

Profiling Prognostic Groups of Breast Cancer Patients Using Clustering

J. D. Martin
I. Jarman
P. J. G. Lisboa
T. A. Etchells

In this paper we use clustering methods to profile a data base of breast cancer patients, as a basis to address two questions: first, whether naturally occurring clusters map onto significantly different survival characteristics; and second, whether conventionally generated prognostic groups comprise heterogeneous populations which can be profiled by the cluster analysis.

PROFILING PROGNOSTIC GROUPS OF BREAST CANCER PATIENTS USING CLUSTERING

J.D. Martín¹, I. Jarman², P.J.G. Lisboa², T.A. Etchells²

¹ Grup de Processament Digital de Senyals, Departament d'Enginyeria Electrònica, Universitat de València, Spain

² The Neural Computation Research Group, School of Computing and Mathematical Sciences, Liverpool John Moores University, United Kingdom

Jose.D.Martin@uv.es, I.H.Jarman@livjm.ac.uk, P.J.Lisboa@livjm.ac.uk

Abstract: In this paper we use clustering methods to profile a data base of breast cancer patients, as a basis to address two questions: first, whether naturally occurring clusters map onto significantly different survival characteristics; and second, whether conventionally generated prognostic groups comprise heterogeneous populations which can be profiled by the cluster analysis. This study applies the Adaptive Resonance Theory (ART) neural network to cluster 931 records of patients from Christie Hospital in Wilmslow, Manchester, who were recruited into a monthly cohort study with 5-year follow-up. An answer to the first question is provided by a Kaplan-Meier plot of the mean survival by cluster, finding that clusters do represent populations with markedly different survival characteristics. The second question is addressed by cross-matching the profile generated by ART against a severity of illness indicator that is internationally used, the Nottingham Prognostic Index (NPI); against other indices derived specifically for these data using Cox regression and the Partial Logistic Artificial Neural Network (PLANN); and against treatment. It is concluded the distribution of input variables among the different clusters is characterised by two explanatory variables, the number of oestrogen receptors and the ratio of axilla nodes affected to nodes removed.

Keywords: breast cancer, prognostic groups, clustering, adaptive resonance theory, survival analysis.

INTRODUCTION

Prognostic groups are normally allocated by survival analysis from information about time to mortality or recurrence of a disease. In practice, oncologists frequently use an algorithm commonly referred to as the Nottingham Prognostic Index (NPI) [1], which was derived using the classical statistical method of

proportional hazards [2] a linear in the parameters approach to the modelling of the censored data.

This paper investigates the characteristics of naturally occurring segments of patient data, and whether they align with the prognostic groups derived by survival analysis.

Patient records are clustered using the Adaptive Resonance Theory (ART) neural network [3], starting by identifying whether patients in different clusters have different expectations of survival. The composition of the clusters is then cross-matched against an array of prognostic indexing methods including NPI, Cox Regression, and the Partial Logistic Artificial Neural Network (PLANN) [4]. The main goal of this work is to analyse whether patients in different clusters but within the same prognostic group are different and, if so, to map this heterogeneity thus finding out whether these differences in patient characteristics affect survival or choice of treatment.

The suitability of ART to solve this problem comes from its two main advantages with respect to other clustering techniques: it overcomes the *stability-plasticity* dilemma, and it does not need to know the number of clusters in advance. The former refers to the ability of ART to adapt clusters to new data without disrupting the already established clusters; in practice, this works by identifying the most appropriate cluster for a given user pattern, then testing whether the cluster prototype is a good-enough representation of the user pattern and, from this, adapting that cluster or starting a new one. In addition, an advantage of this algorithm is that it is not necessary to know the number of clusters in advance; once the degree of similarity is chosen, the algorithm finds the number of clusters corresponding to this choice.

Clustering obtained by ART is utilized to carry out a

knowledge discovery stage, based on analysing the distribution of the different clusters. This analysis can help in achieving the goal previously quoted, i.e., to show similarities and differences among patients who are classified in different clusters, but nevertheless, belong to the same prognostic group. In addition, the common characteristics of patients in a certain cluster can also be observed.

The rest of the paper is outlined as follows: in the next section, data description is shown. Afterwards, ART is described. Results achieved by ART and comparison with other techniques precedes the conclusions of the study and our proposals for further work.

DATA DESCRIPTION

Data used in this work consists of 931 records of patients collected at the Manchester Christie Hospital from 1990 to 1993, in a monthly cohort study with 5-year follow-up. All these patients belong to the group of “low risk”; it means that tumours were smaller than 5 cm, with at most a few mobile affected lymph nodes and with no detectable metastatic spread. The event of interest was defined to be death attributed to breast cancer. Other causes of death or other losses to follow-up were considered as censorship. Censorship is an inherent feature of survival data that arises when follow-up stops before the end of the study period. An example of censorship is an intercurrent death; it means that a patient dies within the follow-up period but from a cause not attributed to breast cancer.

Table 1 shows the 15 explanatory variables recorded for each patient, all of which are categorical and were therefore represented with one-from-N coding. There are a considerable number of missing values but, as the distribution of missing data is not expected to be random, they were coded as an additional attribute. For instance, where the key predictive variable nodes affected is missing this tends to indicate extremely good survival, to be expected of patients for whom axillary clearance was not considered to be required.

The patient data were clustered using the six fields found to have been predictive of survival in a preliminary study using a proportion hazards model with a forward selection stepwise procedure and Akaike’s information criterion (AIC), to measure the significance of adding each variable to the model. Six variables were selected, namely, *age*, *clinical stage nodes*, *histology*, *node ratio*, *pathological size* and *ER status*, which agree with the variables selected in a previous study [4] but adding two more variables, age and ER status. These variables formed a 23-dimensional attribute vector with all binary entries.

Variable	Categories
Menopausal Status	Pre-menopausal Peri-menopausal Post-menopausal
Age Group	20-39 40-59 60+
Predominant Site	Upper Outer Lower Outer Upper Inner Lower Inner Subareolar
Side	Right Left
Maximum Diameter of Tumour	<2cm 2-5cm 5+cm Unknown
Clinical Stage Tumour	T0 (No Tumour) T1 (Tumour < 2cm) T2 (2-5cm) T3 (5+cm) T4 (any size but fixed on the rib cage)
Clinical Stage Nodes	N0 (no nodes found clinically, or node negative by histology) N1 (ipsilateral and mobile axillary nodes) N2 (nodes fixed) N3 (nodes fixed and cannot be removed)
Metastasis Stage	M0 (no distant metastasis) M1 (positive)
Manchester Stage	0 1 2 3 4
Histology	Inv. Duct Inv. Lob/Lob in situ In Situ / Mixed / Medullary / Ucooid / Papillary / Tubular / other Mixed in Situ Unknown
Number of Nodes involved	0 1-3 4+ 98 (too many to count) Unknown
Number of Nodes removed	0-9 10-19 20+ 98 (too many to count) Unknown
Node Ratio	0-20% >20-40% >40-60% 60+% Unknown
Pathological Size	<2cm 2-5cm 5+cm Unknown
ER Status (Oestrogen)	0-10 10+ 8888 (high positive value) Unknown

Table 1: Variables recorded and the attributes description for each one.

ADAPTIVE RESONANCE THEORY

The ART model was originally proposed to model fast adaptive learning in the initial stages of human visual processing [3]. Hence it is termed an artificial neural network. In its initial form, ART1, the model applied only to clustering of binary vectors. It remains among few clustering methods specifically designed for quantized data. The model was then extended to continuous-valued vectors in ART2. These networks cluster inputs by using unsupervised learning. Each time a pattern is presented, an appropriate cluster unit is chosen, and that cluster's weights are adjusted to let the cluster unit learn the pattern. The weights on a cluster unit are considered to be a prototype for the patterns assigned to that cluster.

As a computational tool, ART networks allow the user to control the degree of similarity of patterns placed on the same cluster; once this choice is done, it is not necessary to choose the number of clusters in advance, but the network finds the number corresponding to the degree of similarity chosen.

During training, each data pattern is presented several times. A pattern may be placed on one cluster unit the first time it is presented and then placed on a different cluster when it is presented later (due to changes in the weights for the first cluster if it has learned other patterns in the meantime). A stable network will not return a pattern to a previous cluster, i.e., a pattern oscillating among different cluster units at different stages of training indicates an unstable network.

Some self-organized neural network models achieve stability by gradually reducing the learning rate as the same set of training patterns is presented many times [5]. However, this does not enable the network to learn rapidly a new pattern that is presented for the first time after a number of training epochs have already taken place. The ability of a network to respond to a new pattern equally well at any stage of learning is called plasticity. ART networks are designed to be both stable and plastic.

RESULTS

Clustering of patients

ART2 yielded three different groups to describe the data set. These groups consisted of 488 (C1), 358 (C2) and 85 patients (C3), respectively. A survival analysis of these groups (Fig. 1) revealed that there was a slight difference in terms of survival between two of the clusters (C1 and C2), both presenting a high cumulative of survival for the 5 year follow-up,

whereas C3 showed a behaviour completely different, with much lower values for the cumulative of survival. In spite of the similarity between C1 and C2, if 95% Confidence Intervals (95% CI) are taken into account, there is just a very slight overlap between them. Therefore, the three clusters found corresponded to three different behaviours of patients in terms of survival.

The clear separation of the different clusters showed the suitability of ART2 clustering to describe the different survival characteristics of breast cancer patients. It should be emphasized that clustering did not use survival information at any time. Nevertheless, this group of operable patients shows a remarkable variation in survival just from their clustering characteristics.

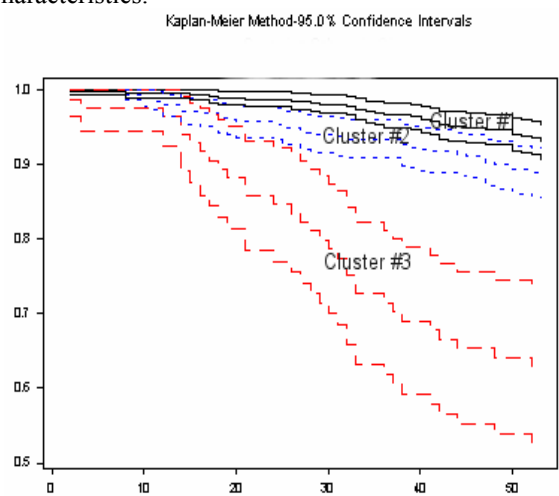


Fig. 1 Kaplan-Meier estimated survival curves, with 95% Confidence Intervals (95% CI), for the three clusters found by ART2. The x-axis represents the time in months, and the y-axis the cumulative of survival. The curve presenting the highest survival corresponds to C1, whereas that presenting the lowest one corresponds to C3.

Comparison between clustering and prognostic indices

The remarkable separation in survival between the clusters suggests that they may be correlated with survival indices. This was tested by plotting the composition of each cluster in terms of a widely accepted survival indicator, NPI, and two other specific models for these data, one linear in the parameters, Cox regression, the other fitted with a neural network, PLANN, regularised with Automatic Relevance Determination (ARD) implementing McKay's Gaussian approximation of the evidence [4]. The cluster composition by prognostic index is profiled in fig. 2.

Particular interest lies in NPI group 3, which is the

most likely group to have genuine heterogeneity among its constituents. It is clear from the figure that the cluster index correlates vaguely with all three indicators of severity of illness, but there is a considerable mix within each cluster. In particular, NPI has arguably the most specific cluster compositions, with only the 290 patients with NPI=3 being significantly split across all three clusters, and a very clear delineation of the cluster with lowest survival with exclusively high NPI scores.

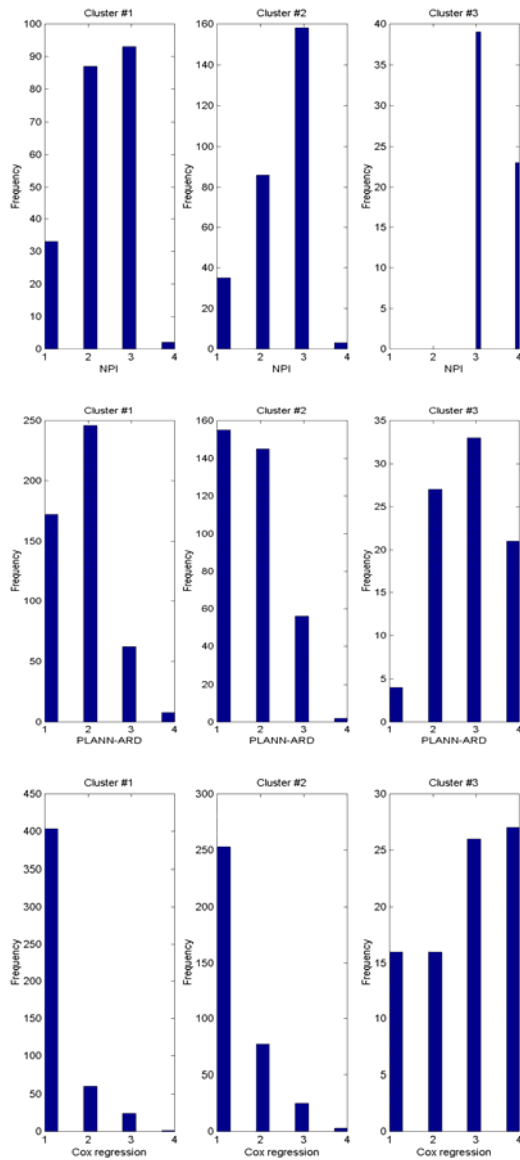


Fig. 2 Histograms showing the frequency for the different clusters of three prognostic indices: NPI, PLANN-ARD and Cox regression. The frequency values of NPI are smaller because there are 372 missing values.

The next stage is to study the distribution of values of the variables in the different clusters, to see whether a relationship can be identified between the prognostic

indices, survival and the choice of treatment.

Cluster profiles for individual variables

A knowledge discovery stage was carried out by analysing whether any of the six input variables peaked for the different clusters. The results are shown in Fig. 3.

We also applied a rule extraction algorithm, the OSRE [6-8] to discover explicit rules that explain the cluster allocation. The rules are shown in table 2a.

A two-dimensional rendition of the cluster composition in terms of the two important explanatory variables identified by the OSRE is shown in Table 2b.

	Specificity	Sensitivity	Rule
Cluster 1	0.982	0.943	(Oestrogen = 3 or 9) and (Node Ratio \neq 4)
	1	0.379	Node Ratio = 9
Cluster 2	1	0.992	(Oestrogen = 1 or 2) and (Node Ratio = 1 or 2)
Cluster 3	0.989	1	Node Ratio = 3 or 4

Table 2a. Rules for each cluster identified using the OSRE, a rule extraction algorithm.

Cluster composition	Nodes ratio					
	1	2	3	4	9	
Oestrogen receptors	1	2	2	3	3	1
	2	2	2	3	3	1
	3	1	1	1 or 3	3	1
	9	1	1	1 or 3	3	1

Table 2b. Cluster labels assigned against just two fields. Attribute '9' denotes 'missing value'. Note that the cluster assignments using this grid accurately describe over 98% of the cluster memberships including all of cluster 3.

In summary, cluster 3 is characterised by high values of the node ratio, cluster 2 has low values of this ratio and oestrogen receptor, cluster 1 has a preponderance of missing values for these two variables as only 8 patients, less than 2%, in cluster 1 have category 3 oestrogen receptor. There is also a transition point between clusters 1 and 3 where the rule extraction has a rule for both clusters that describes that position on the matrix.

Turning to the remaining variables fig.3, the distribution of 'pathological size' seems very logical since cluster 1, the cluster that presents the best survival characteristics, is mainly formed by patients whose pathological size equals 1 (tumours smaller than 2 cm). Cluster 3, the worst cluster in terms of survival, is basically constituted by patients with large tumours (between 2 and 5 cm). Finally, cluster 2 distribution is between the other two, but closer to cluster 1. Therefore, this distribution suggests that clustering segmentation describes appropriately the expected relationship between this variable and the survival outcome.

different clusters but in the same prognostic group. As a first approach, we carried out this analysis for the case of NPI=3 for two reasons; firstly, NPI is the most used prognostic index, and secondly, NPI=3 showed an extremely heterogeneous profile in Fig. 2.

The same trend in values of pathological size, nodes stage and nodes ratio, are apparent in fig.5 for NPI=3 as applied in fig. 3 for the full clusters. The remaining explanatory variables showed no particular trends.

So, while meeting the same prognostic group score using NPI, this particular group of patients is heterogeneous. Survival analysis showed that this

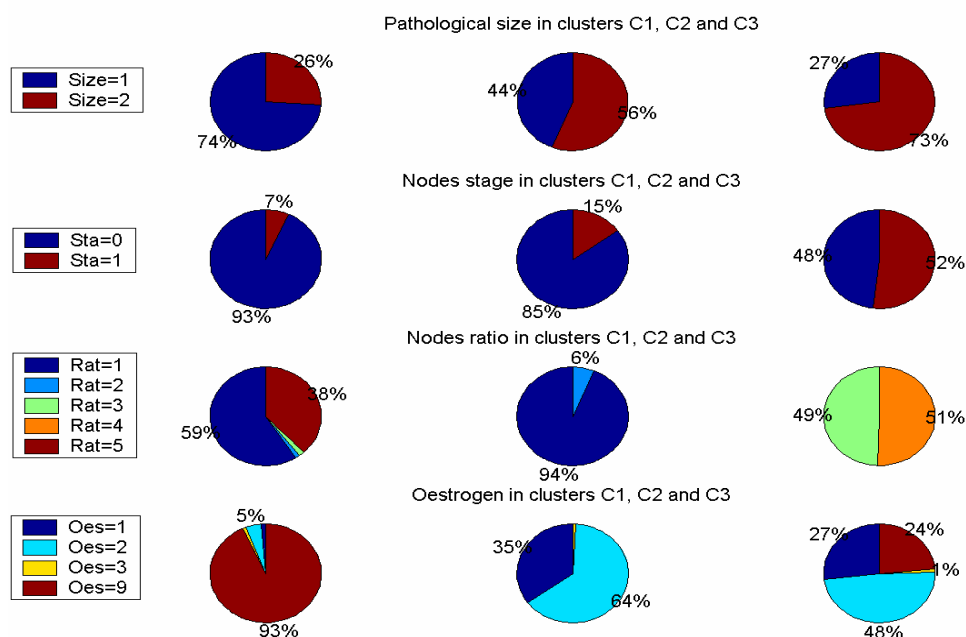


Fig. 3 Pie charts showing the distribution among the three clusters of input variables 'nodes ratio' and 'oestrogen'. Pie charts are referred to clusters C1, C2 and C3, respectively, from left to right.

Similar conclusions can be extracted from the variable 'nodes stage' since cluster 1 is formed mainly by patients who mostly do not present with a palpable spread to the axilla, while for cluster 3 this patients are in the majority.

heterogeneity is reflected in survival, but now the question is if it is also reflected in the choice of treatment

Nevertheless, it is the strong correlation of the cluster indicator label with nodes ratio that goes farthest towards explaining the relative grouped survival behaviour of the different groups, shown earlier in Fig. 1.

Within-cluster survival

Exactly how well correlated the clusters are with prognostic indices can be explored by mapping the two outcome variables of interest, namely:

Input variables' profile in a particular case: patients with NPI=3

- survival within each prognostic group, separating patients by cluster, and
- the distribution of treatments in the same matrix of cluster vs. prognostic index.

It is particularly interesting to analyse the distribution of input variables when comparing clustering with prognostic indices. It can help in discovering which are the differences between the profile of patients in

It is of interest to compare the survival of patients within the same cluster, but with different prognostic index allocations (fig.4). The results are surprising, indicating that the variability in survival by cluster

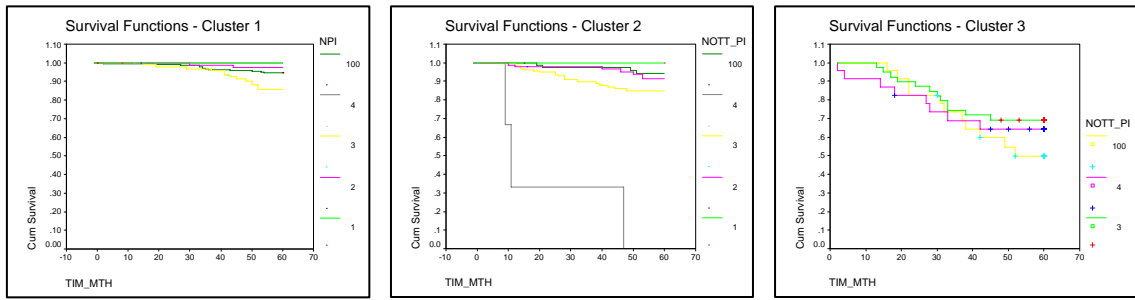


Fig.4 Kaplan-Meier curves showing estimated survival among the three clusters for the Nottingham Prognostic Index, a clinically accepted proportional hazards model. The missing category (100), for NPI with 372 patients, have been included to demonstrate the narrow range of the survival curves within a particular cluster. The only exception being in cluster 2 for NPI group 4 representing three patients. For clarity, confidence intervals have been omitted for the respective survival curves.

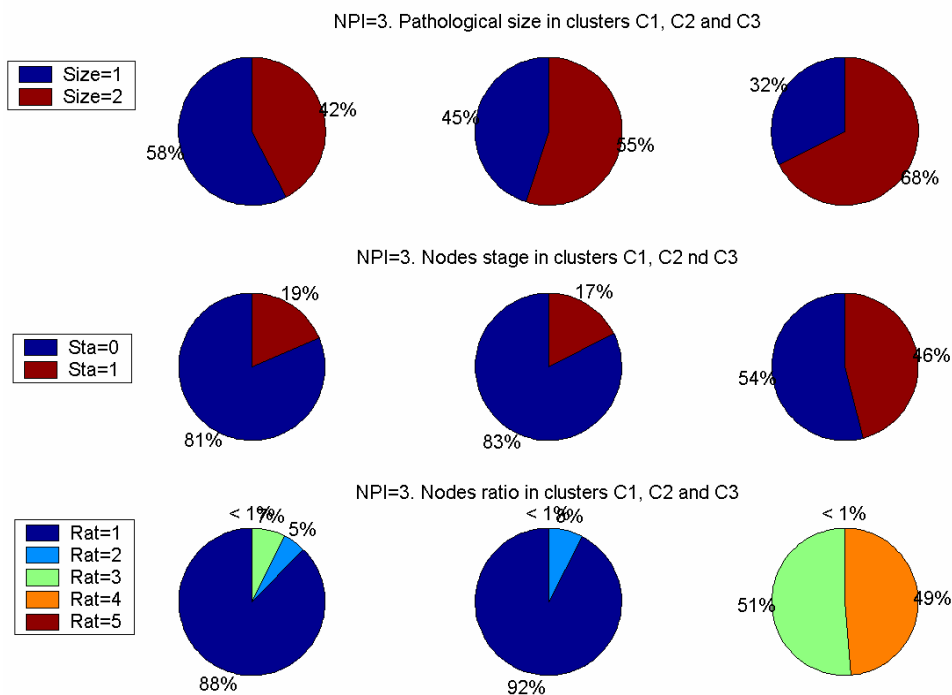


Fig. 5 Pie charts showing the distribution among the three clusters of input variables ‘pathological size’, ‘nodes stage’ and ‘nodes ratio’ for patients with NPI=3. Pie charts are referred to clusters C1, C2 and C3, respectively, from left to right.

dominates the variability by NPI score, with the exception of a small group of three patients NPI 4 in C2.

Treatment profile

The patient database does include information on treatment, of which there are the 4 categories represented in table 3.

Treatment	Label	No. of patients
No Treatment	0	424
Chemotherapy	1	101
Hormone Therapy	2	405
Chemo & Hormone	3	1

Table 3: Treatments administered to the patients of the data set, label used in the data set to code it and number of patients following each treatment.

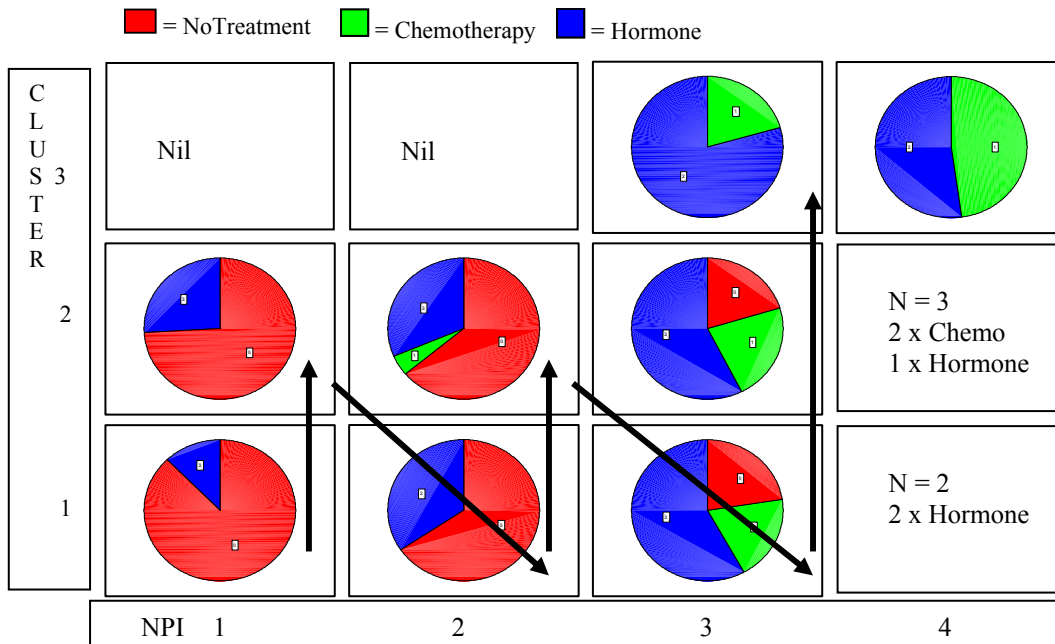


Fig. 6a Pie charts showing patient treatment within a matrix of clusters compared to NPI. This demonstrates the homogeneity of treatment for each cluster, plots in rows correspond to clusters segregated into an NPI survival group.

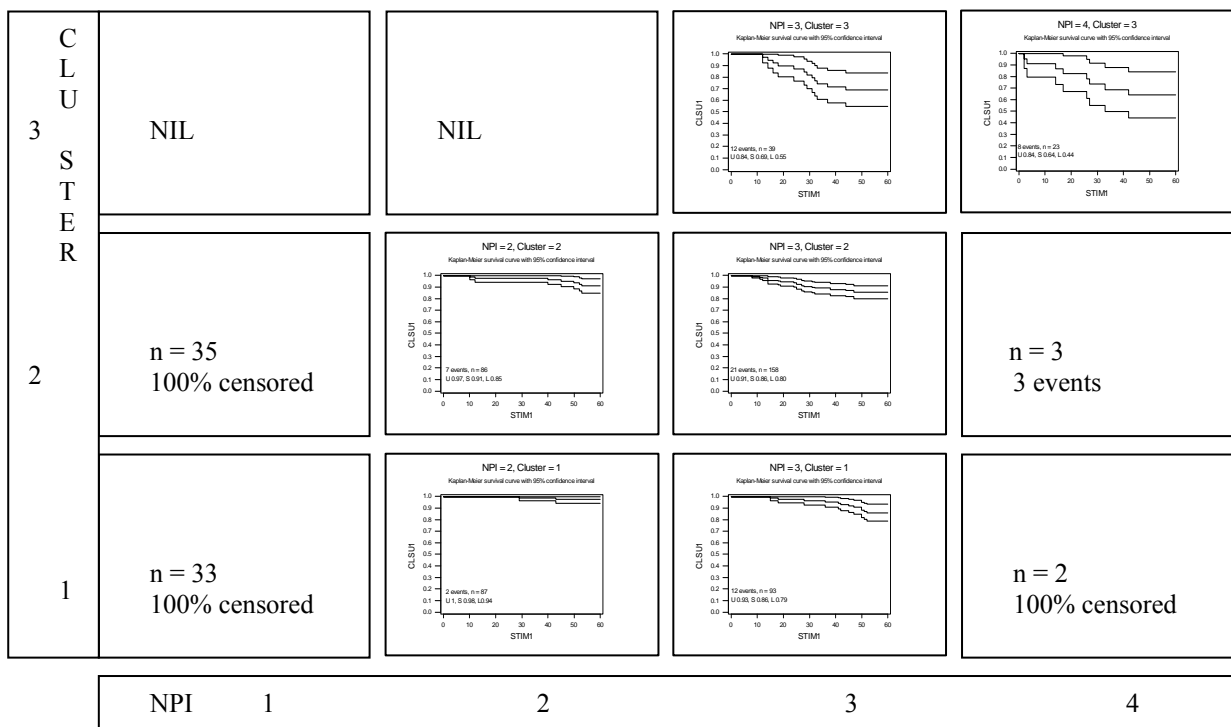


Fig.6b Kaplan-Meier survival curves with 95% confidence intervals showing patient survival within a matrix of clusters compared to NPI. This demonstrates the homogeneity of survival for each cluster, plots in rows correspond to clusters segregated into an NPI survival group.

The pie charts in fig. 6a show the distribution of treatments in a matrix of cluster vs. NPI. It should be noted that we were only able to derive NPI for 559 out of the 931 patients, and figures 6a and 6b reflect this.

Inspecting each row and column in fig 6a, it is apparent that treatment is more consistent for each prognostic index than within each cluster. However, there is a clear

progression in choice of treatment, indicated by the arrows.

This progression, starting from cluster1 and NPI 1, shows a decrease in patients receiving no treatment with increasing treatment of hormone therapy and chemotherapy. The way in which clustering is reflected

in treatment seems logical, since patients who do not receive any of the treatments considered, belong to those combinations of cluster vs. NPI whose survival is highest.

A rational development was to reproduce the same matrix for survival curves that correspond directly to the treatment profiles at each point. Following the same progression as for treatment, we see a gradual decrease in survival as we proceed through the sequence.

By using ART2 in conjunction with NPI we can see gradual treatment changes in the cells of the cross-matching matrix in fig 6a, apparently disaggregating in a meaningful way the patient cohorts allocated to individual prognostic groups. This shows that the information from clustering is complementary to that contained in the supervised prognostic risk model.

CONCLUSIONS

An ART2 clustering approach to profile prognostic groups of breast cancer patients has been proposed in order to identify natural groupings among breast cancer patients presenting for treatment, and to highlight any heterogeneity among patients in the same prognostic group as determined by conventional survival analysis.

The first surprise was that the self-organised clusters separate well by grouped survival. Each cluster crossed different prognostic groups, but it was found that the clusters are characterized predominantly by the value of nodes ratio, except for the highest surviving group for whom the number of oestrogen receptors has largely not been recorded, and sometimes neither has the nodes ratio. This could be because axillary clearance is not routine for patients with little survival risk.

There is particular interest on whether patients in the large middle ranking group by survival, NPI=3, have similar survival expectancy but perhaps for different reasons, and may therefore require different treatments. Differences among these patients are, in fact, apparent. While clusters presenting with high survival in this prognostic group were very similar in the treatment they received, evident differences appeared when comparing with the cluster with the poorest survival in the same prognostic group. This reveals the heterogeneity of patients within the same prognostic group, reflecting the different survival and treatment needs for patients with different ratios of nodes involved and with different oestrogen receptor status, explained by the cluster allocation table 2b.

This heterogeneity, when considered across all NPI groups, showed a gradual and consistent progression both in terms of treatment and survival shown in figures 6a and 6b.

In conclusion, the cross-matching of clusters and prognostic indices is more specific to survival and choice of treatment than the prognostic index alone, demonstrating the potential of clustering methods to disaggregate the heterogeneity inherent in prognostic groups modelled by survival.

REFERENCES

- [1] Galea M.H., Blamey R.W., Elston C.E. and Ellis I.O. (1992): 'The Nottingham Prognostic Index in primary breast cancer', *Breast Cancer Res Treat*, **22**, pp. 207-219.
- [2] Cox D.R. (1972): 'Regression models and life tables', *JR Stat Soc, B*, **74**, pp. 187-220.
- [3] Carpenter G.A. and Grossberg S. (1991): 'ART2: Self-Organization of Stable Category Recognition Codes for Analog Input Patterns', in: Carpenter G.A. and Grossberg S., editors. *Pattern Recognition by Self-Organizing Neural Networks*, MIT Press, Cambridge, MA, USA.
- [4] Lisboa P.J.G., Wong H., Harris P. and Swindell R. (2003): 'A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer', *Artif Intell Med*, **28**, pp. 1-25.
- [5] Kohonen, T. (1997): 'Self-Organizing Maps', 2nd Ed., Springer-Verlag, Berlin, Germany.
- [6] Etchells, T.A., Lisboa, P.J.G., 'On rule extraction from smooth decision surfaces' NNWSMED/CIMED, Proc. 5th International Conference, pp 23-28, 2003.
- [7] Etchells, T.A. 'Rule extraction from Neural Networks: A practical and efficient approach', unpublished PhD thesis. Liverpool John Moores University, (2003). http://www.cms.livjm.ac.uk/etchells/phd/Etchells_thesis.pdf
- [8] Etchells, T.A., Jarman, I.H. and Lisboa, P.J.G. 'Empirically derived rules for adjuvant chemotherapy in breast cancer treatment' *submitted to this conference*.