

Empirically derived rules for adjuvant chemotherapy in breast cancer treatment

T. A. Etchells

I. Jarman

P. J. G. Lisboa

This paper is concerned with how two specialist breast cancer centres interpret guidelines for treatment and to empirically map any differences between two clinical centres, by modelling treatment records using the Orthogonal Search Rule Extraction (OSRE) algorithm for trained neural networks.

Empirically derived rules for adjuvant chemotherapy in breast cancer treatment

T.A. Etchells, I.H. Jarman and P.J.G. Lisboa
School of Computing and Mathematical Sciences, Byrom Street,
Liverpool L3 3AF, England.

Abstract National treatment guidelines allow clinicians considerable latitude for discretion. This paper is concerned with how two specialist breast cancer centres interpret guidelines for treatment and to empirically map any differences between two clinical centres, by modelling treatment records using the Orthogonal Search Rule Extraction (OSRE) algorithm for trained neural networks [1,2]. The algorithm is enhanced with a refining method to reduce the number of explanatory rules conditional on maintaining sensitivity and sensitivity above minimal acceptable thresholds. Results were obtained by cross-validating on 353 routinely acquired patient records from Christie Hospital near Manchester and 248 patient records from the Clatterbridge Centre for Oncology and the Linda McCartney Centre in Liverpool. While the treatment choices between the two centres appear to be much different, for instance only one centre permitting combination treatment of chemotherapy and hormone therapy, the rules describing the allocation of chemotherapy are consistent between the two centres, with high prevalence of chemotherapy among premenopausal women with oestrogen receptor negative tumours and evidence of spread to the axilla.

Keywords: Rule extraction, neural networks

1. Introduction

Supervised neural networks are frequently used in medical applications as inference models [3]. In this paper, we propose their use for knowledge discovery from data (KDD). More specifically, records in a patient database are utilised to identify Boolean rules that, as far as

is possible, explain the allocation of patients to particular choices of treatment.

This is an important step towards assessing compliance of clinical practice with published good practice guidelines. Moreover, the results obtained are also useful to understand the impact of different treatment regimes on patient survival, which will be presented in a separate paper [10]. The focus on chemotherapy is to provide an illustration of the methodology used, which is generic.

Starting with the database of treatment records, an MLP was fitted to the treatment label. This was followed by the OSRE [1] [2] algorithm, a supervised learning process, as a principled method to empirically derive rules for adjuvant treatment given to breast cancer patients following surgery, to meet the requirements of accuracy in representing decision inferences made by the neural network. In addition, an extension to this method was developed that refines the number of rules that OSRE generates subject to preset parameters in terms of a specificity and sensitivity threshold.

Rules were extracted separately for data from two breast cancer centres, Christie Hospital, near Manchester and a from joint database from the Clatterbridge Centre for Oncology (CCO) and the Linda McCartney Centre, in Liverpool. The rules were compared with the rules derived from a benchmark algorithm, the ID3 model [4]. These rules were also analysed between centres, which will enable us to discover the consistencies between centres governing treatment.

Note that the treatment profiles for the two clinical centres show some significant

differences. For instance, Christie Hospital does not administer combination of chemo and hormone therapies, whereas around 30% of patients at the CCO/Linda McCartney Centre received this combination treatment.

2. Patients and Data

For consistency, data from both hospitals went through 2 stages of a filtering process. The first stage is a filter, described in table 1, the data set from which to extract the rules. The second stage, is a refinement of the inclusion criteria in table 2, to enable a fair comparison between the two centres.

The need for the second stage is due to different patient referral procedures, which apply at the two centres, further explained below.

Any records with tumour stage 0 were not included.	
Inclusion criteria	
Follow up	≥ 5 years
NPI [5]	Not missing
Metastasis	= 0
Node Stage	= 0 or 1
Pathological size	< 2cms or 2-5cms

Table 1. The first stage of filtering gives us the data set, from which rules for treatment were extracted.

1.	NPI = 3 or 4
2.	OR Node stage positive
3.	OR All ER negative

Table 2. Inclusion criteria for comparison of rules between hospitals.

2.1 Christie Hospital

The data consist of 1266 records referred to Christie Hospital between 1990 and 1993. There were 559 records that met with the first inclusion criteria (table 1).

The further adjustment to the data for comparison purposes (table 2), left a total of 353 patients.

2.2 CCO and Linda McCartney Centre

The data consist of 808 patients who were on the patient list of the same oncologist at the two centres, who recorded details from date of diagnosis. The patients were referred to the oncologist based on a poorer prognosis, hence the need for a second filtering of data for comparison. Apart from one patient diagnosed in 1957 the rest were referred between 1975 and 2001, of whom 269 were consistent with the first stage of the inclusion criteria.

Variable	Categories	Labelling
Menopausal Status	Pre-menopausal	1
	Peri-menopausal	2
	Post-menopausal	3
Age Group	20-39	1
	40-59	2
	60+	3
Node Stage	N0 (no nodes found clinically, or node negative by histology)	0
	N1 (ipsilateral and mobile axillary nodes)	1
Histology	Inf Duct	1
	Inf Lob/Lob in situ	2
	In Situ / Mixed / Medullary / Ucoid / Papillary / Tubular / other Mixed in Situ	3
Node Ratio	0-20%	1
	20-30%	2
	40-60%	3
	60+%	4
	Unknown	9
Pathological Size	<2cm	1
	2-5cm	2
Oestrogen (ER Status)	0-10	1
	10+	2
	8888 (high positive value)	3
	Unknown	9
Histological Grade	Well differentiated	1
	Moderately differentiate	2
	Poorly differentiated	3
Nodes involved	0	1
	1-3	2
	4+	3
	98 (too many to count)	4
NPI	Very Good	1
	Good	2
	Moderate	3
	Poor	4

Table 3. Input factors used in the OSRE algorithm with the number of coded attributes.

With the second filter applied, 248 patients remained in the study, who were referred between 1984 and 1998.

The explanatory variables (table 3) selected for this study consist of the variables identified as having predictive value in the design of prognostic indices for survival, from studies involving both linear and non-linear models [6]. In addition, we also include the primary variables used by the NPI [7] and additional variables, such as age and nodes ratio, which preliminary analysis suggested are important for assignment of treatment.

In assigning patients with NPI to the ‘moderate’ risk group we follow the same convention as CCO/Linda McCartney Centre of merging groups 3a and 3b.

3 ID3 Approach

The rule induction method ID3 [4] was chosen as the benchmark for a number of reasons. Firstly we were interested in an approach that used a discrimination tree for classification as opposed to OSRE which uses hyper-cubes to fit the response surface generated by an external model, in our case an MLP.

It also classifies patterns that have categorical and ordinal attributes, which is typical of medical data. Finally, the information can be expressed in the form of explicit rules that will allow a comparison to be made with the OSRE rules.

ID3 uses an information-theoretic approach, so at any level in the decision tree we select the factor with the largest increase in information or comparably the largest decrease in entropy, which is empirically estimated by the proportion p of in-class data in each leaf, using the well-known formula

$$S = - \sum_{c=1}^{No.classes} p_c \log(p_c).$$

4 The OSRE algorithm

The OSRE algorithm finds conjunctive rules for classifications of data using an MLP (or any other smooth response surface). In a high dimensional input space the training data occupies only a small fraction of the total space. Outside these regions where the data are present, the response surface is being extrapolated, which for non-linear models is generally unreliable. The OSRE algorithm uses the space occupied by the data and the space in orthogonal directions from each data point to generate rules from the changes in the response of the network.

Tsukimoto [8, reviewed in 2] developed a scalar Boolean model and showed that we can find the optimal Boolean function that describes the outputs of a network with respect to binary inputs. The Boolean function is constructed from the disjunction of the scalar atoms that have an activation from the network greater than the threshold of the output sigmoid (active atoms). Tsukimoto showed that the disjunction of all active atoms provide the optimal Boolean function for the network, where optimality is defined by the Boolean function that is closest in the Euclidean sense to the scalar logic derived from the network response. However, this process was exponential in complexity with respect to the number of inputs into the network. With independent binary data, the exponentiality of the process can be reduced to a polynomial problem by considering the atoms (and their adjacent atoms) represented by the data that trained the network. This results in a subset of rules that best describe the region occupied by the data.

This methodology transcribes to the RULENEG algorithm developed by Pop et al. [9] for binary data, but if there are categorical or ordinal variables RULENEG cannot be applied to the data [2]. In essence the RULENEG algorithm searches for changes in activations between adjacent Boolean atoms to extract rules. However categorical and ordinal variables, when 1-from-N binary encoded,

form a restricted Boolean space for which adjacent Boolean atoms are, generally, excluded by the coding.

In contrast, the OSRE algorithm searches through permitted Boolean space in order to generate rules from the changes in response of the network. The difference between RULENEG and OSRE is demonstrated in figure 1 and figure 2, where black spheres indicate permitted atoms and the dotted arrows show the search pattern.

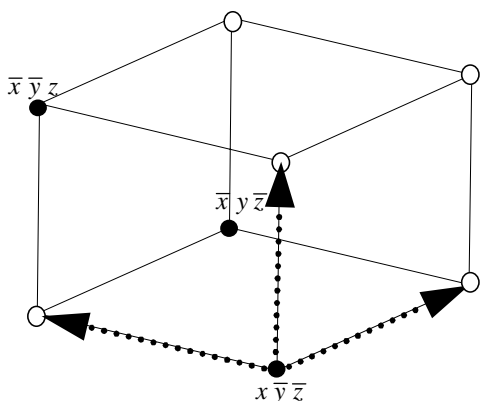


Fig. 1. The RULENEG algorithm inspects nearest neighbours in the atomic hypercube, which violates 1-from-N

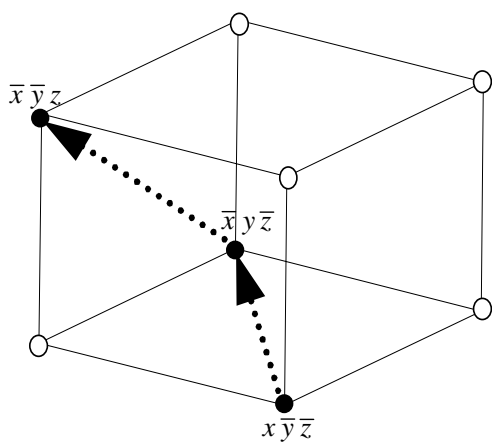


Fig. 2. The OSRE algorithm inspects only the *single variable atomic sub space* of valid codes.

So restricting the search to the data predicted to be within-class and searching in orthogonal directions from the data point reduces the space to a polynomial problem.

In addition the OSRE algorithm overcomes the problem of extrapolating rules beyond the data

space, where fitting a smooth approximation is unreliable, by using the training data to identify the region of the space for which the response surface has been accurately constructed.

4.1 Rule Refinement

When rules are extracted from a dataset with OSRE, the disjunction of all the rules may lead to rules that wholly intersect the disjunction of other rules for a given classification (fig. 3), and is therefore redundant. We introduce a method to reduce the rule set in order to generate a more interpretable explanation of the data, while increasing specificity and controlling any reduction in sensitivity.

Taking the schematic in fig.3 as an illustration, the first rule to drop-out of the set would be R3 as it is fully covered by other rules within the set, the next best candidate for removal would be R4, as the region of data space it spans has the next largest overlap with other existing rules. Algorithmically, this is because the removal of R4 would likely have the least impact on the sensitivity and specificity of the total rule set. Note: Combining rules can only reduce specificity; best sensitivity is achieved from the disjunction of all the individual rules.

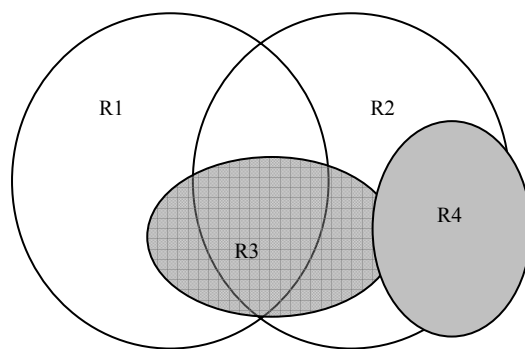


Fig. 3. Rule R3 lies within the data-space of R1 and R2, therefore it is redundant. R4 is a candidate for removal if doing so still meets the required sensitivity and specificity.

To start the rule filtering process we first preset a target value for the false positive rate (1-specificity) for a given rule classification, usually 0.05. Therefore, any individual rule with lower specificity is rejected, as the rule is

then a poor representative of the in-class data since it also identifies an unacceptable proportion of the out-of-class data. This could potentially empty the rule set in the instance that no rule reaches the target specificity. Such a situation could indicate that there are no rules of a conjunctive form that separate the in and out of class data.

In the following description of the algorithm below:

- The global sensitivity and specificity are the sensitivity and specificity of the disjunction of all the rules that remain in the rule set.
- The ROC point is the co-ordinate [1–global specificity, global sensitivity]

The refinement of the rules of high specificity that OSRE extracts is in two stages:

The first stage (steps 1 to 5) eliminates rules that are completely contained within the other rules, for example R3 in figure 3. This first stage finds a smaller set (if it exists) of rules that have exactly the same global sensitivity and specificity as that of the whole rule set.

The second stage involves determining the rules whose removal from the remaining rule set that will increase the global specificity but not reduce the global sensitivity below some predetermined amount, for example R4 in figure 3.

The rules generated by the OSRE are placed in a list call **RuleList**.

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

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

Repeat the process for each rule in the **RuleList**.

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

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

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

(At this stage of the algorithm, R3 in figure 3 would have been identified as redundant and removed)

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

We now have a reduced set of rules, which have the same ROC point as the original set of rules but with any redundant rules filtered out.

Step 6. If the specificity value of the ROC point of these filtered rules has not met a preset target, then remove each rule in turn once again from **RuleSet** and delete the rule that increases the specificity with least cost to sensitivity. Referring to figure 4, the shaded region represents the acceptable region of least cost for a new ROC point; the sensitivity target has been set by the ROC point calculated in step 1. Repeat this process until the specificity target is met.

With reference to fig.4, the theoretical region spanned by the subsets of the complete rule set is the rectangular area. The aim of the rule

optimisation algorithm is to move the ROC point for the rule base to the left of the target specificity threshold, but only if the reduced rule set remains within the circular sector defined by the distance from the ideal classification to the ROC position of the complete rule set, so as not to unduly sacrifice sensitivity.

Using specificity and sensitivity as a vector to measure the distance from [0,1], the maximum [1-specificity, sensitivity] co-ordinate, we have a methodology to give us the least cost movement to the preset threshold (fig.4).

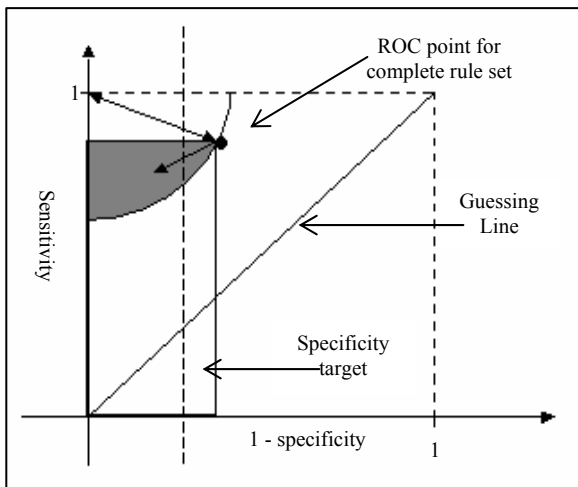


Fig.4. This figure illustrates the purpose of the rule optimisation procedure explained in this text. This is to find the smallest subset of rules for which the distance from the ideal classification in the top left corner of the ROC plot is the same as for the complete rule set.

5 Results

The two algorithms were applied to the two data sets, namely from Christie Hospital and the CCO/Linda McCartney Centre. The rules reported are for choice of chemotherapy treatment, either alone or in combination with another therapy.

5.1 Christie Hospital

As expected, ID3 returned straightforward rules specifying pre-menopausal status and 1-3 or 4+ axillar nodes involved, as listed in table 4.

Specificity	Sensitivity	Rule
0.958	0.564	(Nodes Involved = 2) and (Menopause = 1)
0.988	0.295	(Nodes Involved = 3) and (Menopause = 1)

Global Specificity = 0.946, Global Sensitivity = 0.859

Table 4. Rules for chemotherapy using the ID3 algorithm, extracted from Christie Hospital data, with specificity and sensitivity measures.

In contrast, the OSRE algorithm was applied in a two-stage process, generating first a broad set of overlapping rules, listed in table 5, which were then further refined resulting in the rule set in table 6.

The node ratio, which is the ratio of the number of axillar nodes involved over the number removed, is now preferred to the absolute number of nodes involved, but still in combination with pre-menopausal status. However, there is also a rule that is specific to the younger age group, 20-39 years of age, who receive chemotherapy if at moderate or high risk by NPI score.

Specificity	Sensitivity	Rule
0.992	0.205	(Node Ratio = 2, 3 or 9) and (Menopause = 1)
0.992	0.179	(NPI = 2 or 4) and (Nodes Involved \neq 4)
0.988	0.295	(Nodes Involved = 3) and (Menopause = 1)
0.985	0.359	(Nodes Involved = 2 or 3) and (Grade = 3) and (Oestrogen = 1 or 3) and (Menopause = 1)
0.983	0.346	(Node Ratio \neq 1) and (Menopause = 1)
0.975	0.359	(NPI = 3 or 4) and (Age = 1)
0.975	0.321	(NPI = 3) and (Age = 1)
0.958	0.641	(Nodes Involved = 2 or 3) and (Node Ratio = 1, 3 or 9) and (Menopause = 1)
0.958	0.128	(Grade = 2) and (Oestrogen = 3 or 9) and (Menopause = 1)

Global Specificity = 0.898, Global Sensitivity = 0.897

Table 5. Rules for Chemotherapy using the OSRE algorithm, extracted from the Christie Hospital data, with specificity and sensitivity measures. The highlighted rows are the reduced rule set that have the same coverage as the entire rule set.

Specificity	Sensitivity	Rule
0.983	0.346	(Node Ratio \neq 1) and (Menopause = 1)
0.975	0.359	(NPI = 3 or 4) and (Age = 1)
0.958	0.641	(Nodes Involved = 2 or 3) and (Node Ratio = 1, 3 or 9) and (Menopause = 1)

Global Specificity = 0.933, Global Sensitivity = 0.897

Table 6. After refinement of the rules to reduce specificity whilst controlling sensitivity, we are left with the three rules listed. If the highlighted rule is omitted we are left with the same patient group as in the ID3 algorithm in table 4. However dropping this rule results in a sensitivity value outside the acceptable region (figure 4).

This additional rule may have identified a rule that ID3 has missed. Overall, the rules generated by OSRE have slightly higher sensitivity with marginally lowered specificity, but they do suggest a different clinical interpretation of the same data, which merits further analysis.

5.2 Linda McCartney Centre

Once again the rules generated by the two methods are comparable, but different. Tree-based rule induction identified age and oestrogen receptor status as the main explanatory variables for chemotherapy, while OSRE selected menopausal status, combined with oestrogen receptor status and node ratio.

This time the rules extracted from the trained neural network are fewer, and have greater sensitivity and specificity, than those from ID3.

Specificity	Sensitivity	Rule
0.974	0.110	(Age = 1) and (Oestrogen = 1)
0.954	0.153	(Age = 1) and (Oestrogen = 9)
0.954	0.297	(Age = 2) and (Oestrogen = 1) and (Nodes Involved = 1 or 2)

Global Specificity = 0.881, Global Sensitivity = 0.559

Table 7. Rules for chemotherapy using the ID3 algorithm, extracted from Linda McCartney Centre, with specificity and sensitivity measures.

Specificity	Sensitivity	Rule
0.980	0.137	(Nodes Involved = 2) and (Grade \neq 3) and (Oestrogen \neq 2) and (Menopause \neq 3)
0.974	0.274	(Nodes Involved = 2) and (Oestrogen \neq 2) and (Histology=1) and (Menopause = 1)
0.974	0.137	(Nodes Involved = 2) and (Age = 1)
0.967	0.487	(NPI = 3 or 4) and (Oestrogen \neq 2) and (Node Ratio = 1 or 2) and (Menopause = 1)
0.967	0.222	(Oestrogen \neq 2) and (Node Ratio = 2 or 3) and (Menopause = 1)
0.967	0.137	(Oestrogen \neq 2) and (Node Ratio=3) and (Menopause = 1)
0.960	0.496	(NPI = 3 or 4) and (Nodes Involved = 1 or 2) and (Oestrogen \neq 2) and (Histology = 1) and (Menopause = 1)

Global Specificity = 0.907, Global Sensitivity = 0.709

Table 8. Rules for chemotherapy using the OSRE algorithm, extracted from Linda McCartney Centre, with specificity and sensitivity measures. The highlighted rows are the reduced rule set that have the same coverage as all the rule set.

Specificity	Sensitivity	Rule
0.967	0.487	(NPI = 3 or 4) and (Oestrogen \neq 2) and (Node Ratio = 1 or 2) and (Menopause = 1)
0.967	0.222	(Oestrogen \neq 2) \neq and (Node Ratio = 2 or 3) and (Menopause = 1)

Global Specificity = 0.934, Global Sensitivity = 0.619

Table 9. After refinement of the rules to reduce specificity whilst controlling sensitivity, we are left with the two rules above.

6 Conclusion

Notwithstanding the need to adhere to national clinical guidelines on treatment, the two specialist clinical centres appear to have different therapy regimes but they are consistent with each other with regard to the

administration of chemotherapy. In each case, the empirically derived decision factors are related to young age or pre-menopausal status, oestrogen receptor negative status and evidence of spread to the axilla.

There is an indication that the OSRE method has identified explanatory rules missed by ID3 and the two methods certainly explain the treatment decisions using different variables. Further discussion with clinicians will be required to ascertain which method, if either, is closest to clinical reasoning.

From a statistical point of view, the rules generated by fitting hyper-boxes to a decision surface, OSRE, are either comparable with, or better performing than, those derived by tree-based rule induction using ID3.

7 Acknowledgments

The authors gratefully acknowledge EPSRC for funding this research, as well as R. Swindell of Christie Hospital and S. O'Reilly of the Clatterbridge Centre for Oncology for granting access to patient data.

8 References

1. Etchells, T.A., Lisboa, P.J.G., 'On rule extraction from smooth decision surfaces' NNWSMED/CIMED, Proc. 5th International Conference, pp 23-28, 2003.
2. 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
3. Lisboa, P.J.G. 'A review of evidence of health benefit from artificial neural networks in medical intervention', Neural Networks, Invited Paper, 15, 1, 9-37, 2002.
4. Quinlan, J.R. 'Discovering rules by induction from large classes of examples', Expert Systems in the Microelectronic Age,

pp. 168-201, 1979. Edinburgh: Edinburgh University Press.

5. <http://www.nice.org.uk>
6. 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.
7. J.L. Haybittle et al., 'A Prognostic Index in Primary Breast Cancer', Br. J. Cancer, 1982, 45, 361.
8. Tsukimoto, H., "Extracting rules from trained neural networks", IEEE Transactions on Neural Networks, vol. 11, no. 2, pp. 377-389, March 2000.
9. Pop, E., Hayward, R. and Diederich, J. 'RULENEG: Extracting rules from a trained ANN by stepwise negation' Technical Report, QUT NRC (December, 1994).
10. Jarman, I.H. and Lisboa, P.J.G. 'A comparative study of NPI and PLANN-ARD by prognostic accuracy and treatment allocation for breast cancer patients' *submitted to this conference.*