A Clinical Decision Support System for Breast Cancer Patients

Ana S. Fernandes[1], Pedro Alves[1], Ian H. Jarman[2], Terence A. Etchells[2], José M. Fonseca[1], and Paulo J.G. Lisboa[2]

[1] Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa
{asff,jmf}@uninova.pt
[2] School of Computing and Mathematical Sciences, Liverpool John Moores University,
Byrom Street, Liverpool L3 3AF, UK
{T.A.Etchells,I.H.Jarman,P.J.Lisboa}@ljmu.ac.uk

Abstract. This paper proposes a Web clinical decision support system for clinical oncologists and for breast cancer patients making prognostic assessments, using the particular characteristics of the individual patient. This system comprises three different prognostic modelling methodologies: the clinically widely used Nottingham prognostic index (NPI); the Cox regression modelling and a partial logistic artificial neural network with automatic relevance determination (PLANN-ARD). All three models yield a different prognostic index that can be analysed together in order to obtain a more accurate prognostic assessment of the patient. Missing data is incorporated in the mentioned models, a common issue in medical data that was overcome using multiple imputation techniques. Risk group assignments are also provided through a methodology based on regression trees, where Boolean rules can be obtained expressed with patient characteristics.

Keywords: Breast cancer, survival analysis, decision support systems.

1 Introduction

Prognostic assessments as well as clinical indicators are the key issues to guide clinical oncologists to better define the treatments and to better assess the impact of prognostic factors on survival of operable breast cancer patients.

This paper presents a web decision support system for clinical oncologists, where three different survival models are considered: the commonly used NPI (Nottingham Prognostic Index), Proportional Hazards Modelling and PLANN-ARD (Partial Logistic Artificial Neural Network with Automatic relevance determination), such that each model provides an independent prognostic index. The prognostic indices (PI) for each of the models are derived from prognostic factors and allow stratification of patients by survival outcome. A patient stratification methodology is also introduced to separate the population into significantly different survival risk groups In addition, these risk groups can be characterized using explanatory rules obtained from the prognostic factors used in the analysis. Both, the patient's risk group and the explanatory rules can be incorporated in the Web decision support system. It is important to mention that the aim of the proposed decision support system is to enhance the oncologists'

L.M. Camarinha-Matos, P. Pereira, and L. Ribeiro (Eds.): DoCEIS 2010, IFIP AICT 314, pp. 122–129, 2010.

current practices, rather than to replace them. Therefore, all the previous models are incorporated in the web interface.

Section 2 explains the current study's contribution to technical innovation, section 3 gives a description of the data set used to train the model and defines the predictive variables chosen for the analysis. Section 4 presents the prognostic models and the methodologies used for patient stratification into different survival groups and Section 5 presents the Web decision support system followed by the conclusions.

2 Contribution to Technical Innovation

The present work makes an important contribution to both technical innovation and clinical application as several important novelties were added or changed to current practice. Jarman et al (2008) have already presented a web decision support system as a relevant innovation [1]; this study improves upon this system by resolving and improving, some particular issues. Currently there are several survival models which are in use, such as NPI and other Cox proportional hazards models. It is intended to augment NPI by adding more variables considered to be important in the prognostic model, which selection is explained in section 4. Moreover, it was intended to define a prognostic model to become predictive rather than explanatory as well as modelling non-linear dependences, with PLANN-ARD. Previous research [2] on the dataset used for this study and mentioned on section 3, showed missing data to be missing at random (MAR): hence, it can be successfully imputed. Therefore this work also takes account of missing data and censorship within principled frameworks [2], applying multiple imputation in combination with neural network models for time-to-event modelling, where a new PI was also considered. It is important to note that survival models must take account of censorship, which occurs when a patient drops out of the study before the event of interest is observed or if the event of interest does not take place until the end of the study. Moreover a new stratification methodology was developed, based on decision trees, which adds a more robust path to identify the patient's risk group and the explanatory rules that characterize risk group membership, based on patient's characteristics. Finally, a new web decision support system contributes to technical innovation as it implements both the previously mentioned models, where all can be compared.

3 Data Description

The data set comprise 931 consecutive series of female patients recruited by Christie Hospital in Wilmslow, Manchester, UK, during 1990-94. The current study is specific to early operable breast cancer patients filtered using the standard TNM (Tumour, Nodes, Metastasis) staging system as tumour size less than 5 cm, node stage less than 2 and without clinical symptoms of metastatic spread. The event of interest is death to any cause, being a single risk model, where the study period for analysis is 5 years and the time-to-event was measured in months from surgery. All patients in this study were censored after 5 years of follow-up. 16 explanatory variables in addition to outcome variables were acquired for all patient records.

This study will only focus on Histological type lobular and ductal, therefore some records were withdrawn. Also, two of the 931 records in the training data were identified as outliers and removed. Finally, at the end of the analysis the data set ended up with 743 subjects. Missing data is a common problem in prediction research. After analysing the data set, information has been considered to be Missing at Random (MAR) where a new attribute may be created to denote missing information or the missing values can be imputed. The latter has been shown to be effective [3]. Therefore, the missing covariates were imputed following the method indicated in [3] and repeated 10 times. The choice of this number is a conservative one, as several studies have shown that the required number of repeated imputations can be as low as three for data with up to 20% missing information.

4 Breast Cancer Prognostic Models and Stratification Methods

This section explains both, the different prognostic models and the stratification methodology which were included in the web decision system. The following detailed methods bridge the gap between individual predictions for single patients and allocations of patients into risk groups.

4.1 Breast Cancer Prognostic Models

It is important to mention that historically, the purpose of prognostic models was to stratify patients into cohorts with distinct survival. The most widely used index is TNM, which is purely clinical, as it depends only on clinical investigations and palpation and takes account of metastatic spread of the disease but is not sufficiently detailed for early breast cancer. The Nottingham prognostic index (NPI) [4], a clinical prognostic index for breast cancer patients has been widely applied to inform the choice of adjuvant therapy. It is an indicator of breast cancer outcome and its score is calculated using the following formula:

$$\mathrm{NPI}=0.2\times \textit{pathological size} + \textit{histological grade} + \textit{nodes involved}\ . \tag{1}$$

Subsequently, from their NPI patients are allocated into prognostic groups from excellent prognostic group to poor prognostic group based on the cutpoints: <2.41; <3.41; <5.41 and ≥5.41. Predictive prognostic inference for individual patients was also introduced by the web-based interface for clinical oncologists which has expanded the covariate basis for prognostic inference. www.Adjuvantonline.com [5] is an interface format that appears to be readily accepted by practicing clinicians. This model has the advantage to infer the potential effect of different treatment choices and since its publication on the web, there is greater interest in making individualised predictions of survival. However, its predictions do not include confidence intervals, yet are likely to be subject to substantial uncertainties for particular groups of patients. Advances in therapy, detection technologies and health policy have skewed the patient population and additional prognostic indicators can be added to increase the predictive power of NPI. Consequently, the other two prognostic models are included and compared in this study.

In the survival analysis field, the proportional hazard model, also known as Cox regression is widely used. Cox regression factorises dependence on time and the covariates, where the hazard rate is modelled for each patient with covariates x_p at time t_k, as follows:

$$\frac{h(x_p,t_k)}{1-h(x_p,t_k)}=\frac{h_0(t_k)}{1-h_0(t_k)}.\exp(\sum_{i=1}^{N_i}bx_i). \tag{2}$$

where h_0 is the baseline hazard function and x_i are the patient variables. Here the prognostic score is defined by the traditional linear index *βx*. However, for this study 10 imputed data sets were used, which means that the final prognostic score for each patient was determined as the mean of the 10 prognostic indices identified,

Model selection was carried out through Cox regression (proportional hazards) [2], where six predictive variables were identified: *age at diagnosis*, *node stage, histological type*, *ratio of axillary nodes affected to axilar nodes removed*, *pathological size* (i.e. tumour size in cm) and *oestrogen receptor count.* All variables are binary coded as 1-from-N.

The Partial Logistic Artificial Neural Network is a predictive model, rather than an explanatory model, such as the proportional hazards regression and also has the capability of capturing interactions between covariates as well as fitting the time dependence of the hazard function. It has a strong regularisation framework which has been added to avoid overfitting, using the method of Automatic Relevance Determination, hence the acronym PLANN-ARD [6]. For a single risk, such as overall mortality, this model has the structure of a multi-layer perceptron with a single hidden layer and sigmoidal activations in the hidden and output layer nodes. The number of hidden nodes was determined using cross-validation that is several networks were trained, each one with different hidden nodes and validated with cross-validation. It was concluded that 8 hidden nodes were sufficient to train the network and not lead to over-fitting.

Covariates and discrete time (monthly time increments) are introduced in the network as inputs, where the output is the hazard for each patient and for each time. Estimating the weights requires a likelihood term for the status of one patient at time t_k, by using an indicator when the patient status is observed alive at time t_k (labeled as 0) or have died (label as 1). The papers cited in [6] make a strict theoretical correspondence between this neural network model and classical statistical time-to-event models for censored data. To obtain patient information, that is appropriate to their prognostic risk group, it is important to define first this risk group as well as the relevant prognostic score, appropriated for non-linear models, as with the previous mentioned prognostic models. Therefore, the following expression is proposed:

$$PI(x_p)=(-\ln(1-CCI(t))) = \ln(-\ln(S(t))). \tag{3}$$

where the CCI is the crude cumulative incidence, identified as the probability of the occurrence of a specific event of interest and *S(t)* is the estimated survival at the end of follow up. As a consequence of imputation, PLANN-ARD was computed 10 independent times, one to each imputed data set, which resulted in 10 different trained networks. The mentioned time independent PI was also computed for the 10 networks and averaged for each patient, producing a final PI.

4.2 Group Risk Stratification

In a clinical environment stratification of patients, in different risk groups, based on survival models is frequently used in the evaluation of treatments or to assess the impact of prognostic factors on survival. Therefore, a stratification methodology needs to be defined in order to separate the different patients in statistically significant risk groups by overall mortality. Previous studies [7] presented a comparison between different stratification methodologies, such as the bootstrap log-rank aggregation based on the log-rank test and a regression tree, based on CART algorithm [8], where the PI is now the target for regression using a rule-tree. The bootstrap log-rank aggregation was considered in order to diminish the log-rank cut-points overestimation [9]. All stratification methodologies were applied to both survival models: PLANN-ARD, a neural network for time-to-event data and Cox proportional hazards.

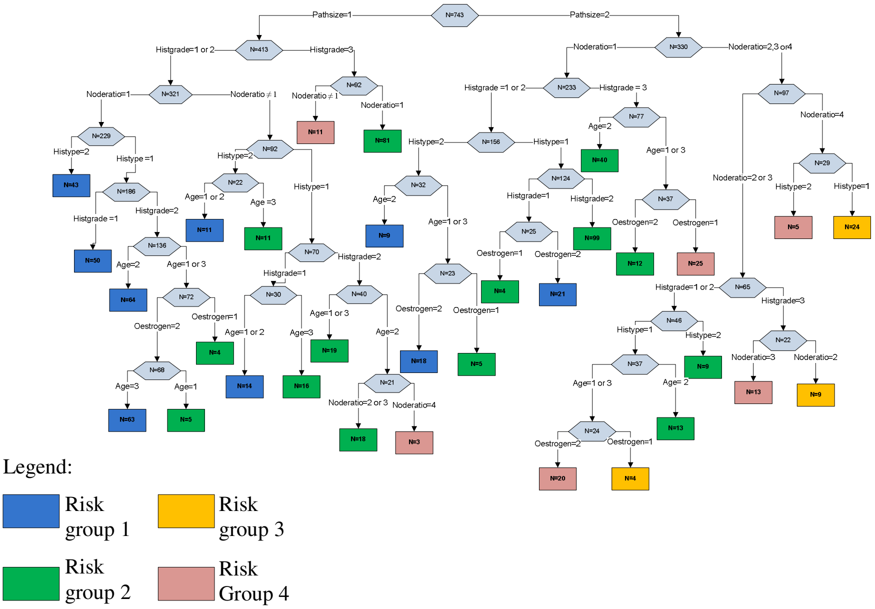


Fig. 1. Final regression tree, indicating the risk groups belonging for all the patients with different predictive variables, using the proportional hazards model as a prognostic model

However, a concern of many clinicians is the 'black box' nature of artificial neural networks (ANN) [10] which raises the important issue of explaining individual inferences by the network. This is a key stage in evaluating the clinical plausibility of inferences made by analytical models to enable clinicians to apply these inferences with confidence. Consequently, the regression tree can be a very well accepted stratification method in a clinical environment, as it gives simple explanatory rules, based on the predictive variables for all the risk groups considering all the existent possibilities, as it can be observed in figure 1. This figure represents the final regression tree, where, thanks to the pruning method, the final leafs define different risk groups. Each risk group is characterized by different rules based on the patient's characteristics which are defined by each branch of the regression tree.

5 Framework for Integrated Decision Support System

All the prognostic models, combined with the stratification methods described above can be integrated with decision supports for clinicians and patients, being used for personalized patient information systems. Figure 2 represents the framework web home page, where the three prognostic models can be computed and the output compared for a single patient. The framework's main goal is to assist the clinicians and patients in defining the appropriate information to their prognostic risk group, by way of a cross-matching matrix for all the different methodologies. The interface combines a group score (the NPI index) with two statistical models (one linear and one non-linear) to estimate breast cancer specific mortality. This provides a 'second opinion' for current users of Adjuvant! but using a single data-based model and hence with the potential to provide not just point estimates of survival but also theoretically derived confidence intervals for those estimates. The Kaplan Meier survival curves can also be available for each cell of the matrix in order to discover heterogeneity within a prognostic risk group. It is important to mention that the basis of this framework is to detail the prognosis risk group using different methodologies, rather then replacing one for another. The same idea of cross-tabulation can be extended to a scatter plot of the PI. It can also inform if some patients are outliers of the model or in the borderline between groups.

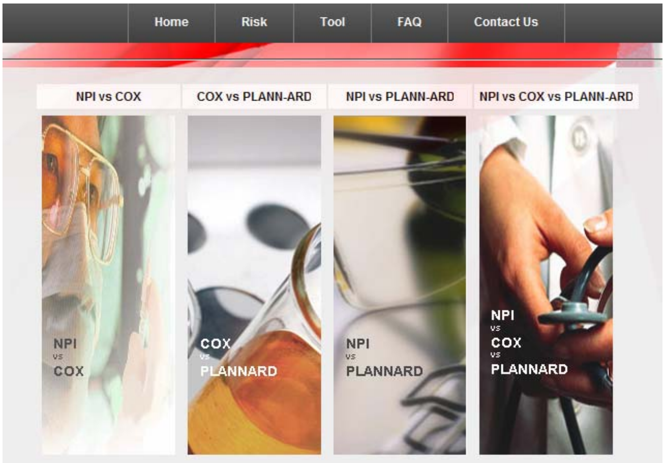
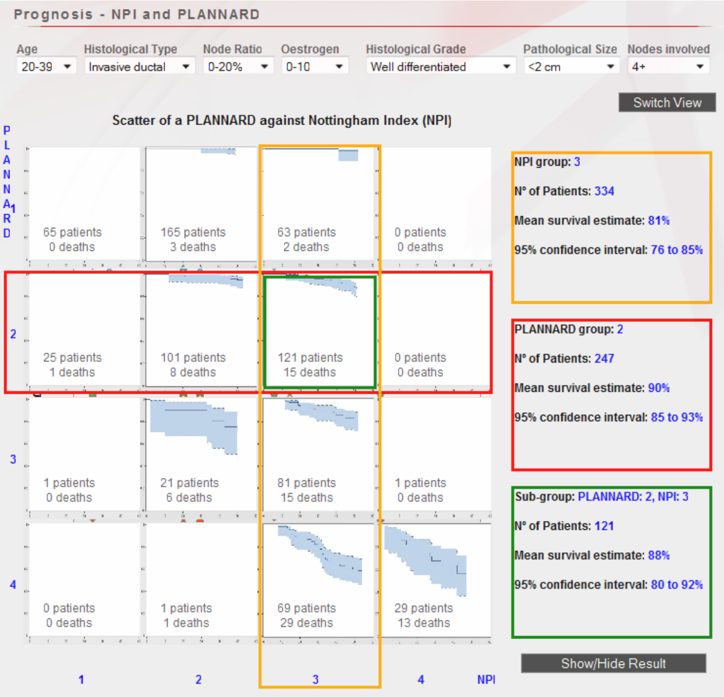


Fig. 2. Home page of a patient information system for breast cancer patients with the possibility to choose a 2D or 3D cross-matching visualisation for different risk group allocation

Patient's characteristics are submitted and the risk scores as well as risk groups are calculated. These are displayed in the web page, figure 3. The cross-matching can be visualized in two dimensional and three dimensional plots. As an example, figure 3 represents in the ordinate the currently used clinical risk score, namely the Nottingham Prognostic Index (NPI), while the abscissa is the PI derived from the PLANN-ARD. The result is that patients in NPI group 3 are shown in the plot to be stratified into different risk groups. Finally, this Web clinical decision support system also allows clinical oncologists to collect and save their patients prognosis as well as their clinical data. Therefore, supported by patient history, clinical oncologists have the possibility to compare prognosis and treatments and improve their medical decisions.



Prognosis - NPI and PLANNARD

Age	Histological Type	Node Ratio	Oestrogen	Histological Grade	Pathological Size	Nodes involved
20-39	Invasive ductal	0-20%	0-10	Well differentiated	<2 cm	4+

Scatter of a PLANNARD against Nottingham Index (NPI)

PLANNARD \ NPI	1	2	3	4
1	65 patients 0 deaths	165 patients 3 deaths	63 patients 2 deaths	0 patients 0 deaths
2	25 patients 1 deaths	101 patients 8 deaths	121 patients 15 deaths	0 patients 0 deaths
3	1 patients 0 deaths	21 patients 6 deaths	81 patients 15 deaths	1 patients 0 deaths
4	0 patients 0 deaths	1 patients 1 deaths	69 patients 29 deaths	29 patients 13 deaths

NPI group: 3
Nº of Patients: 334
Mean survival estimate: 81%
95% confidence interval: 76 to 85%

PLANNARD group: 2
Nº of Patients: 247
Mean survival estimate: 90%
95% confidence interval: 85 to 93%

Sub-group: PLANNARD: 2, NPI: 3
Nº of Patients: 121
Mean survival estimate: 88%
95% confidence interval: 80 to 92%

Fig. 3. User interface for breast cancer oncology, showing the risk group belonging for PLANN-ARD model and NPI PI and a visualisation of the patient data

6 Conclusions

This paper presents a web decision support system for breast oncology, which shows the value of the new prognostic models and stratification methodology in discriminating patients by mortality risk. This tool, after introducing patient variables, identifies the risk group allocation for the three prognostic models presented in this paper (NPI, Cox proportional hazards, PLANN-ARD) and consequently the rules that explain each risk group. A cross-matching matrix of grouped survival and the position a patient resides within the matrix it is also presented, leading to better insights if the risk group allocation for a specific prognostic model is more accurate. Moreover, as Adujantonline is gaining clinical support, a link to this web tool has been placed in the presented web decision support system. Future work on the prognostic model will include a detailed validation of the prognostic predictions by application to out-of-sample data collected by the British Columbia Cancer Agency. This is the same data set that was used to evaluate the Adjuvantonline system, thus enabling benchmarking. In order to improve the presented tool, for future work the different treatment choices for the data base used for training the prognostic model.

Acknowledgments. The authors gratefully acknowledge Mr. R. Swindell from Christie Hospital for making the anonymised modelling data set available for this study and members of the BC Cancer Agency's Breast Cancer Outcomes Unit for assembling the British Columbia data. This study was partially funded by the European Network of Excellence Biopattern (FP6/2002/IST/1-508803) and by Fundação para a Ciência e Tecnologia, through the POS_C(SFRH/BD/30260/2006).

References

1. Jarman, I.H., Etchells, T.A., Martín-Guerrero, J.D., Lisboa, P.J.G.: An integrated framework for risk profiling of breast cancer patients following surgery. Artificial Intelligence in Medicine 42, 165–188 (2008)
2. Fernandes, A.S., Jarman, I.H., Etchells, T.A., Fonseca, J.M., Biganzoli, E., Bajdik, C., Lisboa, P.J.G.: Missing data imputation in longitudinal cohort studies – application of PLANN-ARD in breast cancer survival. In: 2008 Seventh International Conference on Machine Learning and Applications, 2008, ICMLA 2008, pp. 644–649 (2008)
3. Clark, T.G., Altman, D.G.: Developing a prognostic model in the presence of missing data an ovarian cancer case study. Journal of clinical epidemiology 56, 28–37 (2003)
4. Haybittle, J.L., Blamey, R.W., Elston, C.W., Johnson, J., Doyle, P.J., Campbell, F.C., Nicholson, R.I., Griffiths, K.: A prognostic index in primary breast cancer. Brit J. Cancer 45, 3621 (1982)
5. Ravdin, P.M., Siminoff, L.A., Davis, G.J., Mercer, B.M., Hewlett, J., Gerson, N., Parker, H.L.: Computer Program to Assist in Making Decisions about Adjuvant Therapy for Women with Early Breast Cancer. J. Clin. Oncol. 74(4), 980–991 (2001)
6. Lisboa, P.J.G., Wong, H., Harris, P., Swindell, R.: A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. Artificial Intelligence in Medicine 28(1), 1–25 (2003)
7. Fernandes, A.S., Etchells, T.A., Jarman, I.H., Fonseca, J.M.: Stratification methodologies for neural networks models of survival. In: Cabestany, J., et al. (eds.) IWANN 2009. LNCS, vol. 5517, pp. 989–996. Springer, Heidelberg (2009)
8. Breiman, L., Friedman, J.H., Olsen, A.R., Stone, C.J.: Classification and Regression Trees. The Wadsworth & Brooks (1984)
9. Etchells, T.A., Fernandes, A.S., Jarman, I.H., Fonseca, J.M., Lisboa, P.J.G.: Stratification of severity of illness indices: a case study for breast cancer prognosis. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part II. LNCS (LNAI), vol. 5178, pp. 214–221. Springer, Heidelberg (2008)
10. Lisboa, P.J.G.: A review of evidence of health benefit from artificial neural networks in medical intervention. Neural Networks 15(1), 9–37 (2002)