Prognosis in node-negative primary breast cancer: a neural network analysis of risk profiles using routinely assessed factors

E. Biganzoli1*, P. Boracchi2, D. Coradini3, M. Grazia Daidone3 & E. Marubini2

1Unità di Statistica Medica e Biometria, 2Unità Operativa Determinanti Biomolecolari nella Prognosi e Terapia, Istituto Nazionale per lo Studio e la Cura dei Tumori, Milan; 3Istituto di Statistica Medica e Biometria, Università degli Studi di Milano, Milan, Italy

Received 6 February 2003; revised 30 May 2003; accepted 18 July 2003

Background: The present study investigated complex time-dependent effects of routinely assessed factors on the risk of breast cancer recurrence over follow-up time, with a partial logistic artificial neural network (PLANN) model.

Patients and methods: PLANN was applied to data from 1793 patients with node-negative breast cancer, not submitted to any adjuvant treatment and with a minimal potential follow-up of 10 years.

Results: The shape of the hazard function changed according to histology, which showed a time-dependent effect, partly modulated by estrogen receptors (ERs). Age and progesterone receptors (PgR) showed protective effects; the latter was more evident for short follow-up and high ER values. Tumour size and ER content showed time-dependent unfavourable effects at early and long follow-up times, respectively. Predicted values of disease recurrence probability at 2 years of follow-up showed that low steroid-receptor content, young age and large tumour size were associated with the highest risk of relapse. Although the oldest patients with high ER content seem to be those most protected overall, high risk predictions tend to spread also to higher steroid-receptor contents, intermediate ages and small tumour size, with an increase in follow-up time.

Conclusion: PLANN with suitable visualisation techniques provided thorough insights into the dynamics of breast cancer recurrence for improving individual risk staging of node-negative breast cancer patients.

Key words: artificial neural networks, breast cancer, prognostic factors, survival analysis

Introduction

Only a few standardised and routinely assessed factors have recognised prognostic relevance in primary breast cancer [1–3]. Although chemotherapy reduces the risk of breast cancer recurrence in axillary node-negative (N0) patients, the definition of risk categories based on patient and tumour characteristics (age, pathological tumour size, histology, and steroid-receptor content) is useful to identify low risk patients who would receive negligible benefit. Moreover, improvements to current practice might be achieved by accounting for the modulations of the effects of the above-mentioned variables on the rate of disease recurrence in the course of follow-up, aiming at identifying patients considered at low risk at early follow-up (i.e. 2–3 years after primary therapy), but who may have high risk of disease recurrence at a longer follow-up (8–10 years). Statistical models on the hazard function (i.e. the instantaneous risk of disease recurrence) are suited to this task, having an exploratory role in the ‘knowledge phase’ of the disease dynamics. Relevant effects, discovered within this phase, support the development of prognostic classification schemes for the subsequent ‘decision phase’.

For example, the time-dependent prognostic relevance of steroid receptors has long been debated [4–6] and several studies [7–10], including our own performed in a large series of N0 breast cancer patients without adjuvant systemic therapy after surgery [11], have shown that the impact of estrogen receptors (ERs) on the risk of recurrence changes during the course of follow-up. However, the presence of non-additive effects of steroid receptors and other routinely assessed prognostic variables was not investigated, and no information on the shape of the hazard, as a function of time, was available from the Cox model.

The few studies performed so far on the relationship between the hazard function and routinely assessed prognostic factors are generally based on their dichotomisation, with predefined cut-off values [8–10]. A more informative analysis of the dynamics of breast cancer recurrence should thus involve the evaluation of the joint effect of steroid-receptor content and other prognostic factors, accounting for their original scale of measurement. It would therefore be useful to adopt a non-linear modelling approach that is able to identify previously unknown prognostic relationships underlying data, avoiding too restrictive a priori assumptions, and to integrate and verify the consistency of the results obtained with conventional modelling strategies. Methods based on artificial neural networks (ANNs) appear to be suited for this task.

ANNs are statistical models whose mathematical structure reproduces the biological organisation of neural cells for simula-
tion of the learning dynamics of the brain [12]. Considerable attention has been paid in recent years to the application of ANN-based regression methods for the development of prognostic models in oncology. Although some doubts have been raised about the true advantages of ANNs over traditional techniques [13], a recent review highlights their benefits for outcome prediction [14]. A desirable improvement in their use as ‘black box’ predictors is the recognition of the underlying prognostic relationships and risk profiles, which may well be non-linear in nature and difficult to analyse using standard statistical methodology [15].

The possible use of ANNs as exploratory tools for the study of disease dynamics has been neglected because no ANN models suitable for the hazard function of censored survival data were available. The partial logistic ANN (PLANN) approach, based on the extension of the well-known logistic regression model, has been developed for the analysis of the hazard as a function of multiple variables and time [16]. A challenging issue in PLANN is the representation of complex results for their clinical interpretation. With this approach, visualisation techniques are exploited for evaluating covariate effects on the hazard function and model-predicted disease recurrence probability at different follow-up times.

The aim of the present study was to investigate the joint role of ERs and progesterone receptors (PgR), and other clinically relevant patient and tumour characteristics (age, tumour size, histology), on the risk of cancer recurrence during follow-up after surgery of N– patients without adjuvant therapy to avoid any possible confounding effect due to systemic treatment. ER, PgR, age and tumour size were analysed, accounting for their original scale of measurement, in order to improve the accuracy of the picture of their effects on the hazard function and individual risk profiling. The results obtained with the PLANN model integrated previous findings [11] from the time-dependent Cox model, outlining the shape of the hazard function, and provided additional details on the prognostic role of steroid receptors as a function of time, and other factors useful to improve follow-up and/or therapy planning.

Patients and methods

Patients

The case series consisted of 1793 women with primary resectable invasive breast cancer who had no axillary lymph node involvement, and no radiological or clinical evidence of distant metastasis, a synchronous bilateral tumour or a second primary tumour. Cases with these clinicopathological features and with a minimal potential follow-up of 10 years (i.e. the time elapsed from the date of surgery to the last updating of the patient records) were selected from some 7000 women with an operable tumour, consecutive to their median values. Confidence intervals were estimated by fixing other continuous covariates to their median values. To investigate possible high order interactions, multpanel conditioning plots were adopted [23]. These plots display the joint effect of two covariates and follow-up time on the logarithm of the relative hazard (log HR), fixing other continuous covariates to their median values. Confidence intervals were estimated by means of a non-parametric bootstrap procedure [24].

Multiple correspondence analysis (MCA) was adopted to visualise the relationship between model predictions and covariate patterns [25]. For this

### Table 1. Patient and Tumor Characteristics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgery, n (%)</td>
<td></td>
</tr>
<tr>
<td>Mastectomy</td>
<td>864 (48)</td>
</tr>
<tr>
<td>Quadrantectomy + radiotherapy</td>
<td>929 (52)</td>
</tr>
<tr>
<td>Age (years)</td>
<td></td>
</tr>
<tr>
<td>Median (25th percentile to 75th percentile)</td>
<td>52 (45–62)</td>
</tr>
<tr>
<td>Tumor size (cm)</td>
<td></td>
</tr>
<tr>
<td>Median (25th percentile to 75th percentile)</td>
<td>1.7 (1.2–2.3)</td>
</tr>
<tr>
<td>Estrogen receptors (fmol/mg cytosolic protein)</td>
<td></td>
</tr>
<tr>
<td>Median (25th percentile to 75th percentile)</td>
<td>56 (15–158)</td>
</tr>
<tr>
<td>PgR (fmol/mg cytosolic protein)</td>
<td></td>
</tr>
<tr>
<td>Median (25th percentile to 75th percentile)</td>
<td>61 (9–232)</td>
</tr>
<tr>
<td>Histology, n (%)</td>
<td></td>
</tr>
<tr>
<td>Invasive ductal carcinoma (IDC)</td>
<td>1302 (73)</td>
</tr>
<tr>
<td>Invasive lobular carcinoma (ILC)</td>
<td>235 (13)</td>
</tr>
<tr>
<td>IDC + ILC</td>
<td>122 (7)</td>
</tr>
<tr>
<td>Other</td>
<td>134 (7)</td>
</tr>
</tbody>
</table>

### Receptor determination

ER and PgR concentrations were determined using the dextran-coated charcoal method, quantified by multiple-point Scatchard analysis, and expressed as fmol/mg of cytosolic protein as described previously according to the European Organisation for Research and Treatment of Cancer (EORTC) standard assay protocol [17, 18]. ER and PgR levels were determined within the Quality Control Program activated by the Italian Committee for Hormone Receptor Assay Standardization [19].

### Statistical analysis

ANNs [12, 20] have been introduced recently to model censored survival data, accounting for complex prognostic patterns. A general framework for neural network models on censored survival data has recently been proposed [21]. Within this framework, PLANN is an ANN generalisation of partial logistic regression, strictly related to the conventional logistic regression [16, 22], which is suitable for smoothed hazard estimation as a function of time and covariates, and allows for non-linear, non-proportional and non-additive effects. A description and a graphical representation of the model is provided in the Appendix (Figure A1). Since the distributions of steroid receptors were positively skewed and proportional increments of their values were considered relevant for their prognostic impact, the natural logarithmic transformation of these variables was adopted. Models were applied to 1715 (95.6%) cases with complete information, on the considered variables. The analysis was based on 10-year curtailed follow-up information.

To visualise the effect of the continuous covariate (ER, PgR, age, tumour size) and the categorical one (histology), model results were represented by three-dimensional surface plots of estimated conditional discrete hazard as a function of time and of the continuous covariate for each type of histology, after fixing other continuous covariates to their median values. To investigate possible high order interactions, multpanel conditioning plots were adopted [23]. These plots display the joint effect of two covariates and follow-up time on the logarithm of the relative hazard (log HR), fixing other continuous covariates to their median values. Confidence intervals were estimated by means of a non-parametric bootstrap procedure [24].

Multiple correspondence analysis (MCA) was adopted to visualise the relationship between model predictions and covariate patterns [25]. For this
Figure 1. Surface plots of PLANN model results for: (A) logarithm of estrogen receptors log(ER); (B) logarithm of progesterone receptors log(PgR); (C) age; and (D) tumour size. Estimated conditional hazards are plotted as a function of follow-up time (years) for each histology type.
The details of MCA have been described elsewhere [11]. Briefly, MCA is a suitable approach to investigate the association among categorical and continuous variables. Variables and subjects can be plotted onto a plane defined by the first two factorial axes (which mainly contribute to explain the total variability of the original data) according to new coordinates (factorial scores). Such a plot approximates the association patterns underlying data, since, considering the origin of factorial axes, the angular distance among categories and subjects is related to their mutual associations (smaller distances indicate

Figure 2. Multi-panel conditioning plot of PLANN model results for: (A) histology and estrogen receptors (ERs); (B) histology and progesterone (PgR) receptors; and (C) ER and PgR. Estimated log relative hazards are plotted as a function of follow-up time (years) for selected ER and PgR values.
higher associations). For MCA, the categories of the continuous variables defined for PLANN model implementation (see Appendix) and histological type were plotted jointly with subjects, according to their factorial scores. In addition, the size and the colour level of the dots are proportional to the predicted probability of disease recurrence: large dark and small light dots indicate high and low risk subjects, respectively. Therefore, the relative position of the covariate categories indicates the association between model-predicted values and covariate patterns.

Results

The plots of the hazard as a function of follow-up time and log(ER), log(PgR), age and tumour size are reported in Figure 1 for different histologies. Overall, the plots show that the shape of the hazard function changes according to histology. A marked non-monotonic shape over time is observed for invasive ductal carcinoma (IDC) and for the class ‘other’ histology, whereas for invasive lobular carcinoma (ILC) such behaviour is attenuated, and for IDC plus ILC the hazard appears to be steadily increasing. This pattern confirms the time-dependent histological role already uncovered [11]. The effect of ERs is shown in Figure 1A. Little or no effect on the hazard is present early during follow-up, whereas after ∼3 years increasing log(ER) values have an unfavourable prognostic effect. Such behaviour seems to differ according to histology, suggesting the presence of a high-order interaction with follow-up time, which would be difficult to detect using conventional statistical approaches.

The joint effects of histology and ER have been investigated further, resorting to multipanel conditioning plots of the log HR, estimated using the PLANN model for different combinations of the above two variables. Point estimates (solid lines) with 95% bootstrap confidence intervals (dotted lines) are reported in Figure 2A, considering as reference low ER (1 fmol/mg) and IDC histology. The long-term risk effect of ILC and IDC plus ILC is much more evident for intermediate (10 fmol/mg) to high (100 fmol/mg) ER values. Increasing ER values appear to be associated with a slight increase in the log HR at long follow-up; however, this time-dependent effect is less than that of histology.

The protective role of PgR on the risk of disease recurrence is shown in Figure 1B. A multipanel conditioning plot log HR for PgR and histology in Figure 2B show an increase in the protective effect of IDC plus ILC at early follow-up for high (100 fmol/mg) PgR values. The protective role of age is shown in Figure 1C. The effect of variables related to hormone responsiveness was investigated further by means of multipanel conditioning plots of the log HR for ER and PgR, ER and age, and PgR and age. Figure 2C shows relevant results for ER and PgR in IDC histology, considering as reference low ER and PgR (1 fmol/mg). The protective effect of PgR is much more evident at early follow-up for intermediate (10 fmol/mg) to high (100 fmol/mg) ER values.

Figure 1D shows the time-dependent prognostic role of tumour size: the risk of recurrence is greatest for early follow-up time and large tumour size (it increases with an increase in tumour size), but this effect lessens after ∼5 years of follow-up as the hazard decreases.

To obtain an overall view of the association between covariate levels and model-predicted risks, a multivariate visualisation approach (MCA) was adopted [25]. Predicted disease recurrence probabilities for each subject were jointly projected with covariate categories (see Appendix) onto a plane defined by the first two factorial axes. For a dynamic view of the evolution of the risk of disease recurrence, plots of 2-, 5- and 8-year predictions are reported in Figure 3A, B and C, respectively. According to the
The general trend, high risk predictions at 2 years are mainly concentrated in the lower right corner of the graph, being associated with intermediate to large tumour size, low ER and PgR levels, and young age. Conversely, low risk predictions appear to be well separated, being mainly concentrated in the upper left corner and associated with older age, high ER and high PgR levels. The separation tends to diminish from 5–8 years. Although the oldest patients with a high ER content appear to be the most protected overall, high risk predictions progressively spread towards the upper left corner of the graph with the increase in follow-up time, being gradually associated with higher steroid-receptor content, intermediate ages and smaller tumour sizes. The increasing effects of ERs and histology, along with the decreasing effect of tumour size over time, probably contribute to this evolving pattern.

**Discussion**

Although today most primary breast cancer patients without axillary lymph node involvement are candidates for adjuvant systemic therapy (endocrine, cytotoxic or both), the investigation
into prognostic factors in case series of untreated patients is still relevant to identify women with an intrinsically low risk of disease recurrence who could be spared from adjuvant systemic therapy [2, 3].

Prediction of breast cancer recurrence is a difficult task; in fact, although a considerable number of variables have been investigated, only a few have proved to have a clinically relevant prognostic role, even considering their availability in a reliable way on the whole breast cancer population [1–3]. Therefore, clinical practice is still based to a large extent upon information provided by only a few clinical and pathological factors, including patient age, tumour size and histology, together with routinely measured biological markers such as hormone steroid receptors, which are considered more useful for predicting the clinical response to hormone therapy than for discriminating patients at different risk of relapse. An important question is whether the prognostic information provided by the above variables has been fully exploited, particularly for those that are measurable on a quantitative scale. Flexible statistical models play a key role in the discovery and evaluation of complex unknown prognostic relationships over prolonged follow-up (at least 10 years) in breast cancer. In this context, Coradini et al. [11] applied a flexible time-dependent extension of the Cox model to a large series of N–breast cancer patients, focusing on the relative effects of covariates without exploring the shape of the hazard function over time. The PLANN approach, an ANN suitable for the analysis of the hazard as a function of time and covariates for censored survival time data [16], was adopted in the present work to investigate the joint effects of routinely measured prognostic factors on the dynamics of tumour recurrence. Therefore, the present results integrate those achieved with the Cox regression models.

ANN-based models rely on inductive inference. They attempt to discover biologically and clinically relevant prognostic relationships directly from patient and tumour data, thereby avoiding too restrictive a priori assumptions [15]. In their conventional application to outcome prediction, ANNs have been adopted according to this principle, even though they are essentially used as ‘black boxes’. In fact, although they may improve the accuracy of the prediction, they consist of a statistical model not immediately accessible to clinicians and investigators. As we are aware of the need to increase the knowledge of the dynamics of breast cancer recurrence, the PLANN model is not exploited as a black box. Namely, multivariate visualisation techniques are adopted for exploring covariate effects on the hazard function over time, and patient profiles are displayed with the corresponding model-predicted disease recurrence probabilities at different follow-up times. According to this strategy, prognostic relationships underlying data are identified, in order to understand how the model predicts outcomes before further clinical application. With this aim, the identification of time-dependent risk profiles can be useful in improving follow-up and/or therapy planning.

The substantial consistency of the results of the PLANN model with previous ones confirmed the time-dependent prognostic relevance of histology, ER, and pathological tumour size, reveal-

**Figure 3.** Visualisation of predicted disease recurrence probabilities and covariate patterns by means of multiple correspondence analysis on the plane identified by the first two factorial axes (factors 1 and 2). Points coordinates are factorial scores. Dots identify subjects, and dot size and colour level are proportional to the predicted probability of disease recurrence. The corresponding colour scale is reported on the right of each panel. The labels in the plot correspond to the class into which the specific covariate has been grouped. Patterns of the risk of disease recurrence at (A) 2, (B) 5 and (C) 8 years of follow-up, respectively.
Moreover, they constitute the basis for the identification of the prolonged follow-up, because they could be at high risk of disease interrelationships between variables. Patients with recognised the evolution of patient risk of relapsing that better clarifies the function according to tumour histology, thus suggesting different recurrence dynamics as a function of histology. Moreover, such behaviour is partly influenced by ER and PgR. Different to previous results [11], in PLANN analysis the protective role of PgR is much more evident for early follow-up. Overall, present results suggest a different impact of hormonal regulation between ductal and lobular structures. As expected, tumour size showed a time-dependent effect that was similar in all histologies, supporting the evidence for an unfavourable impact on early recurrence, which decreases over time. Overall these effects could be relevant for improving the accuracy of the prediction of disease evolution in each patient.

The joint representation of patient profiles with the corresponding PLANN-predicted disease-free survival probabilities after 2, 5 and 8 years of follow-up provides a comprehensive picture of the evolution of patient risk of relapsing that better clarifies the interrelationships between variables. Patients with recognised unfavourable prognostic features such as large tumour size, young age and low steroid-receptor content appeared to be at particular risk of relapse at the earliest follow-up time. However, patients with small tumours associated with high ER levels and IDC plus ILC histology should also be carefully monitored, even after prolonged follow-up, because they could be at high risk of disease recurrence in the medium to long term. It should be considered that the probabilities of disease recurrence according to the studied prognostic factors may be substantially different following the administration of an adjuvant systemic therapy. Considering the response to such therapy, it would be useful to assess whether the increase in risk at long follow-up for the above-mentioned patients may be overcome. A partial evidence against such a chance is the presence of time-dependent effects for ERs also for case series including treated patients [8]. Although the present results need to be confirmed for other case series, they show the possible advantages of the application of flexible statistical techniques, such as ANNs, for survival analysis in an exploratory framework.

In conclusion, thorough insights into the dynamics of breast cancer recurrence have been obtained by the application of an ANN model to survival time data. The present findings allow us to refine the evidence previously acquired using Cox regression analysis on the same data set and using the same covariates. Moreover, they constitute the basis for the identification of the prognostic contribution of new variables and the improvement of biological hypotheses about the dynamics of breast cancer recurrence. This step is critical for the generation of future prognostic classification schemes that should provide tailored prediction for individual patients, and thus contribute to the improvement in the clinical management of early breast cancer.

**Acknowledgements**

This work was supported in part by grants from the Italian Association for Cancer Research and the Ministero della Sanità.

**Appendix**

**The PLANN model**

The general mathematical structure of the PLANN model [16] is as follows:

$$\lambda(m, x) = f \left[ \beta_0 + \sum h \beta_h g_h \left( \beta_{h0} + \beta_{hm} m + \sum k \beta_{hk} x_k \right) \right]$$

In the corresponding graphical representation (Figure A1), input nodes are assigned to the explanatory variables ($x_1, \ldots, x_K$) for the $i$th observation, namely: age (years), tumour size (mm), ER and PgR content (logarithmic transformation of fmol/mg cytosolic protein), and histology (lobular, ductal, mixed lobular plus ductal, other), whereas time, grouped into $I = 1 \ldots L$ intervals with midpoints ($m_i$), is treated as an additional input. Input values are summed with weights $\beta_h$ and $\beta_{hm}$ transformed by hidden nodes with activation functions $g_h(\cdot)$. The weighted sum of the values of the hidden nodes with weights $\beta_h$ is transformed by a single output node with activation function $f(\cdot)$ to estimate the discrete hazard function $\lambda(m, x_1, \ldots, x_K)$.

The logistic function $g_h(\cdot) = f(\cdot) = \exp(\cdot) / [1 + \exp(\cdot)]$ is used as activation for both the hidden and the single output node of the network. The PLANN model provides smoothed estimates of the discrete hazards, whereas the degree of smoothing is modulated by the number of hidden nodes and penalty terms for the error.

![Graphical representation of the PLANN model.](image-url)

**Figure A1.** Graphical representation of the PLANN model.
function (weight decay). Since weight decay is used, all input variables are rescaled to span, approximately, between 0 and 1 [20]. Estimation of model parameters ($\beta_m$, $\beta_h$, $\beta_p$, $\beta_c$) is attained by minimisation of an error function (Kullback–Leibler distance for binomial data) with a quasi-Newton algorithm. Models with too many estimated parameters would be overfitted (under-smoothed), including idiosyncratic features of the present data not generalisable to other case series, whereas too small (over-smoothed) models would provide excessively biased estimates of the hazard function. For model selection the network information criterion (NIC) was used [20]. To reduce the size of the data matrix the ‘binning’ approach was followed and covariate values were grouped according to the procedure proposed by Gray [23]. Accounting for the marginal distributions, 10 groups were adopted for age, ER and PgR, whereas nine groups were adopted for tumour size. One-year intervals were adopted for follow-up time, which was truncated at 10 years. The cut-point values for grouping are reported as follows: age (39, 43, 46, 49, 52, 57, 60, 65, 70 years); ER (2, 10, 22, 36, 56, 83, 125, 199, 322 fmol/mg of cytosolic protein); PgR (1, 6, 17, 36, 65, 115, 188, 308, 525 fmol/mg of cytosolic protein); and tumour size (10, 11, 13, 15, 17, 20, 25, 30 mm).

Models with 6–12 hidden nodes and 0.01–1 penalty values with 10 different weighting initialisations were evaluated. The selected model had eight hidden nodes and 0.18 as a penalty value, with ~20 free estimated parameters. The number of effectively estimated parameters is much smaller than the theoretical number (81) for the network model with eight hidden nodes, but such a result is expected due to the effect of the penalty term. In fact, for penalty values >0.10 the choice of the number of hidden nodes in the network was found to have less influence than weight decay in modulating the number of parameters actually estimated. Results on model selection obtained with NIC were further confirmed by a leave-five-out, non-linear cross-validation procedure [26]. Confidence intervals on the PLANN model results have been estimated by means of a non-linear version [26] of the non-parametric bootstrap procedure [24].

References