European Journal of Cancer 40 (2004) 1837–1841

European Journal of Cancer

www.ejconline.com

“Good Old” clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers ☆

Patrik Edén [b], Cecilia Ritz [b], Carsten Rose [a], Mårten Fernö [a,*], Carsten Peterson [b]

[a] *Department of Oncology, Jubileum Institute, Lund University, SE-22185 Lund, Sweden*
[b] *Complex Systems Division, Department of Theoretical Physics, Lund University, Sölvegatan 14A, SE-22362 Lund, Sweden*

Received 12 November 2003; received in revised form 24 February 2004; accepted 25 February 2004
Available online 19 May 2004

Abstract

We compared the power of gene expression measurements with that of conventional prognostic markers, i.e., clinical, histopathological, and cell biological parameters, for predicting distant metastases in breast cancer patients using both established prognostic indices (e.g., the Nottingham Prognostic Index (NPI)) and novel combinations of conventional markers. We used publicly available data on 97 patients, and the performance of metastasis prediction was represented by receiver operating characteristic (ROC) areas and Kaplan–Meier plots. The gene expression profiler did not perform noticeably better than indices constructed from the clinical variables, e.g., the well established NPI. When analysing separately subgroups, according to the oestrogen receptor (ER) status both approaches could predict clinical outcome more easily for the ER-positive than for the ER-negative cohort. Given the time it may take before microarray processing is used worldwide, particularly due to the costs and the lack of standards, it is important to pursue research using conventional markers. Our analysis suggests that it might be possible to improve the combination of different conventional prognostic markers into one prognostic index.

Keywords: Breast cancer; Prognosis; Metastases; cDNA microarray; Gene expression; Histopathology; Cell biology; Oestrogen receptor; Artificial neural network

1. Introduction

Breast cancer patients can be cured either by local regional therapy alone or by the addition of systemical medical therapy with increased side effects and costs. Strong prognostic factors are needed to divide patients into different risk group in order to help in making treatment decisions. The choice of postoperative treatment of primary breast cancer is based on clinical (age), histopathological (lymph node status, tumour size, histological grade), and cell biological oestrogen receptor (ER) and progesterone receptor (PgR) parameters [1]. These classical factors are generally considered not to be sufficient for this purpose. New and more powerful markers are therefore called for. In recent exploratory studies, markers from microarray gene expression analyses have shown some promise as breast cancer prognostic tools [2,3]. These gene expression results have been compared with data obtained from using conventional prognostic markers: i.e., age, tumour size, histological grade, angioinvasion, lymphocytic infiltration, ER and PgR status [2,3], both separately, and by using the National Institute of Health (NIH) and St. Gallen prognostic indices [1,4]. Although these comparisons favoured the gene expression markers, this was usually in comparison against single factors and the new technology has not been tested against a combined index, such as the Nottingham Prognostic Index (NPI).

Whether a novel factor gives information about clinical outcome and treatment sensitivity, in addition to that obtained by classical factors, is very important,

☆ Sources of support: Swedish Research Council, Swedish Foundation for Strategic Research, Lund University Medical Faculty and the Swedish Cancer Society.

[*] Corresponding author. Tel.: +46-46-17-75-65; fax: +46-46-14-73-27.

E-mail address: marten.ferno@onk.lu.se (M. Fernö).

0959-8049/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ejca.2004.02.025

since approximately half of all new breast cancers are diagnosed in the third world, where the analyses of prognostic factors need to be inexpensive, easy to perform, and have a good reproducibility. The conventional clinical and histopathological factors fulfill these criteria.

Here, we re-address the question of how the gene expression markers identified in the study by van 't Veer and colleagues [2] compare with conventional markers. This was done by using the high quality, publicly available data-set of 97 breast cancer tumours of node-negative patients, for which both gene expression data and conventional markers have been reported [2]. We specifically compared well established criteria based upon conventional markers, such as the NPI [5], to the gene expression markers suggested by van 't Veer and colleagues [2]. In addition, we investigated if artificial neural networks (ANNs) can be used to find new combinations of conventional markers, with improved prognostic power.

2. Patients and methods

The data-set in the study by van 't Veer and colleagues consists of 97 patients, younger than 55 years of age, with primary sporadic breast cancer, less than 5 cm in size and with no axillary metastasis. In this study, a poor prognosis group (46 patients developing distant metastasis within 5 years) and a good prognosis group (51 patients being distant metastasis-free for a follow-up period of at least 5 years) were selected [2]. Five patients received adjuvant systemic therapy, all in the poor prognosis group. The gene expression measurements were based upon 25,000 human genes. In the study of van 't Veer and colleagues [2], a set of approximately 200 genes correlating to prognotic status was selected for constructing a classifier based on gene expression. The predictor thus obtained was further used on 234 additional patients, with very encouraging results [3]. For the 234 additional patients, survival properties of the gene expression profiler were compared with those of the NIH [1] and St. Gallen [4] criteria in terms of Kaplan–Meier curves, which showed a very clear advantage for the gene expression profiler. Unfortunately, the data for the 234 patient cohort are not publicly available.

The publicly available gene expression data in the study of van 't Veer and colleagues [2] was pre-processed according to a private communication with the authors, which enabled us to reproduce their classifier. van 't Veer and colleagues also reported the more conventional markers age, tumour size, histological grade, ER, PgR, status of angioinvasion, and status of lymphocytic infiltration for the public data-set [2]. We compared the gene expression classifier with the NPI [5] (based on Coxian multivariate analysis), the NIH criterion [4], and the St. Gallen criterion [1], which are based upon different subsets of these conventional markers.

We also constructed a classifier based upon all seven reported conventional factors. To account for potential non-linear dependences in the data, we employed an ANN algorithm [6]. Given the relatively few data points, care must be exercised to avoid overfitting. This was done by sequentially leaving one patient out for a blind test. For each of these blind tests, the remaining patients were used for training using a cross-validation procedure for determining the learning parameters for a committee of ANNs as in [6]. Thus, based on all tumour samples, but one, a committee of 60 ANNs was constructed by repeating a 3-fold cross-validation 20 times. Several of such committees, with different ANN architectures and learning parameters, were constructed, and the committee with the best validation performance was applied to the omitted blind test sample. The whole procedure was repeated with a new test sample selected every time, so that every sample acted as blind test once. Hence, the results for each patient are not plagued with overfitting. This leave-one-out testing procedure is the same as that used for the microarray profiler, making performance comparisons valid.

The pool of ANN designs was constructed by varying the number of hidden nodes (0, 2 or 3), the weight decay parameter (0 or 0.01) and the number of training epochs (50 or 100). This gave a pool of 12 different designs. During training, a cross-entropy error was minimised using back propagation with a learning rate of 0.5, momentum coefficient of 0.1, and a decrease in the learning rate by a factor of 0.98 for each iteration.

3. Results

For classifiers, the balance between sensitivity and specificity depends upon a decision threshold. Following the study of van 't Veer and colleagues [2], we set thresholds corresponding to 10% misclassified in the metastasis group. For the NPI, this gave a threshold of 3.3, remarkably close to the standard threshold 3.4 between good prognosis and moderately good prognosis [7]. The St. Gallen criteria depends on several independent thresholds, and since the choice of those giving 10% false-positives would be rather arbitrary, the standard set of thresholds was used. This gave no false-positives for the St. Gallen criteria.

The odds ratios (OR) for the different approaches are presented in Table 1. As can be seen, the St. Gallen and NIH criteria did not perform very well. This is consistent with the data reported by van de Vijver and colleagues [3], using a larger cohort. It is interesting that among the commonly used indices, the NPI outperformed the other two. It is illustrative to present the comparisons in Kaplan–Meier curves even though

Table 1
Size of good prognosis group and odds ratios (maximum likelihood estimates) with 10% misclassified of the metastasis group

Method	Good prognosis M+/M−	OR	95% CI (Fisher's exact)
All 97			
Gene expression[a]	4/31	15.7	4.7–70
Conventional Markers[b]	4/25	9.8	2.9–43
NPI	4/21	7.2	2.1–32
NIH criteria	4/6	1.4	0.3–7.2
St. Gallen criteria	0/7	Infinity	1.4–infinity
ER+			
Gene expression[a]	3/27	14.7	3.6–90
Conventional Markers[b]	3/26	13.3	3.2–80
NPI	3/22	9.0	2.2–54
NIH criteria	3/4	0.9	0.13–6.4
St. Gallen criteria	0/7	Infinity	1.0–infinity

The good prognosis group is presented for M+ (false-positives) and M− (true-positives) separately. The St. Gallen criteria depends on several independent thresholds, and since the choice of those giving 10% false-positives would be rather arbitrary, the standard set of thresholds was used. This gave no false-positives for the St. Gallen criteria.

OR: odds ratio; CI: confidence interval; ER: oestrogen receptor; NPI: Nottingham prognostic index; ANN: artificial neural network; NIH: National institute of health.

[a] Leave-one-out cross-validation results, using the correlation classifier.

[b] Leave-one-out cross-testing results, using ANN committees constructed with cross-validation.

the sample selection hinders a too detailed analysis of the curves. In Fig. 1, Kaplan–Meier survival curves are shown for the gene expression profiler, ANN-based conventional markers and the NPI. The NIH and St. Gallen criteria are not shown here since so few patients remain in the good prognosis group. As can be seen, the three profilers performed in parity within the error bars.

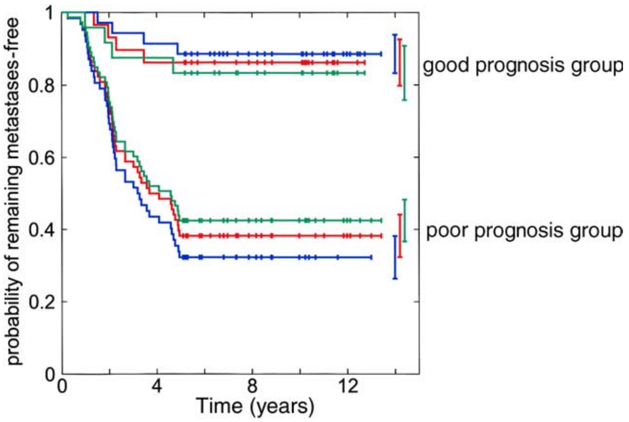

Fig. 1. Kaplan–Meier survival curves for the good and poor prognosis groups of the 97 tumour patients using gene expression measurements (blue curves), ANN-based conventional marker classifier (red curves) and the NPI (green curve), respectively. The classification output thresholds were set so that 10% of the patients with metastasis were misclassified. $P < 0.0001$ (log-rank test) for both blue and red curves and $P = 0.0016$ for the green curves.

In a study with a limited number of samples, performance measures such as OR can depend strongly on the choice of decision threshold. The results from varying such a threshold can be visualised in a receiver operating characteristic (ROC) curve, and the area under the ROC curve represents a performance average [8]. We therefore investigated the ROC areas for the different classifiers. In Table 2, it is clear that the ANN-based conventional marker classifier and the NPI performed as well as the gene expression method, whereas the NIH criteria was not so impressive. Results from the St. Gallen criteria are absent since a ROC curve cannot be uniquely defined for a classifier that depends on several independent thresholds.

Tables 1 and 2, and Fig. 1 show that the ANN-based conventional marker classifier performed well, slightly better than the NPI and in parity with the gene expression classifier. The dominating architecture of the ANNs had no hidden nodes and this corresponds to simple logistic regression. To illustrate how these networks could be transformed into something more useful in a clinic, we chose the most often selected ANN design and constructed a new network with no cross-validation. The resulting logistic expression had the form: $F = 1.0*(\text{histological grade}) + 0.51*(\text{size in cm})\text{-}0.14*(\text{age in years}) + 1.2*(\text{angioinvasion status}) - 0.011*(\text{ERp}) - 0.00026*(\text{PgRp}) - 0.34*(\text{lymphocytic infiltrate}) + 3.0$, where $F > 0$ implies a poor prognosis. Here, ERp and PgRp refer to percentage stained cell nuclei with immunohistochemical staining for ER and PgR, respectively. The final term 3.0 is set to give 10% false-positives with the threshold $F = 0$.

It should be emphasised that this formula is not presented as a final candidate for a better prognostic index. In particular, the function F is independent of axillary node status, which is a natural consequence of the fact that all samples in the study are node-negative. Rather, the presented formula exemplifies how "black box" machine-learning algorithms – such as ANNs – can often be disentangled to a more transparent form.

In the public data-set, 68 patients were ER positive (+) and 29 were ER negative (−). In contrast to the ER + cohort, there was a substantial overlap within the

Table 2
Metastasis prediction performances (ROC areas)

Method	All 97 samples	ER+ samples	ER− samples
Gene expression[a]	0.76	0.79	0.19
Conventional Markers[b]	0.78	0.85	0.42
NPI	0.77	0.81	0.57
NIH criteria	0.67	0.66	0.61

ROC: receiver operating characteristic curve.

[a] Leave-one-out cross-validation results, using the correlation classifier.

[b] Leave-one-out cross-testing results, using ANN committees constructed with cross-validation.

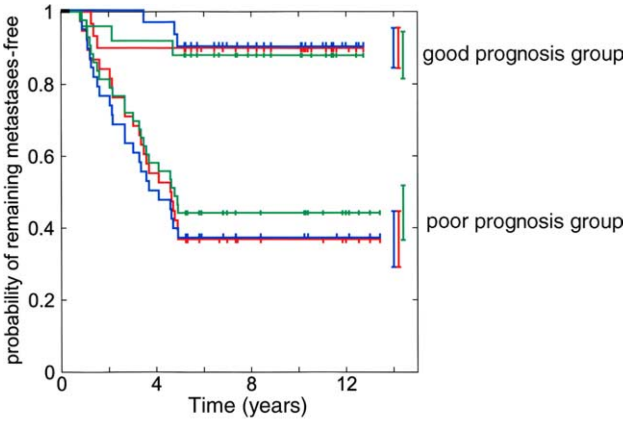

Fig. 2. Kaplan–Meier survival curves for the good and poor prognosis groups of ER+ tumour patients using gene expression measurements (blue curves), ANN-based conventional marker classifier (red curves) and the NPI (green curve), respectively. The classification output thresholds were set so that 10% of the patients with metastasis were misclassified. $P < 0.0001$ (log-rank test) for both blue and red curves, and $P = 0.0008$ for the green curves. To reduce the overlap of the curves and for clarity, the blue curves have been shifted up one line.

ER – cohort where the two methods failed. This observation, together with recent findings that ER+ and ER− tumours represent two distinct disease phenotypes [9], strongly suggests that prognostic predictors should be constructed separately for the two groups. In doing so for the 68 ER+ patients, using the same classifiers and validation procedures as above, the performance increased in terms of ROC areas for all methods (see Table 2). In Fig. 2, the Kaplan–Meier survival curves are shown for the three methods using data from patients with ER+ tumour samples. The exclusion of ER− patients resulted in minor threshold changes. The NPI threshold became 3.4, identical to the conventional one. As can be seen, the methods performed in parity.

Attempts were also made to build a classifier based upon the 29 ER− samples (19 with metastasis and 10 without) with no success for either of the approaches (see Table 2). To investigate if the poor result was an indication that the ER− category represents a more heterogeneous disease, or was merely a consequence of limited statistics, we processed 20 subgroups of 29 ER+ samples with the same proportion of cases with metastasis as in the ER− cohort, using the same procedure as for the other cohorts. In this case, the average ROC performances for the microarray- and the ANN-based conventional marker classifier were rather high, 0.69 and 0.81, indicating that metastasis is indeed more difficult to predict in ER− breast cancer than in a ER+ one.

4. Discussion

In general, proper evaluation of putative new prognostic factors such as gene expression profilers is of utmost importance from a clinical perspective. It may also have an economical impact on the health care system. Such an evaluation should fulfill certain criteria [10]: e.g., confirmative studies by other groups, well-designed, high-powered prospective clinical trial designed specifically to address the question, well-performed meta-analyses, and quality control studies of the assay. Furthermore, one of the most important issues concerns whether the new factor gives information, about clinical outcome and treatment sensitivity, in addition to that obtained by classical factors.

In our study of the public data-set presented by van ’t Veer and colleagues, the classifications based on gene expression and on more conventional markers performed in parity. This is in sharp contrast to the claims of van de Vijver and colleagues [3],where the NIH and St. Gallen indices are used for comparison, rather than the widely used NPI or an optimised use of the individual markers. To some extent the two methods fail for different samples, and they may therefore complement each other in a unified predictor. With the present statistics this cannot be demonstrated convincingly, but opens up the possibility of future studies with larger cohorts. The good results of the ANN classifier (Tables 1 and 2 and Fig. 1) suggest that a similar search for a new prognostic index, in a larger and more diverse cohort, could be worthwhile.

We conclude that despite the emergence of microarray data, the conventional markers are by no means outperformed as prognostic factors in breast cancer. Hence, care should be exercised before drawing too strong conclusions when it comes to the power of microarray gene expression data as prediction tools. However, it should be stressed that microarray tumour expression data brings additional knowledge to the table; i.e., insights into the important genes and pathways that underlie the disease outcome. In addition, one should keep in mind that microarray technology is still relatively new as compared with the use of conventional markers.

References

1. Goldhirsch A, Glick JH, Gelber RD, Senn HJ. Meeting highlights: international consensus panel on the treatment of primary breast cancer. *J Natl Cancer Inst* 1998, **90**, 1601–1608.
2. van ’t Veer LJ, Dai H, van de Vijver MJ, *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002, **415**, 530–536.
3. van de Vijver MJ, He YD, van ’t Veer LJ, *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002, **347**, 1999–2009.
4. Eifel P, Axelson JA, Costa J, *et al.* National institutes of health consensus development conference statement: adjuvant therapy for breast cancer, November 1–3, 2000. *J Natl Cancer Inst* 2001, **93**, 979–989.

5. Blamey RW, Davies CJ, Elston CW, Johnson J, Haybittle JL, Maynard PV. Prognostic factors in breast cancer – the formation of a prognostic index. *Clin Oncol* 1979, **5**, 227–236.
6. Ringnér M, Peterson C, Khan J. Analysing array data using supervised methods. *Pharmacogenomics* 2002, **3**, 403–415.
7. Todd JH, Dowle C, Williams MR, *et al*. Confirmation of a prognostic index in primary breast cancer. *Br J Cancer* 1987, **56**, 489–492.
8. Hanley JA, McNeil BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology* 1982, **143**, 29–36.
9. Gruvberger S, Ringnér M, Chen Y, *et al*. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res* 2001, **61**, 5979–5984.
10. Altman DG, Lyman GH. Methodological challenges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Res Treat* 1998, **52**, 289–303.