrence is also necessarily uncertain, as there is a latency before clinical symptoms occur, with the consequence that the date of recurrence may even coincide with the date of death in cases where the recurrence was found only by post-mortem examinations.

Two variables had missing values in double figures, namely progesterone receptor count (92 occurrences) and pathological tumour size (83 occurrences). The former is a categorical variable. Given past experience in medical data sets where the distribution of missing values is not always at random, a separate attribute was created to denote missing. In the case of tumour size, its distribution is positive definite and thus noticeably skewed. For this reason, it was decided to impute missing values using the median, rather than the mean, of observations.

3. Analysis methodology

Censored data modelling in clinical sciences is predominantly carried out using Cox regression, which is termed proportional hazards model for continuous time on account of the factorisation of the hazard distribution separating out the covariate dependent from the time dependence, which is fitted to a baseline population chosen by the user. The models in this report select the baseline population as consisting of nil values for tumour size (the only continuous variable) and the most prevalent attribute, for categorical variables. This choice maximises the sample size of the reference population. An alternative choice would be to take the attribute with the best outcome.

The primary purpose of Cox regression is to study the relative influence on outcome from the explanatory variables, rather than to make survival predictions for individuals, or indeed groups. In this sense, the choice of baseline distribution should, in principle, be immaterial as the relative effects of different covariates remain approximately stable for different choices. This interpretation of the purpose of Cox regression also justifies its use as a linear-in-the-parameters risk staging index. That is the approach used in this study, where the risk score is extended to a non-linear model.

It is well know that in discrete time, Cox regression reverts to a strictly proportional odds model for the hazard. This lends the model naturally extensible for generic non-linear analysis of covariate effects and time dependency as a Partial Logistic Neural Network (PLANN) (Biganzoli, Boracchi, Mariani, & Marubini, 1998).

In common with all generic non-linear models, whether based on kernel functions (Boracchi, Biganzoli, & Marubini,

2001), splines (Boracchi, Biganzoli, & Marubini, 2003) or the Multi-Layer Perceptron neural network (Lisboa, 2002), PLANN must be appropriately regularised to prevent overfitting the data. While out-of-sample verification of the model inferences can be carried out by cross-validation and longitudinal testing, neural network predictions can be stabilised, for instance, by regularisation using weight decay. However, this requires tuning the regularisation hyper-parameter, typically by cross-validation.

An alternative, principled approach to regularisation, is to adjust the hyper-parameters within a Bayesian framework. The calculation of the evidence is non-analytical, but an effective approximation is possible, which has been applied to the Partial Logistic ANN resulting in an implementation with Automatic Relevance Determination (PLANN-ARD) (Lisboa, Wong, Harris, & Swindell, 2003). The term denotes the allocation of separate regularisation, or smoothing parameters, to each input covariate.

The application of the Bayesian regularisation framework with an analytical approximation of the evidence requires a re-weighting of the log-likelihood function to balance the data, so that there is apparently an equal number of observations for each model outcome (i.e. for risk vs. survival in the single risk model, and across all risks *and* survival, for competing risks). This is because the uncertainty in the calculation of the model estimates is reflected in a shift in model output towards a prior distribution where all outcomes are equally likely. The empirically derived priors are reintroduced by appropriately weighting the model outputs to re-instate the prevalence of events in the data (Lisboa, Vellido, & Wong, 2000).

All attributes share the same value of this hyper-parameter, which controls the convergence of the model coefficients, or weights. Their function is to bias the loading of each covariate, depending on an estimate of how informative that covariate is to fit the data. This estimation is, in essence, the inverse curvature of the Hessian matrix, so that a high curvature indicates a well-determined loading factor, i.e. weight coefficient. Uninformative explanatory variables are forced towards zero, hence the term weight decay is generally used for this type of regularisation, sometimes likened to the use of ridge functions for stabilisation of logistic regression models. An important consequence of this methodology is that an overparameterised model will naturally soft-prune unnecessary covariates leaving a core of statistically significant explanatory variables in this non-linear model.

The PLANN methodology has been extended to competing risks (Boracchi et al., 2003) and, more recently, the ARD framework has been applied to form a Partial Logistic Artificial Neural Network for Competing Risks modelling with Automatic Relevance Determination (PLANN-CR-ARD) (Arsene, Lisboa, Aung, Boracchi, & Biganzoli, 2006).

The application of the ARD framework to PLANN, both for single- and competing risks, enables the estimation of individual predictions of the hazard, calculated over time, with confidence intervals. Given a fixed covariate vector, the confidence intervals arise from the assumed distribution of the weight values, which no longer have point estimates but, rather, are assumed to obey a density function derived from the Bayesian equation for the data fit and regularisation terms. In other words, the confidence intervals reflect the shape of the Hessian function for the model parameters, in the same way as the calculation of the hyper-parameters does, and following a local analytical approximation that exactly mirrors the calculation of predicted standard errors in linear and logistic regression models (Bishop, 1995).

One of the main advantages of the neural network methodology for modelling time-to-event data with right censorship is the ability to infer smooth estimates for the hazard, without requiring *a priori* assumptions about proportionality. This yields useful estimates of the marginal hazards and of covariate effects.

The predictions of hazards for individual cases, which arise by sampling the covariate space using the individual patient's covariate set, form a special case of a method used to explore the covariate space of logistic regression models. The implementation of model predictions, for instance, in the software package GenStat, is weighted using estimated population weights, formed by multiplying together a one-way table of weights for each factor, containing the proportions of cases recorded in each of its levels (Galea et al., 1992). However, individual predictions form an extreme case where the probability density is sampled by a δ-function, so its appropriateness and accuracy requires careful validation.

Performance evaluation was carried out in three ways:

- The estimated cumulative hazard generated by the model was compared with a crude empirical estimate calculated directly from the event-rate in the training data, to verify the accuracy of the model fit.
- The observed survival in the modelling data, described by Kaplan–Meier curves for grouped data, was compared with those for the corresponding risk groups generated by applying the same risk scores to out-of-sample data, used for testing.
- The rules describing the composition of each risk group were verified against clinical expectation.

For this analysis the network was chosen by 5-fold cross-validation on the training data repeated for several models with differing numbers of hidden nodes, from 5 to 12, all other parameters were unchanged. From these, the network parameters were selected and PLANN-ARD run again, this time only once, on all the training data in order to produce just a single model that we could use to test the validation data set. It was found that 10 hidden nodes, 20 iterations for early stopping and all the control parameters set to 0.01 produced stable results.

4. Rule extraction methodology

A principled rule extraction methodology is Orthogonal Search-based Rule Extraction (OSRE) (Etchells & Lisboa, 2006). OSRE extracts conjunctive rules from smooth decision surfaces derived by analytical models, whether they are derived

from traditional statistical models which are linear-in-the-parameters, such as logistic regression, or with generic non-linear approximations to decision surfaces, as is the case for the wide range of ANN architectures.

In this paper the OSRE methodology is used in the derivation of Boolean rules when the explanatory variables are continuously valued. The resulting rule proliferation requires the introduction of a rule refinement strategy to reduce the number of overlapping rules while preserving the cumulative sensitivity and specificity of the original rule set. This naturally leads to a rule hierarchy paradigm to rank the rules and thus generate rule trees. The extended OSRE methodology is a benchmark against alternative data-based rule extraction methodologies by application to publicly available data sets.

The practical value of the rule extraction paradigm is further illustrated with direct visualisation of decision surfaces in well-separated data subspaces, enabling inferences about individual data points to be contextualised within the grouped decisions characterised by Boolean rules.

The OSRE methodology was developed to generate low-order Boolean rules to describe decision surfaces predicted by analytical inference models. In essence, the OSRE efficiently searchers for axis-parallel hyper-cubes to the response surface thresholded at an appropriate value, carving out Boolean regions of data space which fit within the classification regions predicted by the inference model. However, the search mechanism is driven by the data points and so generates a rule for each of them, resulting in a plethora of mutually-overlapping rules. This requires equally efficient algorithms for rule ranking and pruning, which form the core of the OSRE procedures, typically resulting in a small subset of overlapping, low-order rules to explain binary inferences derived from the response surface.

The rule search starts by fitting a classifier to the data. Currently it is achieved either with a logistic regression model or by training a Multi-Layer Perceptron (MLP) regularised with the well-known Bayesian framework with a Gaussian approximation of the evidence (Mackay, 1992). Rules for the data are then extracted by the search engine, initially generating a rule for each data point, and pruned by validating them using the data and known target outcomes.

The present OSRE methodology (Etchells & Lisboa, 2006) involves coding the data into $\{0, 1\}$ for binary variables and 1-from-N binary coding for categorical and ordinal data. This transforms the data into a Boolean space, where each possible input pattern forms a 'restricted' space within this Boolean space (we say 'restricted' space instead of subspace as the sum of two permitted input patterns is not a permitted input pattern, hence by definition not a vector subspace). Each permitted input pattern is a Boolean atom, which is a vertex of a unit hyper-cube of dimension the number of inputs into the network.

Orthogonal search in the context of the OSRE methodology means moving in orthogonal directions in the space of variables before coding. That is, choosing a point in the space and stepping through the values of one variable whilst keeping all others constant. This orthogonal movement when viewed within the Boolean space is the stepping from one permitted atom to one of the next nearest permitted atoms.

Selecting any input pattern and searching in each orthogonal direction and evaluating the ANN at each point, is equivalent to traversing a subset of the permitted atoms in the vicinity of the given input pattern. Every search point that has an in-class decision from the ANN translates to an 'active' atom.

This orthogonal searching results in the formation of hyper-boxes in the space of the binary, categorical and ordinal variables. These hyper-boxes are readily translated into conjunctive (understandable) rules, based on an assumption that the atoms encompassed by this rule are also active. This assumption is validated by testing the rule with the training data and ascertaining its specificity and sensitivity values. Rules with poor specificity and sensitivity are discarded as: poor specificity implies that the assumption that the rules internal atoms are active is not sustainable; very poor sensitivity could imply that the input pattern may well be an outlier or noise and is not representative of the data in general.

A single search will not usually find a rule that describes the behaviour of the network for all possible input patterns. As it is not computationally practicable to investigate all the possible input patterns, a strategy is to use the training data as starting points for the orthogonal searches. Using the training data (and the test data if the data are split for training purposes) for the starting points has two important benefits. Firstly, the ANN is trained using these data and hence the decision surface will be accurately constructed about them, unless the data contains severe outliers or noise. Secondly, the training data accounts for only a tiny fraction of the possible input patterns, making the search process tractable to a high number of inputs, whilst retaining representation of the underlying logic of the data. A full description of the algorithm can be found in Etchells and Lisboa (2006).

A Continuous-Valued Variable (CVV) has values/observations that belong to a finite (or infinite) interval. A possible strategy is to split the finite interval of the values of the CVVs into a number of concurrent sub-intervals. Then a continuous value can then be relabelled dependent on which interval they belong and assigning an ordinal value accordingly. These coded variables can then be further 1-from-N coded and used within the present OSRE methodology. This strategy is problematic in that the ANN will be forced to find a decision boundary based on some arbitrary division of the original continuous interval. Should the classification of the data change from one end of a large interval to the other, the ANN would have great difficulty in finding a decision boundary that separates the data accurately. If this is to be avoided then the sub-interval widths would have to be small, this would lead to many ordinal values and this results in the proliferation of inputs to the networks when 1-from-N coding is performed.

When training neural networks that have CVVs, it is good practice to standardise the data; say scale into the interval $[-1, 1]$ or a standardised normal distribution $N(0, 1)$, so that continuous variables with large values do not dominate the training process or the networks' inferences. This leads to a single input to the network for each CVV, hence keeping

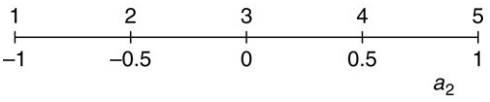

Fig. 1. Categorising the CVV.

the number of inputs to a minimum. A consequence of the standardisation is that the interval of each CVV will usually be no bigger than $[-4, 4]$ if $N(0, 1)$ or $[-1, 1]$ if min–max scaling to $[-1, 1]$ is used.

As the OSRE methodology involves stepping through values of one variable, whilst leaving all others constant, a practical and pragmatic approach to orthogonal searching in the direction of CVVs is to step through the CVV from its minimum to maximum value by some constant distance, searching for a change in the network's response. Each step point can be considered as an ordinal point, so for example stepping from $[-1, 1]$ in a step size of 0.5, leads to 5 ordinal points where each ordinal point can be interpreted as a 1-from-N code Boolean code. This means that each step point can be mapped into the restricted Boolean space, constructed from this and all the other variables.

As an example, a two variable problem with a categorical variable a_1 with values 1, 2 & 3 and a CVV a_2 standardised $[-1, 1]$ and step size 0.5. The variable a_1 values would be mapped to a 1-from-3 coded binary vector and a_2 variable would be mapped to 1-from-5 coded vector (Fig. 1):

Hence, any two-dimensional point in the input space of (a_1, a_2) would transform to a eight-dimensional restricted Boolean space. For example the point $(2, -0.5)$ would map to the atom

$$[\overbrace{0, 1, 0}^{a_1}, \overbrace{0, 1, 0, 0, 0}^{a_2}].$$

This example illustrates that the process of stepping through a CCV is equivalent to categorising the interval and hence each point can be mapped into a restricted Boolean space so the principled approach of OSRE is preserved. Importantly, even though the input space can be mapped into a higher-dimensional Boolean space for theoretical justification, this need not be done for training and rule extraction purposes. Hence there is no proliferation of inputs through 1-from-N encoding even if the step size is reduced to 0.01 or smaller.

This makes the orthogonal search efficient and scalable to small step sizes and very small step sizes allow the orthogonal searches to find the decision boundary in the direction of the CVV to within the accuracy of the step size.

An unfortunate side effect of stepping through small step sizes is that OSRE can produce a plethora of rules, sometimes as many rules as there are data. This is because the orthogonal searching is performed relative to a data point and the distance from a particular data point to the decision boundary may be unique to that data point to within the tolerance of the step size. The principle goal of OSRE is to produce rules that are comprehensible to a human rule; this comprehensibility is compromised if the rule extraction method produces too many rules. A rule refinement technique is introduced that reduces the number of rules by deleting rules that are encompassed wholly by other rules.

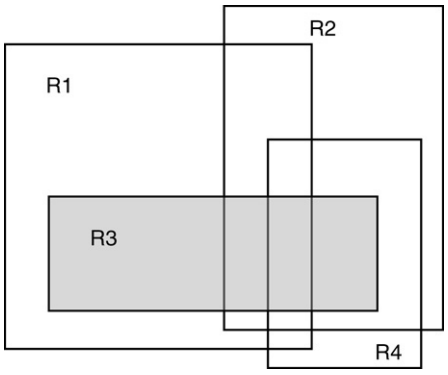

Fig. 2. Rule $R3$ lies within the data space of $R1$, $R2$ and $R4$, therefore can be removed from the rule list.

The disjunction (addition) of all the rules extracted using OSRE form a number of Non-Mutually Exclusive intersecting hyper-boxes in the input space forming a 'hyper-shape'. It is likely that many of the rules may be wholly submerged within parts of this hyper-shape and do not add any extra information/coverage (Fig. 2), and is therefore redundant. A rule refinement method is introduced to reduce the rule set in order to generate a more interpretable explanation of the data, while increasing specificity and controlling any reduction in sensitivity.

Taking the schematic in Fig. 2 as an illustration, the rule R3 can be removed from the rule set as R3 is fully covered by other rules within the set. Furthermore, in the search for comprehensibility the next best candidate for removal would be R4, as the additional coverage (or added value) of the rule does not add much more information. This aspect is covered in the Rule Hierarchy section below.

Each step of the rule refinement process relies heavily on the Receiver Operator Characteristic plot, a two-dimensional plot that has 1-specificity as the horizontal axis and sensitivity on the vertical axis. The ROC point of a rule is a measure of a conjunctive rule to successfully classify the data, sensitivity measuring how much of the in-class data the rule covers and specificity measuring how well it does not classify the out-of-class data. The Global ROC point refers to the ROC point of the disjunction of a set of rules.

The first step in the rule refinement strategy is to filter out all rules whose individual specificity value is below some pre-determined value e.g. 0.9. This will leave a rules set of size n, where n could still be large. A filter of deleting rules with poor sensitivity, say less than 0.1, could be applied here to possibly reduce the rule list further.

The second step of the rule refinement process is the evaluation of the specificity value for each individual rule that will give the best global ROC point for a set of rules. That is, say, filter out all rules with specificity less than 0.91 and evaluate this rule list's global ROC point, then filter out all rules with specificity less than 0.92 etc., up to rules with specificity equal to one. On completion of the procedure a Receiver Operator Curve (ROC) can be produced that connects the global

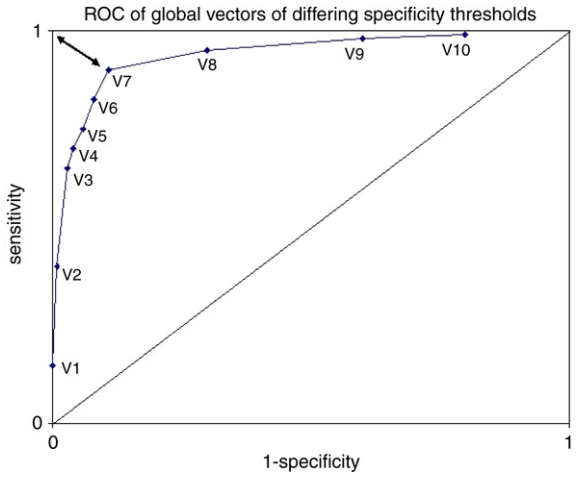

Fig. 3. The first stage in rule refinement finds the best specificity threshold from 10 iterations in this example. Selecting the global coverage, in terms of the true and false positive classifications that is closest to the vector [0 1], the ideal point. Vector V7 in this representation.

vectors for each iteration. By selecting the specificity threshold that has a ROC point closest to the vector [0 1], the ideal point, a disjunction of rules that best describe the classification for maximum coverage whilst controlling for false positives has been chosen (Fig. 3).

At this point the number of rules has been reduced to those below the false positive threshold whose disjunction gives the best global coverage. However, the number of rules can still be prohibitive and present a 'black-box' of rules in terms of understanding the classification. Therefore, further steps are required in the refinement of the rules that OSRE extracts to reduce the number of rules to a minimum and still maintain an acceptable coverage of the classification.

The third step of the rule refinement strategy eliminates rules that are completely contained within the other rules, for example R3 in Fig. 2. This first stage finds a smaller set (if it exists) of rules that have exactly the same global sensitivity and specificity as that of the whole rule set selected in the second stage of refinement.

The rules generated by the OSRE after the second step are placed in a list called **RuleList**.

Step 1. Find the ROC point of **RuleList.**

Step 2. Remove the first rule from the **RuleList** and determine whether there is a change in the ROC point of this reduced **RuleList**. If there is no change then this rule is added to a list called **RemoveRuleList**. The removed rule is replaced back into **RuleList**. Repeat the process for each rule in the **RuleList**.

Step 3. Remove the rules that belong to **RemoveRuleList** from **RuleList**.

Step 4. Re-calculate the ROC point of **RuleList**. If this ROC point is equal to the ROC point in step 1, go to step 6.

If we have reached step 5, *we need to reintroduce rules from* **RemoveRuleList** *to* **RuleList** *so as to move the ROC point of* **RuleList** *back to the point calculated in step* 1.

(At this stage of the algorithm, R3 in Fig. 2 would have been identified as redundant and removed.)

Step 5. Select the rule in **RemoveRuleList** that, when reintroduced to **RuleList**, moves the ROC point of **RuleList** closest to the ROC point calculated in step 1. If there is a tie, select one of the tied rules arbitrarily. Remove this rule from **RemoveRuleList** and add it to **RuleList**. Repeat this process until the ROC point of **RuleList** is equal to that of the ROC point calculated in step 1.

There is now a reduced set of rules, with the same ROC point as the original set of rules with the redundant rules filtered out.

At this stage in the rule refinement process we have a list of Non-Mutually Exclusive rules that will have varied sensitivity and minimum specificity determined in the second step of the refinement. If this list is still long then the rules need to be ordered into a hierarchy where a rule's position is dependent on how 'accurate' it is. The term 'accurate' could refer to the sensitivity or the specificity of the rule, or some other indicator of how well the rule separates the in- and out-of-class data. A number of rule hierarchy methodologies are now introduced that rank the rules so that the position of the rule in the hierarchy is indicative of how much it 'adds value' to the 'accuracy' measure.

Simply ordering the rule list in terms of the sensitivity values does not necessarily give a rule hierarchy that improves the global sensitivity as it is traversed, as there may be significant overlap of the rules. For example, a rule list has a number of rules and the rule R_i has the highest sensitivity value of 0.62. The rule R_j has the next highest sensitivity value of 0.54, however as there is a lot of in-class data that obeys both rules the sensitivity of $R_i \vee R_j$ is only 0.68. However, the sensitivity of a third rule R_k is only 0.3, but it happens that most of the data covered by R_k is not covered by R_i and $R_i \vee R_k$ is 0.79. As a consequence R_k is chosen over R_j for the next rule in the hierarchy. The following algorithm orders the rules in a sensitivity first hierarchy:

Given a data set D and a rule set R and an empty list H

1. Using D calculate the sensitivity of each rule in R; choose the rule, R_i, with the highest sensitivity (pick one at random in the event of a tie) and add to the end of H.
2. Delete that data that obeys R_i from D.
3. Delete R_i from R.
4. If R is not empty go to step 1.
5. Return H.

The algorithm above can be modified to construct a *specificity* first hierarchy by replacing step 1 with:

1. Using D calculate the specificity of each rule in R; choose the rule, R_i, with the highest sensitivity (pick one at random in the event of a tie) and add to the end of H.

Similarly a *Positive Predicted Value* (PPV) first hierarchy can be constructed by changing step 1 to:

1. Using D calculate the PPV of each rule in R; choose the rule, R_i, with the highest sensitivity (pick one at random in the event of a tie) and add to the end of H.

A further measure of rule 'accuracy' is the distance from its ROC point to the point $(0, 1)$. A rule, or a set of rules,

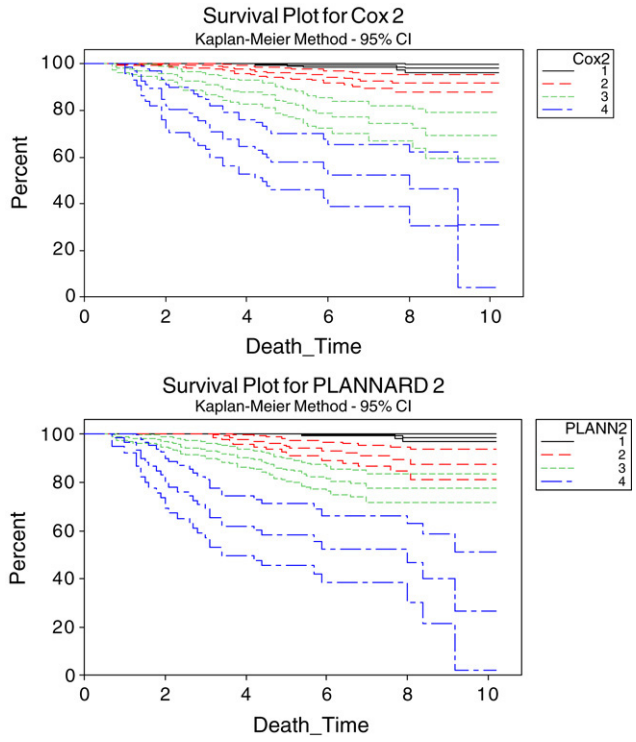

Fig. 4. Grouped survival for breast cancer specific mortality. The model type is indicated in the figure title.

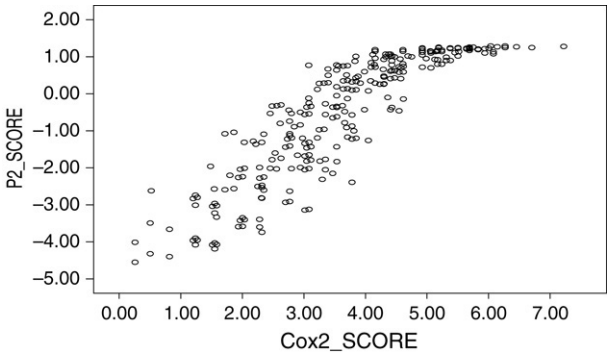

Fig. 5. Cross-matching the risk scores for Cox regression and PLANN-ARD.

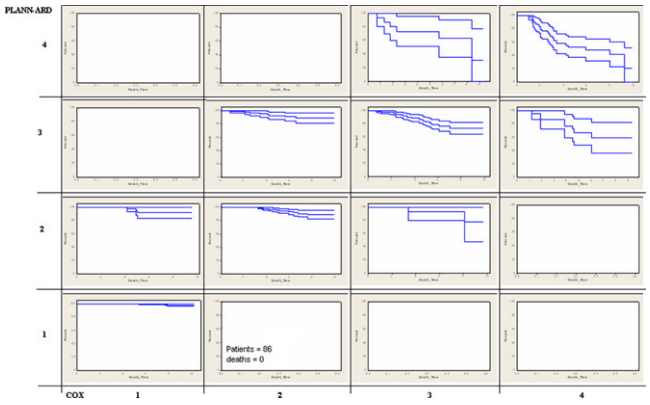

Fig. 6. Cross-matching the survival for Cox regression (abscissa) and PLANN-ARD (ordinate).

with a ROC point (0, 1) represents a perfect rule, that is covers all the in-class data and none of the out-of-class data. As the distance is calculated from the ROC point of the rules, both sensitivity and specificity values are jointly considered in this rule hierarchy. The distance hierarchy algorithm is a modification of the sensitivity first algorithm by replacing step 1 with:

[1] Using D calculate the ROC point (x, y) of each rule in R and evaluate $d = \sqrt{x^2 + (1 - y)^2}$; choose the rule, R_i, with the highest value for d (pick one at random in the event of a tie) and add to the end of H.

5. Breast cancer specific mortality

The result sections are presented in self-contained figures and tables, following a brief critical commentary. In all single risk studies, univariate significance tests were utilised first, to identify a pool of covariates, from which multivariate Cox regression identified statistically significant groups of covariates by forward and backward stepwise feature selection.

All of the available variables were permitted in the model, resulting in the following set of selected covariates:

DCL_T: Tumour stage (clinical)
SBR: Histological grade
GENV: Axillar nodes involved
RPCELL: Progesterone receptor count
ATT_CUTA: Skin invasion

The three variables typically used in quantitative prognostic assessment of breast cancer risk are used, namely: tumour size, generally considered as a measure of the likelihood of the disease metastasizing; histological grade, which is a measure of tumour aggressiveness; the number of axillary nodes affected, which is the most likely route for the disease to spread.

The separation between the mean observed survival of the highest and lowest survival groups identified without surgical variables is around 75% dropping the group at highest risk to around 25% at 10 years from diagnosis, shown in Fig. 4.

A cross-matching plot of the neural network and Cox regression risk scores, shown in Fig. 5, indicates the extent to which non-linearities are present.

This is evident for the patients at highest risk, where a 'tail' with a range of risk scores predicted by the Cox model has a much narrower risk score prediction by the neural network. These patients are more clearly identified as high risk by PLANN-ARD, and preferentially so when surgical variables are *not* included in the model.

A cross-correlation of risk group allocations by the two models is shown in Fig. 6. Cells with small sample sizes are conspicuous for their large confidence intervals. Some of the off-diagonal cells are empty, as would be expected. In particular, both cells in the top row have very low survival, indicating the specificity of the neural network.

Table 1 shows the population sizes for the combined group allocations by both methods and Fig. 7 is a verification test

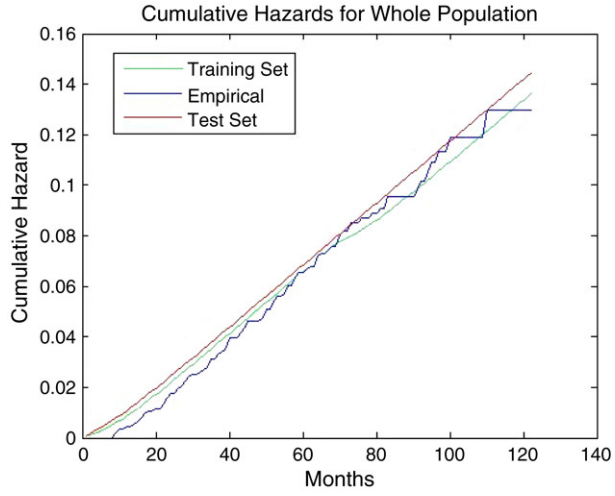

Fig. 7. Verification of the predicted vs. observed cumulative hazards for breast cancer specific mortality applied to the total cohort.

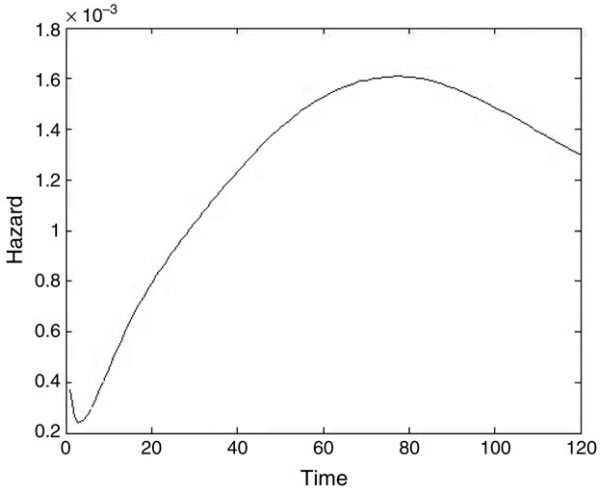

Fig. 8. The marginal hazard shows the typical shape for malignant pathologies.

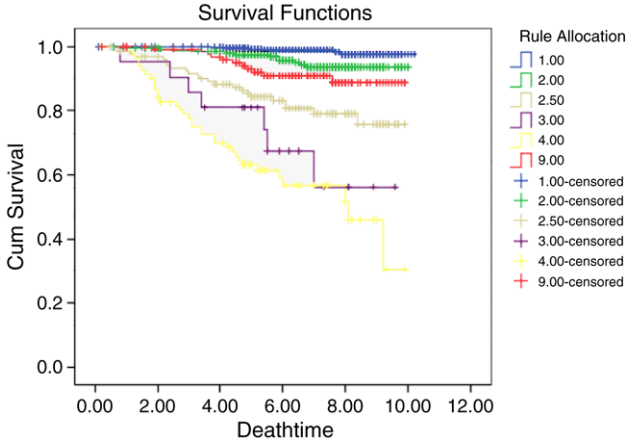

Fig. 9. Breast cancer specific survival predictions for groups allocated by the rules derived from the analytical risk scores.

comparing the predicted cumulative hazard for the training and test data sets, with the crude estimate of the empirical cumulative hazard calculated over the combined data set. The marginal hazards averaged over the complete population are in Fig. 8.

Table 1
Composition of the cells in Fig. 3

PLANN group	Cox group				
	1	2	3	4	Total
4	0	0	15	45	60
3	3	87	143	22	255
2	46	167	14	0	227
1	528	86	0	0	614
Total	577	340	172	67	1156

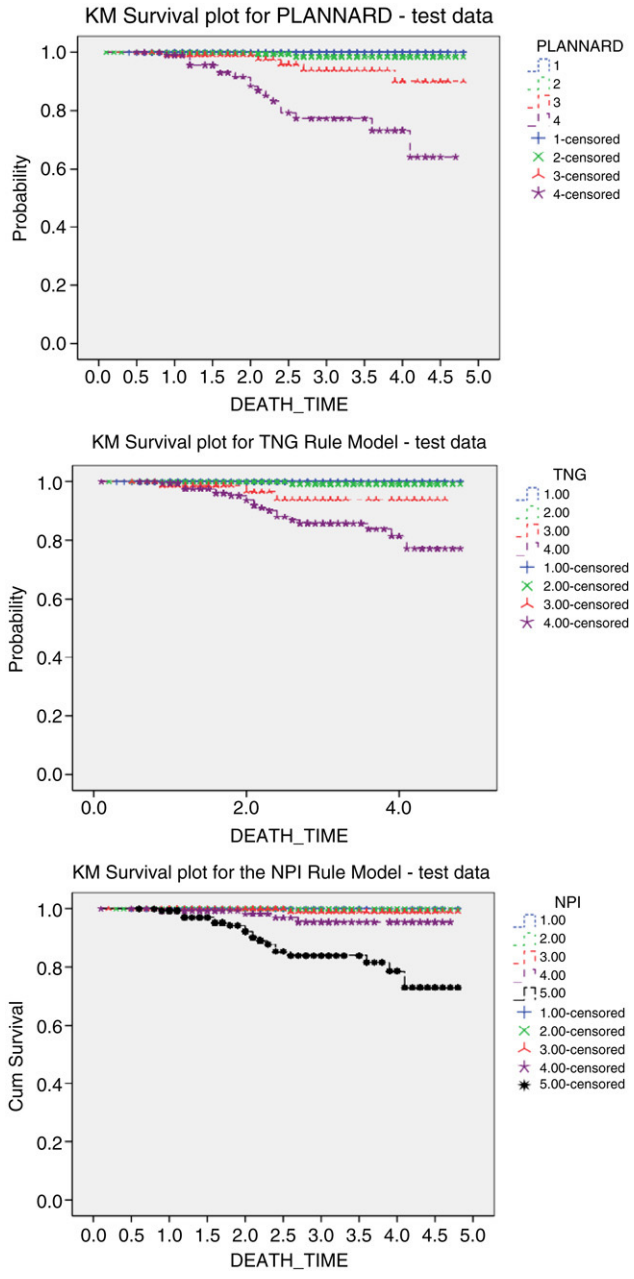

Fig. 10. Kaplan–Meier curves for, from top to bottom, PLANNARD, TNG staging and the NPI rule model.

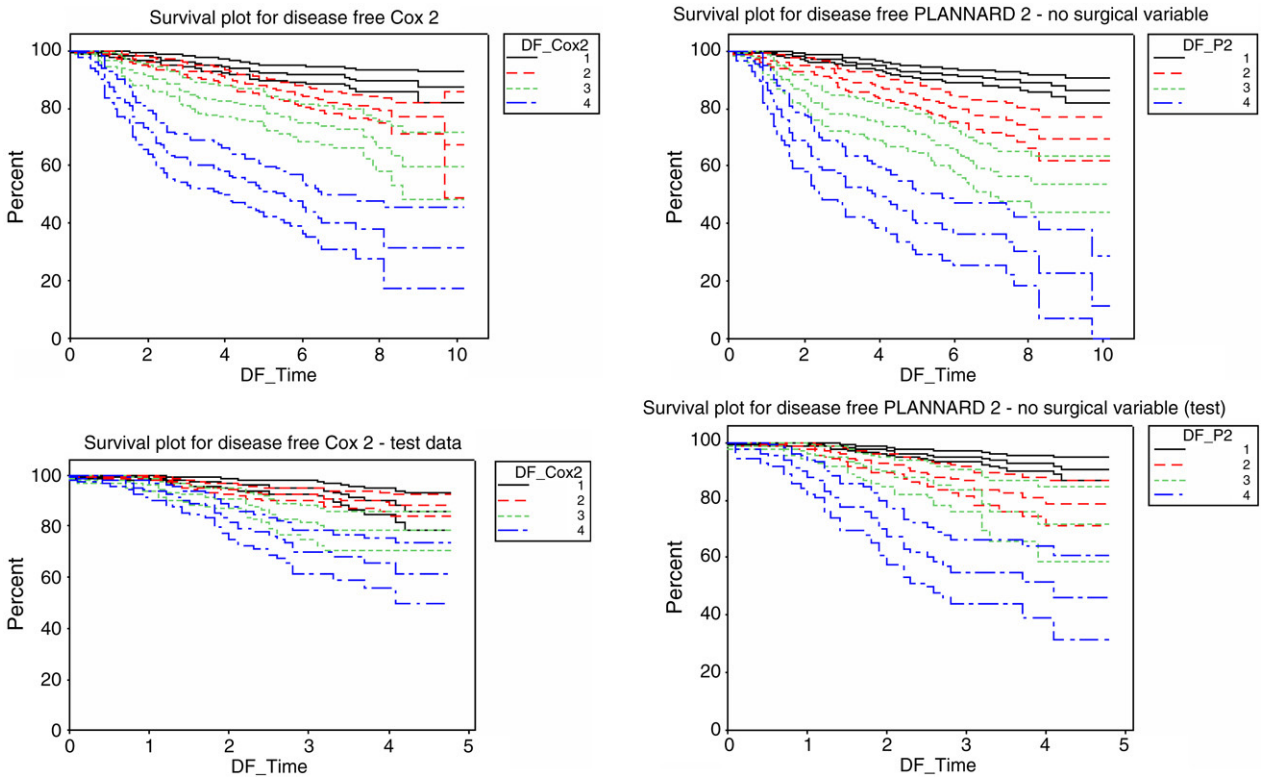

Fig. 11. Kaplan–Meier curves with 95% confidence intervals for grouped data, comparing training and test data (top and bottom) for Cox regression and the PLANN-ARD (left and right).

Explanations for the risk group allocations have been derived for each of the preceding models, using the methodology of Orthogonal Search Rule Extraction (OSRE). This method is designed to automatically generate from the data Boolean rules of low-order, i.e. involving a small number of covariates. This is achieved at the expense of losing mutual exclusivity, so that one patient's allocation to a risk group may be explained by more than one rule. Example rules are shown in Table 2. These rules were used to validate the model using clinical expertise.

Generally speaking, group 1 is characterised by combinations of good prognostic values, rising through mixtures of variables associated with different typical outcomes, to group 4 which combines specific values all associated with poor prognosis, i.e. large tumours with high grade and axillary invasion.

The neural network (and equally Cox regression) can be turned into a 'white-box' by replacing the risk group allocations from analytical scores into strictly defined logical Boolean rules. A few patients fall outside of the rules selected, suggesting that they are outliers in the probability distribution of the covariates. The rule-based group allocation is shown in Fig. 9, separating out patients who fall specifically into rules assigned to each group, but also those whose covariates fit rules from risk groups 2 and 3, which is indicated as '2.5'. The group of outliers is labelled '9' and its survival is roughly in the middle of the Kaplan–Meier plot. This figure achieves a clinically transparent, rule-based, risk allocation from the baseline indicators for breast cancer patients.

Further analysis was then carried out to compare the PLANNARD mortality model with two rule-based models, TNG staging and the NPI rule model (Jarman et al., 2007). These models are a subset of the data made up of operable breast cancer patients with the full set of rules and patient numbers for the training and test data presented in Table 3. All comparisons made are for the test data. The new model uses the same three variables as these models: DCL_T, GENV and SBR plus two additions namely: RPCELL and ATT_CUTA.

The aim is to discover if the additional variables further discriminate the prognostic risk groups in terms of survival. As can be seen from Fig. 10 all three models separate into 3 distinct survival groupings with the new PLANNARD model displaying better separation than the two rule-based models. This suggests that for this data the two extra variables are adding to the discriminatory power of the new model. To discover the generality of these results a future work would involve a multi-centre study.

6. Disease-free survival

Disease-Free Survival (DFS) was also modelled as a single risk, the event of interest being the first documented recurrence, whether local or distal. Significant covariates were identified by Cox regression with forward and backward stepwise feature selection, resulting in the following variables being selected:

DCL_T: Tumour stage (clinical)
SBR: Histological grade
GENV: Axillar nodes involved
RECELL: Estrogen receptor count
AGEPAT: Patient's age

Table 2
The rule sets for cancer specific mortality per group

Group 1 ($N = 606$)	Group 2 ($N = 304$)	Group 3 ($N = 168$)	Group 4 ($N = 78$)
Rule 1: SBR grade<= 2[a] RPCEL>= 10% T staging = T1 No. Invaded Ganglia<= 2[a]	Rule 1: SBR grade= 3[a] T staging = T0 or T1 No. Invaded Ganglia<= 2	Rule 1: SBR grade= 3 RPCEL < 50%[a] T staging = T2 or T3 No. Invaded Ganglia<= 2[a]	Rule 1: SBR grade = 3 RPCEL< 50%[a] T staging = T2 or T3 or T4 No. Invaded Ganglia >= 3
Rule 2: SBR grade<= 2[a] RPCEL > 50% T staging = T0 or T2 No. Invaded Ganglia<= 2[a]	Rule 2: SBR grade<= 2[a] RPCEL < 50%[a] T staging = T2 or T3 No. Invaded Ganglia <= 2[a]	Rule 2: SBR grade<= 2[a] RPCEL < 50%[a] T staging = T2 or T3 No. Invaded Ganglia >= 3	Rule 2: SBR grade = 3 T staging = T3 or T4
Rule 3: SBR grade<= 2[a] T staging = T0 No. Invaded Ganglia<= 2[a]	Rule 3: SBR grade= 2 or 3[a] RPCEL< 10% T staging = T1 or T2 or T3 No. Invaded Ganglia = 0 or >= 3	Rule3: SBR grade= 3 RPCEL< 50%[a] T staging = T1 No. Invaded Ganglia >= 3[a]	Rule 3: RPCEL < 50%[a] T staging = T4 No. Invaded Ganglia>= 3
Number of cases true by rule set in group 1 = 536	Number of cases true by rule set in group 2 = 201	Number of cases true by rule set in group 3 = 111	Number of cases true by rule set in group 4 = 78
Number of cases true by rule set not in group 1 = 0	Number of cases true by rule set not in group 2 = 94	Number of cases true by rule set not in group 3 = 11	Number of cases true by rule set not in group 4 = 8

[a] Indicates that this rule statement is also inclusive of cases where the variable is missing.

NB_NOD: N-stage (clinical)
MTUMINF: Nipple infiltrating tumour
MENOPAUS: Menopausal status

The structure of the significant explanatory variables follows an expected pattern comprising the three core variables associated with NPI, together with hormone receptor count, now represented by estrogen rather than progesterone as was the case in the mortality study.

The other half of the selected variables includes age and clinical stage nodes, but interestingly also menopausal status. This is unexpected alongside age, as there is a significant overlap in the breast cancer relevant information that these two variables contain. An additional factor, nipple infiltrating tumour, is new compared with other studies, but this may be because this variable was not appropriately represented in previously published data sets.

One further remark to make is that stepwise feature selection with the Cox model is known to be prone to select a few too many variables. It is therefore possible that the model is slightly overfitted, in particular that as many as two of the selected variables may be in excess of the optimal set for robust future predictions. The PLANN-ARD neural network is designed to suppress the effect of uninformative variables in the model.

The robustness lent to the neural network by regulation with ARD may explain why the results in Fig. 11 show better separation and spread of mean group survivals for the neural network compared with Cox regression, using the same covariate sets.

Table 4 lists some of the derived rules that explain risk group allocations for the vast majority of patients. As before, the rules are a validation tool to ensure that the operation of the neural network model (or Cox model) is consistent with clinical knowledge, but also to derive new insights from patient

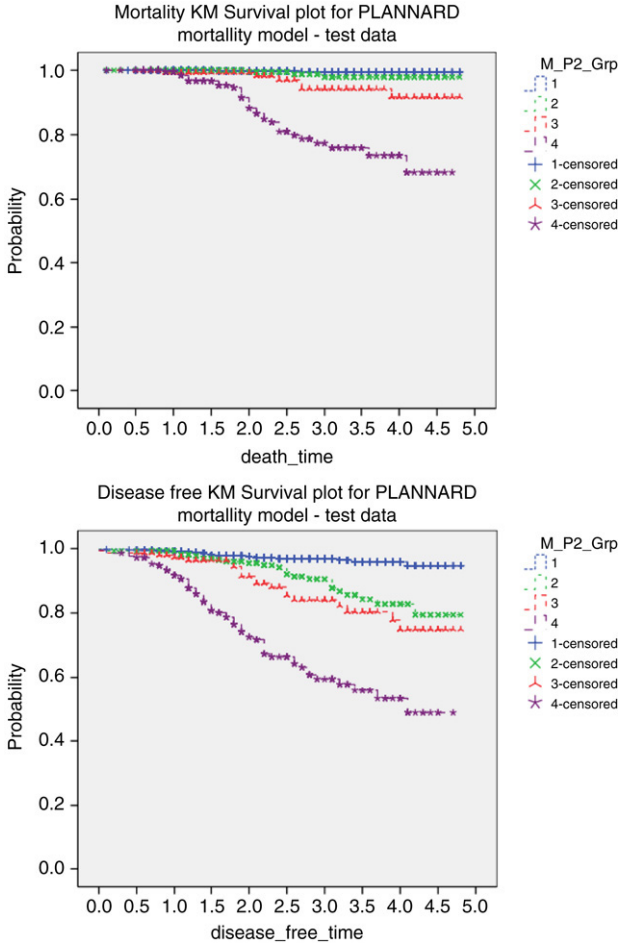

Fig. 12. Kaplan–Meier survival plots for the PLANNARD mortality risk groups for separate observations: (top) mortality and (bottom) disease-free survival.

Table 3
The rule sets for TNG staging and the NPI rule model

Group	Rule no.	T size[a]	Nodes[a]	Grade[a]		Train (n)	Test (n)
TNG staging							
I	1.1	T1	N1	G1 or G2		224	231
	1.2	T1	N2	G1		46	45
	1.3	T2	N1	G1		16	12
					Sub-total	**286**	**288**
II	2.1	T1	N1	G3		67	68
	2.2	T1	N2	G2		81	103
	2.3	T1	N3	G1		10	12
	2.4	T2	N1 or N2	G2		77	84
	2.5	T2	N2	G1		12	10
					Sub-total	**247**	**277**
III	3.1	T1	N2	G3		28	42
	3.2	T2	N1	G3		34	44
					Sub-total	**62**	**86**
IV	4.1	T1	N3	G2 or G3		75	61
	4.2	T2	N2	G3		44	30
	4.3	T2	N3	Any G		78	84
					Sub-total	**197**	**175**
					Total	**792**	**826**
NPI rule model							
I	1.1	T1	N1	G1		85	84
					Sub-total	**85**	**84**
II	2.1	T1	N1	G2		139	147
	2.2	T1	N2	G1		46	45
	2.3	T2	N1	G1		16	12
					Sub-total	**201**	**204**
III	3.1	T1	N1	G3		67	68
	3.2	T1	N2	G2		81	103
	3.3	T1	N3	G1		10	12
	3.4	T2	N1	G2		44	50
	3.5	T2	N2	G1		12	10
					Sub-total	**214**	**243**
IV	4.1	T1	N2	G3		28	42
	4.2	T1	N3	G2		43	25
	4.3	T2	N1	G3		34	44
	4.4	T2	N2	G2		33	34
	4.5	T2	N3	G1		8	4
					Sub-total	**146**	**149**
V	5.1	T1	N3	G3		32	36
	5.2	T2	N2 or N3	G3		82	79
	5.3	T2	N3	G2		32	31
					Sub-total	**146**	**146**
					Total	**792**	**826**

[a] Tsize = DCL_T, Nodes = GENV, Grade = SBR.

groups whose survival is unexpected but who can, nevertheless, be strictly characterised by a well-defined Boolean profile.

However, the statistically significant separation between the prognostic scores of groups 2 and 3, found in the training data, does not generalise to the test set. This is clear from Fig. 11. Interestingly, the poor performance in risk group allocation by the rule set for groups 2 and 3 also suggests that this separation is not reliable, therefore the groups should be merged.

In the introduction it was noted that in general doctors consider disease-free survival and mortality to be equivalent when they used a prognostic model and often would talk in terms of being disease free when using a mortality risk model. To investigate whether this practise is appropriate two KM survival plots, Fig. 12 were produced for the same risk model namely: the new PLANNARD mortality risk model, for separate observation, first the usual target; death, the second with disease-free survival as the target.

Although the order of the risk groups remain the same for the two targets of interest the proportional difference for groups 1 and 2 are quite different and overall survival over time is markedly lower for disease-free survival. This suggests that mortality and disease-free survival should not be used interchangeably when discussing prognosis.

7. Conclusion

The breast cancer specific mortality study confirmed that successful risk-staging can be carried out both with Cox

Table 4
The rule sets for disease-free survival for the group

Group 1 ($N = 349$)	Group 2 ($N = 481$)	Group 3 ($N = 191$)	Group 4 ($N = 135$)
Rule 1: SBR grade= 1[a] T staging not T4 No. Invaded Ganglia<= 2 23 < Age < 70	Rule 1: SBR grade >= 2[a] RECEL>= 10% T staging = T1 or T3 No. Invaded Ganglia<= 2 23 < Age < 76 No. Nodules >= 1[a]	Rule 1: SBR grade >= 2 RECEL >= 10% T staging= T1 or T3 No. Invaded Ganglia <= 2[a] 58 < Age < 86	Rule 1: SBR grade >= 2 T staging = T3 or T4[a]
Rule 2: RECEL >= 10% T staging = T0 or T2 No. Invaded Ganglia <= 2 23 < Age <= 60	Rule 2: SBR grade >= 2[a] T staging = T0 or T2 No. Invaded Ganglia <= 2[a] 63 < Age < 90	Rule 2: SBR grade >= 2 RECEL >= 10% T staging = T1 or T3 No. Invaded Ganglia = 3 35 < Age < 60	Rule 2: SBR grade >= 2 RECEL <= 50%[a] T staging = T1 or T3 or T4[a] Ganglia >= 3[a]
Rule 3: SBR grade = 1[a] T staging = T0 or T2 No. Invaded Ganglia<= 2	Rule 3: SBR grade>= 2[a] T staging = T1 or T3 No. Invaded Ganglia <= 2 23 < Age < 59 No. Nodules >= 1[a]	Rule3: SBR grade>= 2 RECEL <= 50%[a] T staging = T1 or T3 No. Invaded Ganglia <= 2[a] 36 < Age < 64	Rule 3: SBR grade >= 2 T staging = T1 or T3 or T4[a] No. Invaded Ganglia >= 3[a] 57 < Age < 90
Number of cases true by rule set in group 1 = 303	Number of cases true by rule set in group 2 = 286	Number of cases true by rule set in group 3 = 115	Number of cases true by rule set in group 4 = 116
Number of cases true by rule set not in group 1 = 50	Number of cases true by rule set not in group 2 = 87	Number of cases true by rule set not in group 3 = 125	Number of cases true by rule set not in group 4 = 45

[a] Indicates that this rule statement is also inclusive of cases where the variable is missing.

regression and with the PLANN-ARD neural network. The neural network appears to be more specific to identify patients at the extremes of high and low risk. Model selection includes three widely accepted prognostic indicators together with additional covariates known to have prognostic significance.

Disease-free survival, treated as a single risk, yielded models and risk groups that are consistent with those derived in the mortality study. Interestingly, PLANN-ARD seems to generalise better than Cox regression, perhaps because of the explicit use of complexity regulation in the neural network model, which enable soft-pruning and hence reduces any possible overfitting to the training data. The neural network makes smooth predictions of the mean and cumulative hazards. These results were verified by comparing with crude empirical estimates. Moreover, Boolean rules can be generated to explain the allocation of individuals to risk groups, providing a validation tool to ensure that the operation of the risk allocation model is consistent with clinical judgement. In addition, there is evidence that disease-free survival and mortality should not be used interchangeably when discussing prognosis. A further step can be taken to produce a fully rule-based risk allocation system, which has good specificity for group survival predictions and is fully transparent for clinicians.

Further evaluation of the predictive accuracy of time-to-event models for censored data can be carried out within the extension of the AUROC into a time-dependent performance index in Antolini, Boracchi, and Biganzoli (2005). The PLANN-ARD model is also currently being extended for the analysis of multiple competing risks.

Acknowledgments

The work was carried out as a collaboration between Liverpool John Moores University, Centre Leon Bérard, and Themis-ICTA Group. Financial support from Pfizer France is gratefully acknowledged.

References

Antolini, L., Boracchi, P., & Biganzoli, E. (2005). A time-dependent discrimination index for survival data. *Statistics in Medicine*, *24*, 3927–3944.

Arsene, C. T. C., Lisboa, P. J. G., Aung, M. S. N, & Boracchi, P. Biganzoli, E. (2006). A Bayesian neural network for competing risk models with covariates'. In *IET 3rd int. conf advances in medical. signal and information processing*.

Biganzoli, E., Boracchi, P., Mariani, L., & Marubini, E. (1998). Feed forward neural networks for the analysis of censored survival data: A partial logistic regression approach. *Statistics in Medicine*, *17*, 1169–1186.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Clarendon Press.

Boracchi, P., Biganzoli, E., & Marubini, E. (2001). Modelling cause-specific hazards with radial basis function artificial neural networks: applications to 2233 breast cancer patients. *Statistics in Medicine*, *30*, 3677–3394.

Boracchi, P., Biganzoli, E., & Marubini, E. (2003). Joint modelling of cause-specific hazard functions with cubic splines: An application to a large series of breast cancer patients. *Computational Statistics & Data Analysis*, *42*, 243–262.

Etchells, T. A., & Lisboa, P. J. G. (2006). Orthogonal search-based rule extraction (OSRE) for trained neural networks: A practical and efficient approach. *IEEE Transactions on Neural Networks*, *17*(2), 374–384.

Galea, M. H., Blamey, R. W., Elston, C. E., & Ellis, I. O. (1992). The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Research and Treatment*, *22*, 207–219.

Haybittle, J. L., Blamey, R. W., Elston, C. W., Johnson, J., Doyle, P. J., Campbell, F. C., et al. (1982). A prognostic index in primary breast cancer. *British Journal of Cancer*, *45*, 3621.

Jarman, I. H., Etchells, T. A., Ellis, I. O., Green, A. R., & Lisboa, P. J. G. (2007). A rule-based alternative to the Nottingham Prognostic Index: A breast cancer prognostic model. In *3rd int. conf. computational intelligence in medicine and healthcare*.

Lisboa, P. J. G. (2002). A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural Networks*, *15*(1), 9–37.

Lisboa, P. J. G., Vellido, A., & Wong, H. (2000). Bias reduction in skewed binary classification with Bayesian neural networks. *Neural Networks*, *13*, 407–410.

Lisboa, P. J. G., Wong, H., Harris, P., & Swindell, R. (2003). A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. *Artificial Intelligence in Medicine*, *28*(1), 1–25.

Lisboa, P., Etechells, T., Jarman, I., Aung, H., & Perol, D. (2007). Time-to-event analysis with artificial neural networks: An integrated analytical and rule-based study for breast cancer. In *Proceedings of IJCNN '07*.

Mackay, D. J. C. (1992). A practical Bayesian framework for back propagation networks. *Neural Computation*, *4*(3), 448–472.