Stratification of Severity of Illness Indices: A Case Study for Breast Cancer Prognosis

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

[1] 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
[2] Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa
{asff, jmf}@uninova.pt

Abstract. Prognostic modelling involves grouping patients by risk of adverse outcome, typically by stratifying a severity of illness index obtained from a classifier or survival model. The assignment of thresholds on the risk index depends of pairwise statistical significance tests, notably the log-rank test. This paper proposes a new methodology to substantially improve the robustness of the stratification algorithm, by reference to a statistical and neural network prognostic study of longitudinal data from patients with operable breast cancer.

Keywords: Risk modelling, log-rank test.

1 Introduction

Stratification of patients by risk of adverse outcome is central to clinical practice. This begins with modelling empirical data either with a classifier or a failure time model, depending on whether the data represent a snapshot in time of the patient's condition at diagnosis, or evolution of the disease over time in a longitudinal cohort study. Either way, the equivalent of the linear argument *ß.x* in a Generalised Linear Model defines a prognostic index that ranks patient data by severity of the illness. In the case of breast cancer, typically a piecewise linear model is used [1] from which the prognostic index can be derived. A good example of this is the Nottingham Prognostic Index (NPI) which is widely used in clinical practice [2] and takes the form: *NPI score = 0.2*Tumour size (cm) + Node Stage (1...3) + Histological Grade (1...3).*

The same principles apply when flexible models are used, such as generic non-linear algorithms including artificial neural networks.

In the case of discrete time models of longitudinal data, the main variable is the event rate, also called the hazard rate

$$h(x_p, t_k) = P(t \le t_k \mid t > t_{k-1}, x_i) \tag{1}$$

which is the probability that patient p with characteristics x_p survives to the end of time interval t_k given that the patient is known to have entered that interval without

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 214–221, 2008.

experiencing the event of interest. This is the output of the Partial Logistic Artificial Neural Network (PLANN) model [3]. This index is usually averaged over time to define a time-independent risk score over a predefined time interval.

The equivalent index to the NPI score is now the log-odds ratio

$$\textit{PLANN Risk Index} = \log\left(\frac{h(x_p,t_k)}{\left(1-h(x_p,t_k)\right)}\right) \tag{2}$$

where the conditional probability of class membership is directly estimated by the neural network output.

Once the risk score is defined, the population of patients at risk needs to be stratified for the purpose of tailoring adjuvant therapy and to enable comparisons between to be made between patient cohorts from different clinical centres, or subject to different clinical interventions, to be made between patients at similar risk by outcome. In survival analysis the most widely used statistic to test significant differences is the log-rank statistic.

The next section reviews current practice in the application of the log-rank test to stratification of patient data. This is followed by a case study of prognostic modelling of data from patients with operable breast cancer, comparing a statistical methodology (Cox regression) with PLANN-ARD [4] that uses the ARD framework to regularise the PLANN model and performed well in comparison with alternatives in a recent double blind benchmark of linear and generic non-linear survival models [5]. In section 4 the prognostic scores obtained on the same data by these two methods are stratified, comparing the standard application of the log-rank test with a novel methodology proposed in this paper, to resolve significant issues of robustness in the allocation of patients into risk groups.

2 Application of the Log-Rank Test to Stratification of Patient Data

In the literature the approach to splitting risk indices into risk groups is not always stated clearly, sometimes stating the cut-off points of the respective risk scores without a clear indication of how these were obtained [6-7]. Where the split of the indices is at all explained, expert knowledge has been a factor as in the case for the widely used NPI. This index is designed for ease of use and is derived by rounding a more cumbersome Cox regression calculation, the cut-off points being chosen to best match the risk groups from the original model which, in turn, was split on the basis of best match with known clinical groups.

In another approach the indices are split into equal sized groups as suggested by Harrell et al [8]. This tutorial in biostatistics suggests using deciles as a starting choice and in a prognostic model for ovarian cancer Clark et al [9] used quartiles to partition the risk score.

A suggestion for an automated method is to use successive top-down splits by maximising the log-rank test statistic [10].

3 Prognostic Modelling of Breast Cancer Patients

The reference data set for this case study consists of routinely acquired clinical records for patients recruited by Christie Hospital, Manchester, during 1983-89. The specific cohort of interest is patients with early, or operable breast cancer, defined 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. This defines a case series (n=917) for a longitudinal cohort study with 5-year follow-up. The date of recruitment is the date of surgery and the event of interest is cancer specific mortality.

Earlier studies identified the following six predictive variables: age at diagnosis, node stage, histological type (lobular, ductal or in situ), ratio of axillary nodes affected to axilar nodes removed, pathological size (i.e. tumour size in cm) and oestrogen receptor count. All of these variables are banded and binary coded as 1-from-N. For details of the attribute assignment see [4].

Two analytical models were fitted to the data, starting with a piecewise linear model Cox regression, also known as proportional hazards. This model factorises dependence on time and the covariates, modelling the hazard rate for patient with clinical characteristics x_p at time t_k as follows:

$$\frac{h(x_p,t_k)}{1-h(x_p,t_k)} = \frac{h_0(t_k)}{1-h_0(t_k)}.\exp\left(\sum_{i=1}^{N_i}\beta x_i\right) \tag{3}$$

where h_0 denotes the empirical hazard for a reference population with covariate attributes all equal to zero and x_i is the static covariate vector. This was taken to be the standard used by the software that implemented the piecewise linear model Cox regression (SAS), which is the last attribute for each covariate. It was found that the risk group allocation is not sensitive to the choice of reference population. In contrast, the PLANN model is semi-parametric and with the following model for the hazard.

$$\frac{h(x_p,t_k)}{1-h(x_p,t_k)} = \exp\left(\sum_{h=1}^{N_h} w_h \cdot \;\; g(\sum_{i=1}^{N_i} w_{ih}.x_{pi} + w.t_k + b_h) \;\; +b\right) \tag{4}$$

where the indices i and h denote the input and hidden node layers and the non linear function $g(\cdot)$ is a sigmoid.

Both models optimise the same objective function, namely the log-likelihood summed over the observed status of the patient sampled over of 60 months, with an indicator variable that is 1 for death attributed to breast cancer and 0 if the patient is observed alive. Using target values τ_{pk} as indicator labels and t_l as the time index

$$G = -\sum_{p=1}^{No.patients}\sum_{k=1}^{t_l}\left[\tau_{pk}\log\left(h\left(x_p,t_k\right)\right)+\left(1-\tau_{pk}\right)\log\left(1-h\left(x_p,t_k\right)\right)\right] \tag{5}$$

Data for patients who are lost to follow-up are said to be right-censored and the patients no longer counts as part of the set of patients at risk.

In addition, over-fitting is avoided by regularisation, in the case of Cox regression using Akaike's Information Criterion (AIC) and in the neural network model with Automatic Relevance Determination (ARD) [11].

The prognostic index is defined in both cases as the covariate dependent term in eq. (2) and they are compared for the two models in fig. 1.

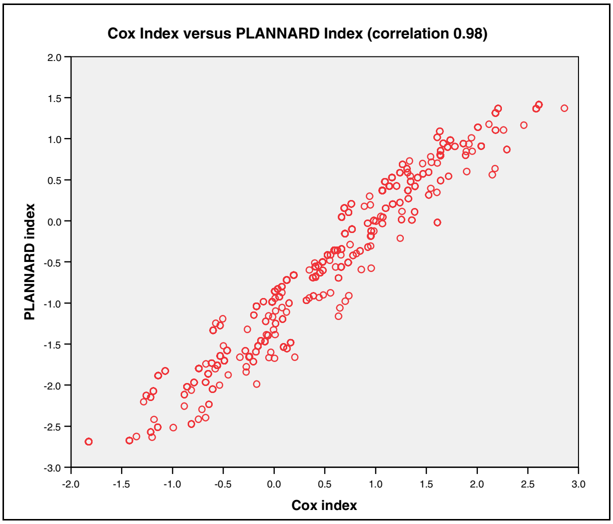

Fig. 1. Correlation between the prognostic derived with Cox regression and PLANN-ARD for a 5-year study of a patient cohort with early breast cancer. The proportionality of the hazards is borne out by the high correlation between the methods. In general, higher risk cohorts, longer follow-up times, studies of recurrence and models for other diseases will only be suitable for piecewise linear modelling if the proportionality of the hazards is observed.

The next stage in the modelling process is to use the prognostic index as the basis to partition the patient cohort into groups at similar risk of adverse outcome. This is the subject of the next section.

4 Robust Methodology for Stratification of Severity of Risk

The most widely used method for stratification of an empirical distribution of prognostic indices is to apply the log-rank test statistic from which the statistical significance for pairwise data partitions can be measured. Given that the test only applies in a pairwise manner, that is to say, for separating two cohorts at a time, this requires a search for the most appropriate threshold to divide the distribution of prognostic index scores.

An accepted strategy was implemented in SAS. It starts by sorting all the records by the value of the prognostic index. Next, the total data are divided into two groups at a threshold value that sweeps the full range of prognostic indices from minimum to

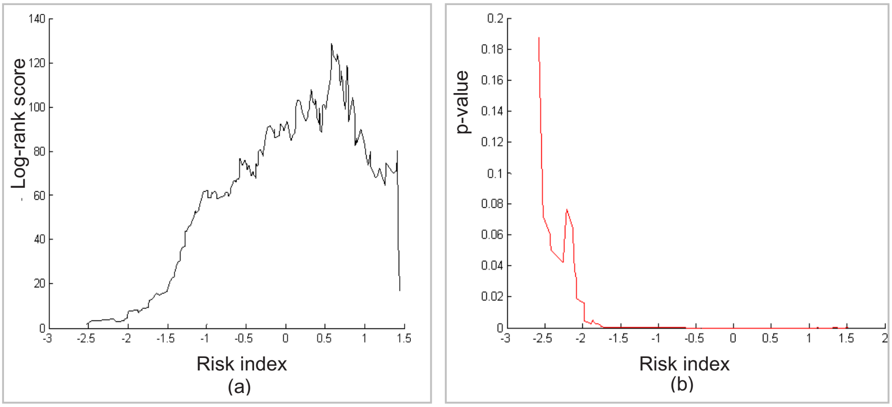

Fig. 2. The significance of data partitions in the top-down approach that is generally applied to stratify patient data in medical statistics detects the global maximum in (a), but this does not take into account that the statistical significance is high for a wider range of possible cut-off thresholds as shown in (b)

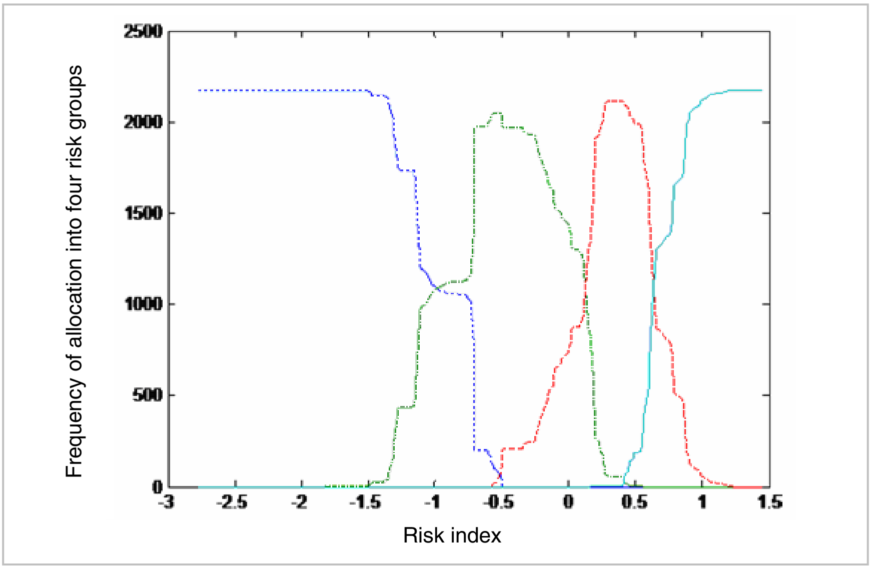

Fig. 3. Results from the proposed methodology for allocating risk groups from a risk index for severity of illness. Note that the group allocation frequencies vary smoothly, in contrast with the spot values of the log-rank test statistic in fig. 2.

maximum. For each threshold, the *log-rank statistic* is calculated and hence a *p-value* results. The maximum of the log-rank statistic determines the first cut-off point. The same method is then repeated in each of the separated cohorts until no further

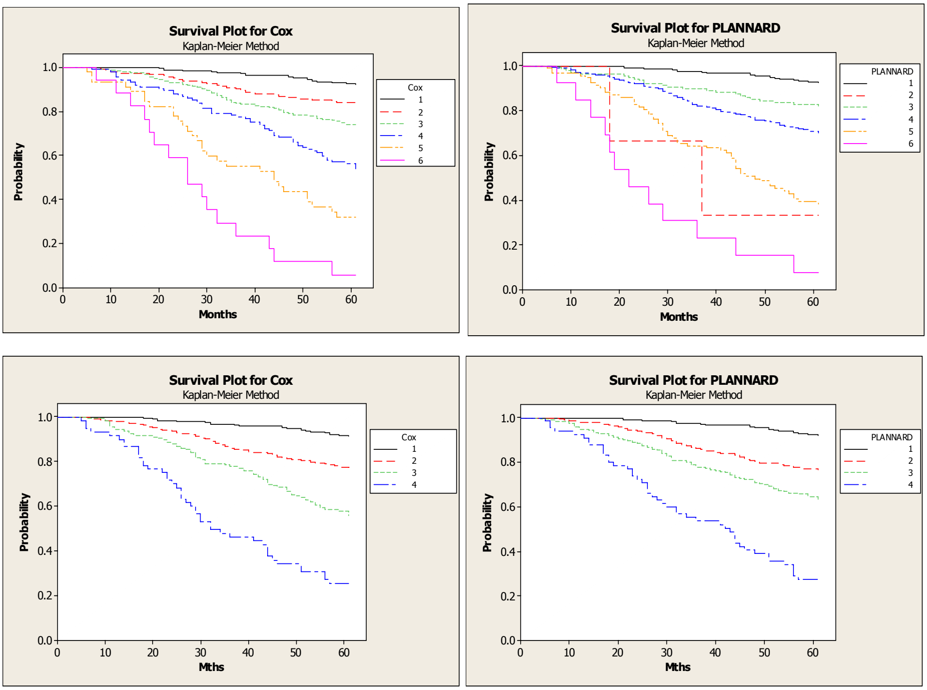

Fig. 4. Actuarial estimates of survival obtained with the Kaplan-Meier method, for the same cases (n=917), stratified using the log-rank test over a 60 month period. The top row uses the standard method and the bottom row uses the proposed method for increasing robustness in the risk stratification. The left column uses Cox regression modelling and the right column the PLANN-ARD neural network. The two modelling algorithms should be consistent, shown in fig. 1 but this is only apparent when the bootstrap method was applied.

partitioning exceeds a pre-set confidence level which, for this study, is as p-value of 0.01 (99% of confidence), corresponding to a test statistic value of around seven.

In practice, the test statistic very much exceeds this value across a wide range of thresholds with the associated p-values forming a plateau indicating that there are a wide range of candidate cutpoints in addition to the maximum log rank statistic that has been selected as can be seen in fig. 2.

A new methodology is proposed to make the stratification of risk indices more robust. The new approach is bottom-up according to the following procedure which involves two nested loops:

Inner loop

i. Bin the risk indices into discrete intervals each containing a minimum number of cases (e.g. n_{min}=10).
ii. Calculate the log-rank statistic for each pair of adjacent cells and aggregate together the two cells with the smallest value of this test statistic.

iii. Repeat the process until the long-rank statistic is significant for all remaining cell pairs. *Outer loop*

i. Draw a sample of the risk indices, with replacement, of size equal to the original data size – this is a bootstrap re-sample of the data.
ii. Apply the *inner loop* to convergence using the re-sampled data.
iii. Allocate each value in the full range of the risk index to a risk group, from 1..N_{groups}
iv. Repeat from *i.* a given number of times (e.g. $n_{\text{resamples}}$ =3000)
v. Identify the distribution of values of N_{groups} and discard all group assignments different from the mode of this distribution.
vi. For each value in the full range of the risk index, build a distribution of risk group allocations – this clearly indicates the cases that fit firmly into a risk group and those that are near the boundary between adjacent groups.
vii. Allocate each case in the original sample to the mode of the distribution of risk groups.

In the current case study, the risk group distributions obtained by this method are plotted in fig. 3.

The robustness of this approach to risk group identification is illustrated in fig. 4. The small size sample causing the unexpected outcome profiles in the solution with 6 risk groups may be an indication that this methodology is over-fitted to the training data.

5 Conclusions

The application of the log-rank test statistic to stratify patients by risk of adverse outcome is subject to variability due to the high prevalence of similar scores for many different risk thresholds. This results in unstable boundaries between strata, causing unwanted variability in allocation of patients into risk groups.

This paper proposes a robust methodology for risk group allocation which exploits bootstrap re-sampling in order to stabilise the distribution of risk groups predicted for each value of the risk score index. The effectiveness and robustness of this methodology are shown by reference to a case study for operable breast cancer, using data from a longitudinal cohort study with 5-year follow-up.

In addition, the generic applicability of the proposed methodology is illustrated using both piecewise linear and neural network models of survival. While the results are consistent with earlier studies of the same data, the current findings are regarded as definitive on account of the robustness that has been added to the stratification process.

Acknowledgments. This work was carried out with partial funding from the Biopattern Network of Excellence FP6/2002/IST/1; N° IST-2002-508803 (www.biopattern.org) and from the Fundação para a Ciência e Tecnologia, through the POS_C(SFRH/BD/30260/2006). Availability of data by Christie Hospital is gratefully acknowledged.

References

1. Cox, D.R.: Regression models and life tables. Journal of the Royal Statistical Society, B 74, 187–220 (1972)
2. 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. British Journal of Cancer 45, 3621 (1982)
3. Biganzoli, E., Boracchi, P., Mariani, L., Marubini, E.: Feed forward neural networks for the analysis of censored survival data: A partial logistic regression approach. Statistics in Medicine 17, 1169–1186 (1998)
4. 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)
5. Taktak, A., Antolini, L., Aung, M., Boracchi, P., Campbell, I., Damato, B., Ifeachor, E., Lama, N., Lisboa, P., Setzkorn, C., Stalbovskaya, V., Biganzoli, E.: Double-blind evaluation and benchmarking of survival models in a multi-centre study. Comput. Biol. Med. 37, 8 (2007)
6. Guerra, I., Algorta, J., Diaz de Otazu, R., Pelayo, A., Farina, J.: Immunohistochemical prognostic index for breast cancer in young women. J. Clin. Pathol: Mol. Pathol. 56, 323–327 (2003)
7. Ortiz Sebastian, S., Rodrıguez Gonzalez, J.M., Parilla Paricio, P., Sola Perez, J., Perez Flores, D., Pinero Madrona, A., Ramirez Romero, P., Tebar, F.J.: Papillary Thyroid Carcinoma: Prognostic Index for Survival Including the Histological Variety. Arch. Surg. 135 (March 2000)
8. Harrell, F.E., Lee, K.L., Mark, B.D.: Tutorial in Biostatistics Multivariate Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. Statistics in Medicine 15, 361–387 (1996)
9. Clark, T.G., Stewart, M.E., Altman, D.G., Gabra, H., Smyth, J.F.: A prognostic model for ovarian cancer. Br. J. Cancer. 85, 944–952 (2001)
10. Williams, B.A., Mandrekar, J.N., Mandrekar, S.J., Cha, S.S., Furth, A.F.: Finding Optimal Cutpoints for Continuous Covariates with Binary and Time-to-Event Outcomes. Technical Report Series #79, Mayo Clinic, Rochester, Minnesota (June 2006)
11. MacKay, D.J.C.: Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. Network: Computation in Neural Systems 6, 469–505 (1995)