Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer

Yixin Wang, Jan G M Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer-van Gelder, Jack Yu, Tim Jatkoe, Els M J J Berns, David Atkins, John A Foekens

Lancet 2005; 365: 671–79

See Comment

Veridex LLC, a Johnson & Johnson Company, San Diego, CA, USA (Y Wang PhD, Y Zhang PhD, F Yang MSc, D Talantov MD, J Yu PhD, T Jatkoe BSc); **Veridex LLC, a Johnson & Johnson Company, Warren, NY, USA** (D Atkins PhD); **and Department of Medical Oncology, Erasmus MC–Daniel den Hoed, Rotterdam, Netherlands** (Prof J G M Klijn MD, A M Sieuwerts BSc, M P Look MSc, M Timmermans BSc, M E Meijer-van Gelder MD, E M J J Berns PhD, J A Foekens PhD)

Correspondence to: Dr John Foekens, Erasmus MC, Josephine Nefkens Institute, Rm BE-426, Dr Molewaterplein 50, 3015 GE Rotterdam, Netherlands **j.foekens@erasmusmc.nl**

Summary

Background **Genome-wide measures of gene expression can identify patterns of gene activity that subclassify tumours and might provide a better means than is currently available for individual risk assessment in patients with lymph-node-negative breast cancer.**

Methods **We analysed, with Affymetrix Human U133a GeneChips, the expression of 22 000 transcripts from total RNA of frozen tumour samples from 286 lymph-node-negative patients who had not received adjuvant systemic treatment.**

Findings **In a training set of 115 tumours, we identified a 76-gene signature consisting of 60 genes for patients positive for oestrogen receptors (ER) and 16 genes for ER-negative patients. This signature showed 93% sensitivity and 48% specificity in a subsequent independent testing set of 171 lymph-node-negative patients. The gene profile was highly informative in identifying patients who developed distant metastases within 5 years (hazard ratio 5·67 [95% CI 2·59–12·4]), even when corrected for traditional prognostic factors in multivariate analysis (5·55 [2·46–12·5]). The 76-gene profile also represented a strong prognostic factor for the development of metastasis in the subgroups of 84 premenopausal patients (9·60 [2·28–40·5]), 87 postmenopausal patients (4·04 [1·57–10·4]), and 79 patients with tumours of 10–20 mm (14·1 [3·34–59·2]), a group of patients for whom prediction of prognosis is especially difficult.**

Interpretation **The identified signature provides a powerful tool for identification of patients at high risk of distant recurrence. The ability to identify patients who have a favourable prognosis could, after independent confirmation, allow clinicians to avoid adjuvant systemic therapy or to choose less aggressive therapeutic options.**

Introduction

About 60–70% of patients with lymph-node-negative breast cancer are cured by local or regional treatment alone.[1,2] The most widely used treatment guidelines are the St Gallen[3] and the US National Institutes of Health[4] consensus criteria. These guidelines recommend adjuvant systemic therapy for 85–90% of lymph-node-negative patients. There is a need for specific definition of an individual patient's risk of disease recurrence to ensure that she receives appropriate therapy. Currently, few diagnostic tools are available to identify at-risk patients. To date, gene-expression patterns have been used to classify breast tumours into clinically relevant subtypes.[5–21] We report a comprehensive genome-wide assessment of gene expression to identify broadly applicable prognostic markers.[5,6] In this study, we aimed to develop a gene-expression-based algorithm and to use it to provide quantitative predictions on disease outcome for patients with lymph-node-negative breast cancer.

Methods

Patients' samples

We selected from our tumour bank at the Erasmus Medical Center (Rotterdam, Netherlands) frozen tumour samples from patients with lymph-node-negative breast cancer who were treated during 1980–95, but who did not receive systemic neoadjuvant or adjuvant therapy. Tumour samples were submitted to our reference laboratory from 25 regional hospitals for measurements of steroid-hormone receptors. Guidelines for primary treatment were similar for all hospitals. Selection of tumours aimed to avoid bias. On the assumption of a relapse rate of 25–30% in 5 years, and a substantial loss of tumours for quality-control reasons, 436 samples of invasive tumours were processed. Patients with poor, intermediate, and good clinical outcome were included. Samples were rejected on the basis of insufficient tumour content (53), poor RNA quality (77), or poor chip quality (20); thus, 286 samples were eligible for further analysis. The study was approved by institutional medical ethics committee (number 02·953). The median age of the patients at surgery was 52 years (range 26–83). 219 had undergone breast-conserving surgery and 67 modified radical mastectomy. Radiotherapy was given to 248 patients (87%) according to our institutional protocol. The proportions of patients who underwent breast-conserving therapy and radiotherapy are normal for lymph-node-negative disease. Patients were included irrespective of radiotherapy status because this study did not aim to investigate the effects of a specific type of surgery or adjuvant radiotherapy. Furthermore, other studies have shown that

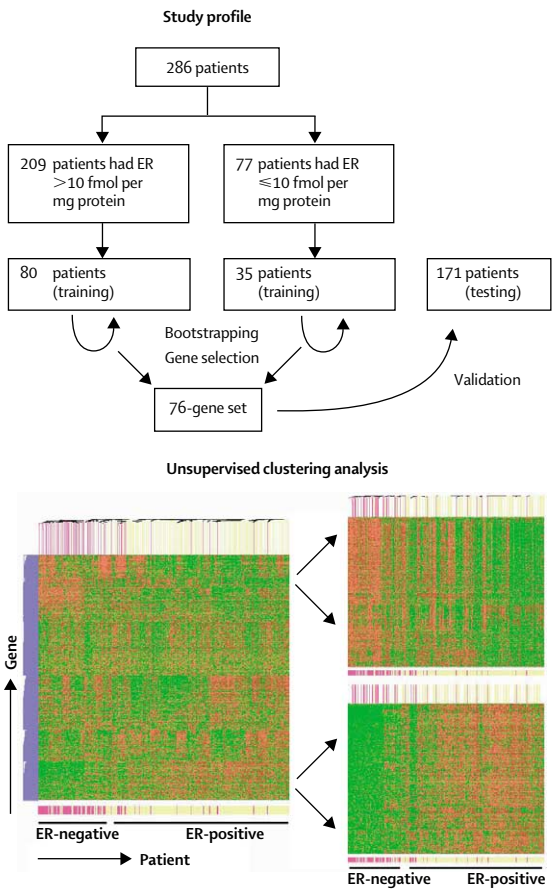

Figure 1: **Profile for selection of samples for analysis and unsupervised clustering analysis of gene-expression data for 286 patients with lymph-node-negative breast cancer**

ER status was used to identify subgroups. Each subgroup was then analysed separately for selection of markers. The patients in a subgroup were assigned to a training set or a testing set. The markers selected from each subgroup were combined to form a single signature to predict tumour recurrence for all patients in the testing set as a whole. The left panel of the clustering analysis is a view of the 17 819 informative genes. Red indicates high relative expression, green relative low expression. Each column is a sample and each row is a gene. The right panel shows enlarged views of two dominant gene clusters that had drastic differential expression between the two subgroups of patients. The upper gene cluster has a group of 282 downregulated genes in the ER-positive subgroup, and the lower gene cluster is represented by a group of 339 upregulated genes in the ER-positive subgroup. The label bar at the foot of each dendrogram indicates the patient's ER status measured by routine assays.

radiotherapy has no clear effect on distant disease recurrence.[22] Lymph-node negativity was based on pathological examination by regional pathologists.[23] All 286 tumour samples were confirmed to have sufficient (>70%) tumour and uniform involvement of tumour in 5 µm frozen sections stained with haematoxylin and eosin. Amounts of oestrogen receptors (ER) and progesterone receptors (PR) were measured by ligand-binding assay, EIA,[24] or immunohistochemistry (nine tumours). The cut-off value for classification of patients as positive or negative for ER and PR was 10 fmol per mg protein or 10% positive tumour cells. Postoperative follow-up involved examinations every 3 months for 2 years, every 6 months for years 3–5, and every 12 months from year 5. The date of diagnosis of metastasis was defined as that at confirmation of metastasis after symptoms reported by the patient, detection of clinical signs, or at regular follow-up.

Gene-expression analysis

Total RNA was isolated from 20–40 cryostat sections of 30 µm thickness (50–100 mg) with RNAzol B (Campro Scientific, Veenendaal, Netherlands). Biotinylated targets were prepared by published methods (Affymetrix, Santa Clara, CA, USA)[25] and hybridised to the Affymetrix oligonucleotide microarray U133a GeneChip. Arrays were scanned by standard Affymetrix protocols. Each probe set was treated as a separate gene. Expression values were calculated by use of Affymetrix GeneChip analysis software MAS 5.0. Chips with average intensity of less than 40 or background signal of more than 100 were rejected. For chip normalisation, probe sets were scaled to a target intensity of 600, and scale mask files were not selected.

Statistical methods

17 819 genes were "present" in two or more samples and were eligible for hierarchical clustering. Before clustering, the expression level of each gene was divided by its median expression level in the patients. This standardisation step limited the effect of the magnitude of expression of genes, and grouped together genes with similar patterns of expression in the clustering analysis. To identify subgroups of patients, we carried out average linkage hierarchical clustering on both the genes and the samples using GeneSpring 6.0. To identify genes that discriminated patients who developed distant metastases from those remaining metastasis-free for 5 years, we used two supervised class prediction approaches. In the first approach, 286 patients were randomly assigned to training and testing sets of 80 and 206 patients, respectively. Kaplan-Meier survival curves[26] for the two sets were examined to ensure that there was no significant difference and that no bias was introduced

Panel: Calculation of relapse scores

$$\text{Relapse score} = A \cdot I + \sum_{i=1}^{60} I \cdot w_i x_i + B \cdot (1-I) + \sum_{j=1}^{16} (1-I) \cdot w_j x_j$$

I=1 if ER is more than 10 fmol per mg protein; I=0 if ER is 10 fmol per mg protein or less; w_i is the standardised Cox's regression coefficient for an ER-positive marker; x_i is the expression value of the ER-positive marker on a $\log_2$ scale; w_j is the standardised Cox's regression coefficient for an ER-negative marker; x is the expression value of the ER-negative marker on a $\log_2$ scale; A and B are constants.

by the random selection of the training and testing sets. In the second approach, patients were allocated to one of two subgroups stratified by ER status (figure 1). Each subgroup was analysed separately for selection of markers. Patients in the ER-positive subgroup were randomly allocated into training and testing sets of 80 and 129 patients, respectively. The ER-negative subgroup was randomly divided into training and testing sets of 35 and 42 patients, respectively. Markers selected from each subgroup training set were combined to form a single signature to predict tumour metastasis for both ER-positive and ER-negative patients in a subsequent independent validation.

The sample size of the training set was determined by a resampling method to ensure its statistical confidence level. Briefly, the number of patients in the training set started at 15 patients and was increased in steps of five. For a given sample size, ten training sets with randomly selected patients were made. A gene signature was constructed from each of the training sets and tested in a designated testing set of patients by analysis of the receiver operating characteristic (ROC) curve with distant metastasis within 5 years as the defining point. The mean and the coefficient of variation of the area under the curve (AUC) for a given sample size were calculated. A minimum number of patients required for the training set was chosen at the point at which the average AUC reached a plateau and the coefficient of variation of the ten AUC was less than 5%.

Genes were selected as follows. First, univariate Cox's proportional-hazards regression was used to identify genes for which expression (on a $\log_2$ scale) was correlated with the length of distant-metastasis-free survival. To reduce the effect of multiple testing and to test the robustness of the selected genes, the Cox's model was constructed with bootstrapping of the patients in the training set.[27] Briefly, 400 bootstrap samples of the training set were constructed, each with 80 patients randomly chosen with replacement. A Cox's model was run on each of the bootstrap samples. A bootstrap score was created for each gene by removing the top and bottom 5% p values and averaging the inverses of the remaining bootstrap p values. This score was used to rank the genes. To construct a multiple gene signature, combinations of gene markers were tested by adding one gene at a time according to the rank order. ROC analysis with distant metastasis within 5 years as the defining point was done to calculate the AUC for each signature with increasing number of genes until a maximum AUC value was reached.

The relapse score was used to calculate each patient's risk of distant metastasis (panel). The score was defined as the linear combination of weighted expression signals with the standardised Cox's regression coefficient as the weight.

The threshold was determined from the ROC curve of the training set to ensure 100% sensitivity and the highest specificity. Values of constants A of 313·5 and B of 280 were chosen to centre the threshold of relapse score to zero for both ER-positive and ER-negative patients. Patients with positive or negative relapse scores were classified as those with poor or good prognosis, respectively. The gene signature and the cut-off were validated in the testing set. Kaplan-Meier survival plots and log-rank tests were used to assess the differences in time to distant metastasis of the predicted high-risk and low-risk groups. Odds ratios were calculated as the ratio of the odds of distant metastasis between the patients predicted to experience relapse and those predicted to remain relapse free.

Univariate and multivariate analyses with Cox's proportional-hazards regression were done on the individual clinical variables with and without the gene signature. The hazard ratio and its 95% CI were derived from these results. Statistical analyses used S-Plus software (version 6.1).

Pathway analysis

A functional class was assigned to each prognostic signature gene. Pathway analysis was done with Ingenuity software (version 1.0). Affymetrix probes were used as input to search for biological networks built by the software. Biological networks identified by

Characteristics	All patients (n=286)	ER-positive training set (n=80)	ER-negative training set (n=35)	Validation set (n=171)
Age, years				
Mean (SD)	54 (12)	54 (13)	54 (13)	54 (12)
≤40	36 (13%)	12 (15%)	3 (9%)	21 (12%)
41–55	129 (45%)	30 (38%)	17 (49%)	82 (48%)
56–70	89 (31%)	28 (35%)	11 (31%)	50 (29%)
>70	32 (11%)	10 (13%)	4 (11%)	18 (11%)
Menopausal status				
Premenopausal	139 (49%)	39 (49%)	16 (46%)	84 (49%)
Postmenopausal	147 (51%)	41 (51%)	19 (54%)	87 (51%)
T stage				
T1	146 (51%)	38 (48%)	14 (40%)	94 (55%)
T2	132 (46%)	41 (51%)	19 (54%)	72 (42%)
T3/4	8 (3%)	1 (1%)	2 (6%)	5 (3%)
Grade				
Poor	148 (52%)	37 (46%)	24 (69%)	87 (51%)
Moderate	42 (15%)	12 (15%)	3 (9%)	27 (16%)
Good	7 (2%)	2 (3%)	2 (6%)	3 (2%)
Unknown	89 (31%)	29 (36%)	6 (17%)	54 (32%)
ER status*				
Positive	209 (73%)	80 (100%)	0	129 (75%)
Negative	77 (27%)	0	35 (100%)	42 (25%)
PR status*				
Positive	165 (58%)	59 (74%)	5 (14%)	101 (59%)
Negative	111 (39%)	19 (24%)	29 (83%)	63 (37%)
Unknown	10 (3%)	2 (2%)	1 (3%)	7 (4%)
Metastases within 5 years				
Yes	93 (33%)	24 (30%)	13 (37%)	56 (33%)
No	183 (64%)	51 (64%)	17 (49%)	115 (67%)
Censored	10 (3%)	5 (6%)	5 (14%)	0

Data are number of patients unless otherwise stated. *Positive=>10 fmol per mg protein or >10% positive tumour cells.

Table 1: **Clinical and pathological characteristics of patients and their tumours**

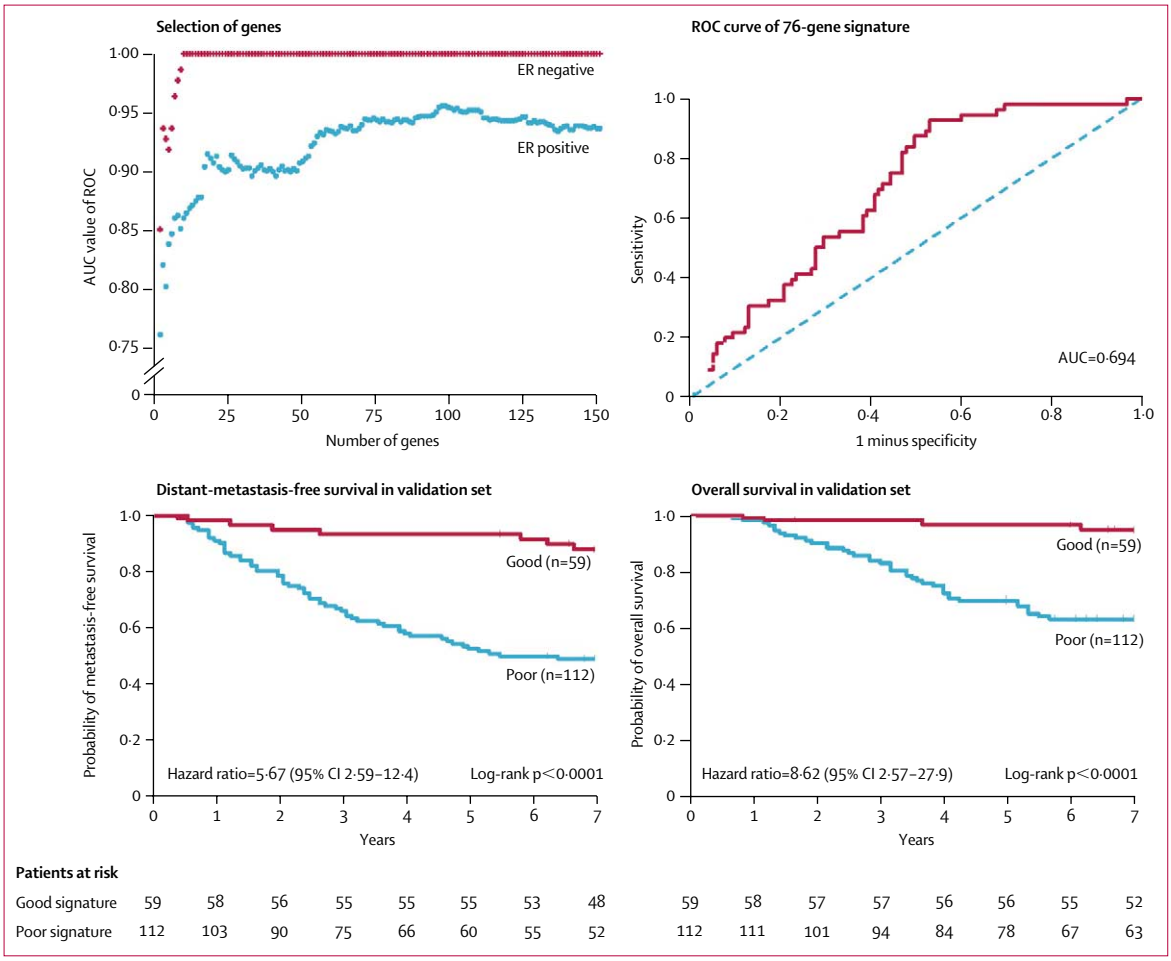

Figure 2: **Establishment of the 76-gene profile and Kaplan-Meier analysis for distant-metastasis-free and overall survival**

the program were assessed in the context of general functional classes by GO ontology classification. Pathways with two or more genes in the prognostic signature were selected and investigated.

Role of the funding sources

This study was supported partly by the Dutch Cancer Society and the Netherlands Genomics Initiative/Netherlands Organisation for Scientific Research. These organisations had no role in study design; the collection, analysis, or interpretation of data; writing of the paper; or in decisions relating to publication. The Erasmus Medical Centre was financially supported by Veridex LLC, a Johnson & Johnson Company, for tissue processing and isolating RNA for Affymetrix chip analysis. The corresponding author had full access to all the data in the study and took final responsibility for the decision to submit the paper for publication.

Results

The median follow-up for the 198 patients who survived was 101 months (range 20–171). Of the 286 patients included, 93 (33%) showed evidence of distant metastasis within 5 years and were counted as failures in analysis of distant-metastasis-free survival. Five (2%) patients died without evidence of disease and were censored at last follow-up. 83 (29%) died after previous relapse. Therefore, 88 patients (31%) were failures in the analysis of overall survival.

Clinical and pathological features of 286 patients are summarised in table 1. There were no differences among the groups in age or menopausal status. The ER-negative training group had a slightly higher proportion of larger tumours and, as expected, more poor-grade tumours than the ER-positive training group. The validation group of 171 patients (129 ER-positive, 42 ER-negative) did not differ from the total group of 286 patients in any of the characteristics of patients or tumours.

Two approaches were used to identify markers predictive of disease relapse. First, we randomly divided all the 286 patients (ER-positive and ER-negative combined) into a training set and a testing set. 35 genes were selected from 80 patients in the training

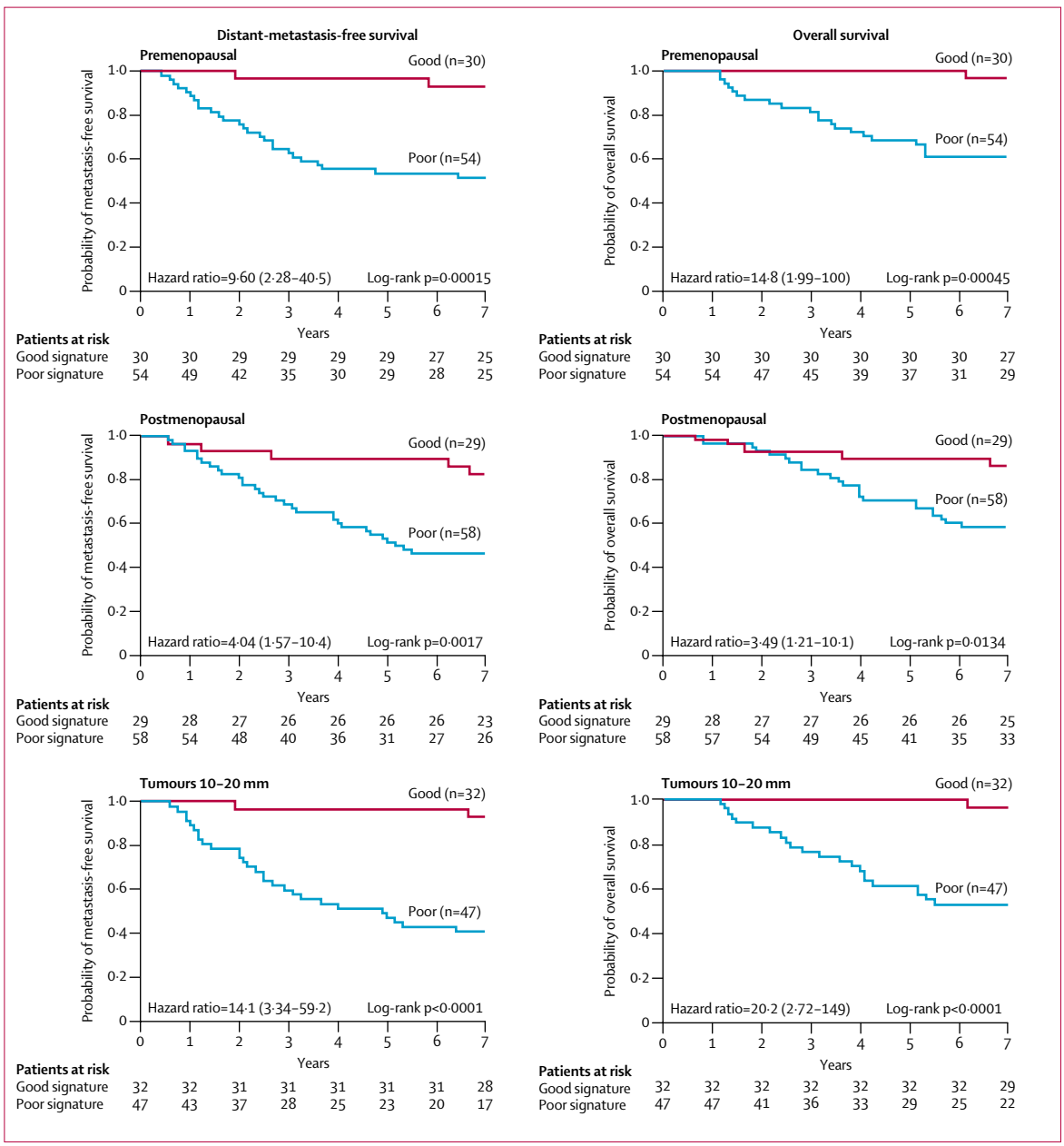

Figure 3: **Analysis of distant-metastasis-free and overall survival in subgroups of patients with lymph-node-negative breast cancer**

set and a Cox's model to predict the occurrence of distant metastasis was built. Moderate prognostic value was observed (data not shown). Unsupervised clustering analysis showed two distinct subgroups highly correlated with the tumour ER status (χ^2 test, $p<0·0001$; figure 1), which supported our second approach in which patients were first grouped on the basis of ER status. Each subgroup was analysed for selection of markers. 76 genes were selected from patients in the training sets (60 for the ER-positive group, 16 for the ER-negative group; figure 2). With the selected genes and ER status taken together, a Cox's model to predict recurrence of cancer was built for all lymph-node-negative patients. Validation of the 76-gene predictor in the testing set of 171 patients produced an ROC with an AUC of 0·694, sensitivity of 93% (52/56), and specificity of 48% (55/115; figure 2). Patients with a relapse score above the threshold of the prognostic signature have an odds ratio of 11·9 (95% CI 4·04–35·1; $p<0·0001$) to develop distant metastasis within 5 years. As the control, randomly selected 76-gene sets were generated. These produced ROC with an average AUC value of 0·515, sensitivity of 91%, and specificity of 12% in the testing group.

	Univariate analysis		Multivariate analysis*	
	Hazard ratio (95% CI)	p	Hazard ratio (95% CI)	p
Age 41–55 years vs ≤40 years	1·16 (0·51–2·65)	0·7180	1·14 (0·45–2·91)	0·7809
Age 56–70 years vs ≤40 years	1·32 (0·56–3·10)	0·5280	0·87 (0·26–2·93)	0·8232
Age >70 years vs ≤40 years	0·95 (0·32–2·82)	0·9225	0·61 (0·15–2·60)	0·5072
Postmenopausal vs premenopausal	1·24 (0·76–2·03)	0·3909	1·53 (0·68–3·44)	0·3056
Stages II and III vs stage I	1·08 (0·66–1·77)	0·7619	2·57 (0·23–29·4)	0·4468
Differentiation†	0·38 (0·16–0·90)	0·0281	0·60 (0·24–1·46)	0·2590
Tumour >20 vs ≤20 mm	1·06 (0·65–1·74)	0·8158	0·34 (0·03–3·90)	0·3849
ER positive vs negative	1·09 (0·61–1·98)	0·7649	1·05 (0·54–2·04)	0·8935
PR positive vs negative	0·83 (0·51–1·38)	0·4777	0·85 (0·47–1·53)	0·5882
76-gene signature	5·67 (2·59–12·4)	<0·0001	5·55 (2·46–12·5)	<0·0001

*The multivariate model included 162 patients, owing to missing values in nine. †Grade: moderate/good vs poor; unknown grade was included as a separate group.

Table 2: **Univariate and multivariate analyses for distant-metastasis-free survival in the testing set of 171 patients**

Patients stratified by such a gene set would have an odds ratio of 1·3 (0·50–3·90; p=0·8) for development of metastases, indicating a random classification. In addition, the Kaplan-Meier analyses for distant-metastasis-free and overall survival as a function of the 76-gene signature showed highly significant differences in time to metastasis between the groups predicted to have good and poor prognosis (figure 2). At 60 months and 80 months, the respective absolute differences in distant-metastasis-free survival between the groups with predicited good and poor prognosis were 40% (93% *vs* 53%) and 39% (88% *vs* 49%), and those in overall survival were 27% (97% *vs* 70%) and 32% (95% *vs* 63%) respectively.

The 76-gene profile also represented a strong prognostic factor for the development of distant metastasis in the subgroups of 84 premenopausal patients (hazard ratio 9·60), 87 postmenopausal patients (4·04), and 79 patients with tumour sizes of 10–20 mm (14·1; figure 3).

Univariate and multivariate Cox's regression analyses are summarised in table 2. Other than the 76-gene signature, only grade was significant in univariate analyses and moderate/good differentiation was associated with favourable distant-metastasis-free survival. Multivariate regression estimation of hazard ratio for the occurrence of tumour metastasis within 5 years was 5·55 (p<0·0001), indicating that the 76-gene set represents an independent prognostic signature strongly associated with a higher risk of tumour metastasis. Univariate and multivariate analyses were also done separately for ER-positive and ER-negative patients; the 76-gene signature was also an independent prognostic variable in the subgroups stratified by ER status (data not shown).

The function of the 76 genes (table 3) in the prognostic signature was analysed to relate the genes to biological pathways. Although 18 of the 76 genes have unknown function, several pathways or biochemical activities were identified that were well represented, such as cell death, cell cycle and proliferation, DNA replication and repair, and immune response (table 4). Genes implicated in disease progression were found, including calpain2, origin recognition protein, dual-specificity phosphatases, Rho-GDP dissociation inhibitor, tumour necrosis factor (TNF) superfamily protein, complement component 3, microtubule-associated protein, protein phosphatase 1, and apoptosis regulator BCL-G. Furthermore, previously characterised prognostic genes such as cyclin E2[28] and CD44[29] were in the gene signature.

The dataset has been submitted to the NCBI/Genbank GEO database (series entry GSE2034).

Discussion

We provide results of an analysis of primary tumours from 286 patients with lymph-node-negative breast cancer of all age-groups and tumour sizes. The patients had not received adjuvant systemic therapy, so the multigene assessment of prognosis was not subject to potentially confounding contributions by predictive factors related to systemic treatment.

The study revealed a 76-gene signature that accurately predicts distant tumour recurrence. This signature could be applied to all lymph-node-negative patients independently of age, tumour size and grade, and ER status. In Cox's multivariate analysis for distant-metastasis-free survival, the 76-gene signature was the only significant variable, superseding clinical variables, including grade. After 5 years, absolute differences in distant-metastasis-free and overall survival between the patients with the good and poor 76-gene signatures were 40% and 27%, respectively. Of the patients with good-prognosis signatures, 7% developed distant metastases and 3% died within 5 years. If further validated, this signature will yield a positive predictive value of 37% and a negative predictive value of 95%, on the assumption of a 25% rate of disease recurrence in lymph-node-negative patients. In particular, this signature could be valuable for defining the risk of recurrence for the increasing proportion of T1 tumours (<2 cm). Comparison with the St Gallen and National Institutes of Health guidelines was instructive. Although ensuring that the same number of high-risk patients would receive the necessary treatment, our 76-gene signature would recommend systemic adjuvant chemotherapy to only 52% of low-risk patients, compared with 90% and 89% by the St Gallen and National Institutes of Health guidelines (table 5). Our gene signature, if further confirmed, could result in a reduction of the number of low-risk lymph-node-negative patients who would be recommended to have unnecessary adjuvant systemic therapy (table 5).

The 76 genes in our prognostic signature belong to many functional classes, which suggests that different paths could lead to disease progression. The signature

Gene	Standard Cox coefficient	Cox p value	Gene description
For ER-positive group			
219340_s_at	−3·83	0·00005	gb:AF123759.1 /DEF=*Homo sapiens* putative transmembrane protein (CLN8) mRNA, complete cds
217771_at	−3·865	0·00001	gb:NM_016548.1 /DEF=*Homo sapiens* golgi membrane protein GP73 (LOC51280)
202418_at	3·63	0·00002	gb:NM_020470.1 /DEF=*Homo sapiens* putative transmembrane protein; homologue of yeast Golgi membrane protein Yif1p
206295_at	−3·471	0·00016	gb:NM_001562.1 /DEF=*Homo sapiens* interleukin 18 (interferon-γ-inducing factor) (IL18)
201091_s_at	3·506	0·00008	Consensus includes gb:BE748755 /heterochromatin-like protein 1
204015_s_at	−3·476	0·00001	gb:BC002671.1 /DEF=*Homo sapiens*, dual specificity phosphatase 4
200726_at	3·392	0·00006	gb:NM_002710.1 /DEF=*Homo sapiens* protein phosphatase 1, catalytic subunit, γ isoform (PPP1CC)
200965_s_at	−3·353	0·0008	gb:NM_006720.1 /DEF=*Homo sapiens* actin binding LIM protein 1 (ABLIM), transcript variant ABLIM-s
210314_x_at	−3·301	0·00038	gb:AF114013.1 /DEF=*Homo sapiens* TNF-related death ligand-1γ
221882_s_at	3·101	0·00033	Consensus includes gb:AI636233 five-span transmembrane protein M83
217767_at	−3·174	0·00128	gb:NM_000064.1 /DEF=*Homo sapiens* complement component 3 (C3)
219588_s_at	3·083	0·0002	gb:NM_017760.1 /DEF=*Homo sapiens* hypothetical protein FLJ20311
204073_s_at	3·336	0·00005	gb:NM_013279.1 /DEF=*Homo sapiens* chromosome 11open reading frame 9 (C11ORF9)
212567_s_at	−3·054	0·00063	Consensus includes gb:AL523310 putative translation initiation factor
211382_s_at	−3·025	0·00332	gb:AF220152.2 /DEF=*Homo sapiens* TACC2 mRNA
201663_s_at	3·095	0·00044	gb:NM_005496.1 /DEF=*Homo sapiens* chromosome-associated polypeptide C (CAP-C)
221344_at	−3·175	0·00031	gb:NM_013936.1 /DEF=*Homo sapiens* olfactory receptor, family 12, subfamily D, member 2 (OR12D2)
210028_s_at	−3·082	0·00086	gb:AF125507.1 /DEF=*Homo sapiens* origin recognition complex subunit 3 (ORC3)
218782_s_at	3·058	0·00016	gb:NM_014109.1 /DEF=*Homo sapiens* PRO2000 protein (PRO2000)
201664_at	3·085	0·00009	gb:AL136877.1 /SMC4 (structural maintenance of chromosomes 4, yeast)-like 1 /FL=gb:AB019987.1 gb:NM_005496.1 gb:AL136877.1
219724_s_at	−2·992	0·0004	gb:NM_014796.1 /DEF=*Homo sapiens* KIAA0748 gene product (KIAA0748)
204014_at	−2·791	0·0002	gb:NM_001394.2 /DEF=*Homo sapiens* dual specificity phosphatase 4 (DUSP4)
212014_x_at	−2·948	0·00039	Consensus includes gb:AI493245 /CD44 antigen (homing function and Indian blood group system)
202240_at	2·931	0·0002	gb:NM_005030.1 /DEF=*Homo sapiens* polo (Drosophila)-like kinase (PLK)
204740_at	−2·896	0·00052	gb:NM_006314.1 /DEF=*Homo sapiens* connector enhancer of KSR-like (Drosophila kinase suppressor of ras) (CNK1)
208180_s_at	2·924	0·0005	gb:NM_003543.2 /DEF=*Homo sapiens* H4 histone family, member H (H4FH)
204768_s_at	2·915	0·00055	gb:NM_004111.3 /DEF=*Homo sapiens* flap structure-specific endonuclease 1 (FEN1)
203391_at	−2·968	0·00099	gb:NM_004470.1 /DEF=*Homo sapiens* FK506-binding protein 2 (13kD) (FKBP2)
211762_s_at	2·824	0·00086	gb:BC005978.1 /DEF=*Homo sapiens*, karyopherin α 2 (RAG cohort 1, importin α 1)
218914_at	−2·777	0·00398	gb:NM_015997.1 /DEF=*Homo sapiens* CGI-41 protein (LOC51093)
221028_s_at	−2·635	0·0016	gb:NM_030819.1 /DEF=*Homo sapiens* hypothetical protein MGC11335 (MGC11335)
211779_x_at	−2·854	0·00053	gb:BC006155.1 /DEF=*Homo sapiens*, clone MGC:13188
218883_s_at	2·842	0·00051	gb:NM_024629.1 /DEF=*Homo sapiens* hypothetical protein FLJ23468 (FLJ23468)
204888_s_at	−2·835	0·00033	Consensus includes gb:AA772093 /neuralised (Drosophila)-like /FL=gb:U87864.1 gb:AF029729.1 gb:NM_004210.1
217815_at	2·777	0·00164	gb:NM_007192.1 /DEF=*Homo sapiens* chromatin-specific transcription elongation factor, 140 kDa subunit (FACTP140)
201368_at	−2·759	0·00222	Consensus includes gb:U07802 /DEF=Human Tis11d gene
201288_at	−2·745	0·00086	gb:NM_001175.1 /DEF=*Homo sapiens* Rho GDP dissociation inhibitor (GDI) β (ARHGDIB)
201068_s_at	2·79	0·00049	gb:NM_002803.1 /DEF=*Homo sapiens* proteasome (prosome, macropain) 26S subunit, ATPase, 2 (PSMC2)
218478_s_at	2·883	0·00031	gb:NM_017612.1 /DEF=*Homo sapiens* hypothetical protein DKFZp434E2220 (DKFZp434E2220)
214919_s_at	−2·794	0·00139	Consensus includes gb:R39094 /KIAA1085 protein
209835_x_at	−2·743	0·00088	gb:BC004372.1 /DEF=*Homo sapiens*, Similar to CD44 antigen (homing function and Indian blood group system)
217471_at	−2·761	0·00164	Consensus includes gb:AL117652.1 /DEF=*Homo sapiens* mRNA
203306_s_at	−2·831	0·00535	gb:NM_006416.1 /DEF=*Homo sapiens* solute carrier family 35 (CMP-sialic acid transporter), member 1 (SLC35A1)
205034_at	2·659	0·00073	gb:NM_004702.1 /DEF=*Homo sapiens* cyclin E2 (CCNE2)
221816_s_at	−2·715	0·00376	Consensus includes gb:BF055474 / putative zinc finger protein NY-REN-34 antigen
219510_at	2·836	0·00029	gb:NM_006596.1 /DEF=*Homo sapiens* polymerase (DNA directed), θ (POLQ)
217102_at	−2·687	0·00438	Consensus includes gb:AF041410.1 /DEF=*Homo sapiens* malignancy-associated protein
208683_at	−2·631	0·00226	gb:M23254.1 /DEF=Human Ca2-activated neutral protease large subunit (CANP)
215510_at	−2·716	0·00089	Consensus includes gb:AV693985 /ets variant gene 2
218533_s_at	2·703	0·00232	gb:NM_017859.1 /DEF=*Homo sapiens* hypothetical protein FLJ20517 (FLJ20517)
215633_x_at	−2·641	0·00537	Consensus includes gb:AV713720 /*Homo sapiens* mRNA for LST-1N protein
221928_at	−2·686	0·00479	Consensus includes gb:AI057637 /Hs.234898 ESTs, Weakly similar to 2109260A B cell growth factor *Homo sapiens*
214806_at	−2·654	0·00363	Consensus includes gb:U90030.1 /DEF=*Homo sapiens* bicaudal-D (BICD) mRNA, alternatively spliced, partial cds
204540_at	2·695	0·00095	gb:NM_001958.1 /DEF=*Homo sapiens* eukaryotic translation elongation factor 1 α 2 (EEF1A2)
221916_at	−2·758	0·00222	Consensus includes gb:BF055311 / hypothetical protein
216693_x_at	2·702	0·00084	Consensus includes gb:AL133102.1 /DEF=*Homo sapiens* mRNA; cDNA DKFZp434C1722
209500_x_at	−2·694	0·00518	gb:AF114012.1 /DEF=*Homo sapiens* TNF-related death ligand-1β mRNA
209524_at	2·711	0·00049	*Homo sapiens* cDNA FLJ10418 fis, clone NT2RP1000130, moderately similar to hepatoma-derived growth factor
207118_s_at	−2·771	0·00156	gb:NM_004659.1 /DEF=*Homo sapiens* matrix metalloproteinase 23A (MMP23A)
211040_x_at	2·604	0·00285	gb:BC006325.1 /DEF=*Homo sapiens*, G-2 and S-phase expressed 1
For ER-negative group			
218430_s_at	−3·495	0·00011	gb:NM_022841.1 /DEF=*Homo sapiens* hypothetical protein FLJ12994 (FLJ12994)
217404_s_at	3·224	0·00036	Consensus includes gb:X16468.1 /DEF=Human mRNA for α-1 type II collagen.
205848_at	−3·225	0·00041	gb:NM_005256.1 /DEF=*Homo sapiens* growth arrest-specific 2 (GAS2)
214915_at	−3·145	0·00057	*Homo sapiens* cDNA FLJ11780 fis, clone HEMBA1005931, weakly similar to zinc finger protein 83
216010_x_at	−3·055	0·00075	Consensus includes gb:D89324 /DEF=*Homo sapiens* DNA for alpha (1,31,4) fucosyltransferase
204631_at	−3·037	0·00091	gb:NM_017534.1 /DEF=*Homo sapiens* myosin, heavy polypeptide 2, skeletal muscle, adult (MYH2)
202687_s_at	−3·066	0·00072	gb:U57059.1 /DEF=*Homo sapiens* Apo-2 ligand mRNA

Continued

Continued Gene	Standard Cox coefficient	Cox p value	Gene description
221634_at	3·06	0·00077	gb:BC000596.1 /DEF=*Homo sapiens*, Similar to ribosomal protein L23a, clone MGC:2597
220886_at	−2·985	0·00081	gb:NM_018558.1 /DEF=*Homo sapiens* GABA receptor, θ (GABRQ)
202239_at	−2·983	0·00104	gb:NM_006437.2 /DEF=*Homo sapiens* ADP-ribosyltransferase (NAD+; poly (ADP-ribose) polymerase)-like 1 (ADPRTL1)
204218_at	−3·022	0·00095	gb:NM_014042.1 /DEF=*Homo sapiens* DKFZP564M082 protein (DKFZP564M082)
221241_s_at	−3·054	0·00082	gb:NM_030766.1 /DEF=*Homo sapiens* apoptosis regulator BCL-G (BCLG)
209862_s_at	−3·006	0·00098	gb:BC001233.1 /DEF=*Homo sapiens*, Similar to KIAA0092 gene product, clone MGC:4896
217019_at	−2·917	0·00134	Contains a novel gene and the 5 part of a gene for a novel protein similar to X-linked ribosomal protein 4 (RPS4X)
210593_at	−2·924	0·00149	gb:M55580.1 /DEF=Human spermidinespermine N1-acetyltransferase
216103_at	−2·882	0·0017	Consensus includes gb:AB014607.1 /DEF=*Homo sapiens* mRNA for KIAA0707 protein

Table 3: **76 genes from the prognostic signature**

included well-characterised genes and 18 unknown genes. This finding could explain the superior performance of this signature compared with other prognostic factors. Although genes involved in cell death, cell proliferation, and transcriptional regulation were found in both groups of patients stratified by ER status, the 60 genes selected for the ER-positive group and the 16 selected for the ER-negative group had no overlap. This result supports the idea that the extent of heterogeneity and the underlying mechanisms for disease progression could differ for the two ER-based subgroups of breast-cancer patients.

Comparison of our results with those of Van de Vijver and colleagues[12] is difficult because of differences in patients, techniques, and materials used. Their study included node-negative and node-positive patients, who had or had not received adjuvant systemic therapy, and only women younger than 53 years. Furthermore, the microarray platforms used in the studies differ—Affymetrix and Agilent. Of the 70 genes in the study by van't Veer and co-workers,[11] 48 are present on the Affymetrix U133a array, whereas only 38 of our 76 genes are present on the Agilent array. There is a three-gene overlap between the two signatures (cyclin E2, origin recognition complex, and TNF superfamily protein). Despite the apparent difference, both signatures included genes that identified several common pathways that might be involved in tumour recurrence. This finding supports the idea that although there might be redundancy in gene members, effective signatures could be required to include representation of specific pathways.

Functional class	76-gene signature
Cell death	TNFSF10, TNFSF13, MAP4, CD44, IL18, GAS2, NEFL, EEF1A2, BCLG, C3
Cell cycle	CCNE2, CD44, MAP4, SMC4L1, TNFSF10, AP2A2, FEN1, KPNA2, ORC3L, PLK1
Proliferation	CD44, IL18, TNFSF10, TNFSF13, PPP1CC, CAPN2, PLK1, SAT
DNA replication, recombination, and repair	TNFSF10, SMC4L1, FEN1, ORC3L, KPNA2, SUPT16H, POLQ, ADPRTL1
Immune response	TNFSF10, CD44, IL18, TNFSF13, ARHGDIB, C3
Growth	PPP1CC, CD44, IL18, TNFSF10, SAT, HDGFRP3
Cellular assembly and organisation	MAP4, NEFL, TNFSF10, PLK1, AP2A2, SMC4L1
Transcription	KPNA2, DUSP4, SUPT16H, DKFZP434E2220, PHF11, ETV2
Cell-to-cell signalling and interaction	CD44, IL18, TNFSF10, TNFSF13, C3
Survival	TNFSF10, TNFSF13, CD44, NEFL
Development	IL18, TNFSF10, COL2A1
Cell morphology	CAPN2, CD44, TACC2
Protein synthesis	IL18, TNFSF10, EEF1A2
ATP binding	PRO2000, URKL1, ACACB
DNA binding	HIST1H4H, DKFZP434E2220, PHF11
Colony formation	CD44, TNFSF10
Adhesion	CD44, TMEM8
Neurogenesis	CLN8, NEURL
Golgi apparatus	GOLPH2, BICD1
Kinase activity	CNK1, URKL1
Transferase activity	FUT3, ADPRTL1

Table 4: **Pathway analysis of the 76 genes from the prognostic signature**

Method	Patients guided to receive adjuvant chemotherapy in the testing set	
	Metastatic disease at 5 years	Free of metastatic disease at 5 years
St Gallen	52/55 (95%)	104/115 (90%)
National Institutes of Health	52/55 (95%)	101/114 (89%)
76-gene signature	52/56 (93%)	60/115 (52%)

St Gallen consensus criteria: tumour ⩾2 cm, ER negative, grade 2–3, patient <35 years (any one of these criteria). National Institutes of Health: tumour >1 cm.

Table 5: **Comparison of the 76-gene signature and the current conventional consensus on treatment of breast cancer**

The strengths of our study compared with the study of Van de Vijver and colleagues[12] are the larger number of untreated lymph-node-negative patients (286 *vs* 141), and the independence of our 76-gene signature with respect to age, menopausal status, and tumour size. The validation set of patients is completely without overlap with the training set, in contrast to 90% of other reports.[30] In conclusion, since only 30–40% of untreated lymph-node-negative patients develop tumour recurrence, our prognostic signature could provide a powerful tool to identify those patients at low risk preventing overtreatment in substantial numbers of patients. If confirmed in subsequent studies, the recommendation of adjuvant systemic therapy in patients with lymph-node-negative primary breast cancer could be guided by this prognostic signature. The predictive value of our gene signature with respect to the efficacy of different modes of systemic therapy could be tested in the adjuvant setting or in patients with metastatic disease.

Contributors

Y Wang, J G M Klijn, E M J J Berns, D Atkins, and J A Foekens designed the study, interpreted the data, and wrote the report. Y Zhang, J Yu, and T Jatkoe analysed the data and developed the prognostic signature. M Timmermans and D Talantov were

responsible for laboratory experiments and pathological assessment of the tissue samples. F Yang and A M Sieuwerts did laboratory experiments on the isolation of RNA and quality assessment. M P Look and M E Meijer-van Gelder collected and handled the patients' data and contributed to the survival analyses.

Conflict of interest statement

YW, YZ, FY, DT, JY, TJ, and DA are employed by Veridex LLC, a Johnson & Johnson Company, which is in the business of commercialising diagnostic products. The other authors declare no conflicts of interest.

Acknowledgments

We thank Anneke Goedheer, Anita Trapman-Jansen, Miranda Arnold, and Roberto Rodriguez-Garcia for technical assistance, and the surgeons, pathologists, and internists of the St Clara Hospital, Ikazia Hospital, St Franciscus Gasthuis at Rotterdam, and Ruwaard van Putten Hospital at Spijkenisse for the supply of tumour tissues, for their assistance in the collection of the clinical follow-up data, or both.

References

1 Early Breast Cancer Trialists' Collaborative Group. Polychemotherapy for early breast cancer: an overview of the randomised trials. *Lancet* 1998; **352:** 930–42.

2 Early Breast Cancer Trialists' Collaborative Group. Tamoxifen for early breast cancer: an overview of randomised trials. *Lancet* 1998; **351:** 1451–67.

3 Goldhirsch A, Wood C, Gelber RD, Coates AS, Thürlimann B, Senn HJ. Meeting highlights: updated international expert consensus on the primary therapy of early breast cancer. *J Clin Oncol* 2003; **21:** 3357–65.

4 Eifel P, Axelson JA, Costa J, et al. National Institutes of Health Consensus Development Conference Statement: adjuvant therapy for breast cancer, Nov 1–3, 2000. *J Natl Cancer Inst* 2001; **93:** 979–89.

5 Ntzani E, Ionnidis JPA. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet* 2003; **362:** 1439–44.

6 Wang Y, Jatkoe T, Zhang Y, et al. Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer. *J Clin Oncol* 2004; **22:** 1564–71.

7 Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature* 2000; **406:** 747–52.

8 Sørlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* 2001; **98:** 10869–74.

9 Sørlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci USA* 2003; **100:** 8418–23.

10 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:** 5979–84.

11 Van 't Veer L, Dai H, Van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; **415:** 530–36.

12 Van de Vijver MJ, Yudong HE, Van't Veer L, et al. A gene expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002; **347:** 1999–2009.

13 Ahr A, Kam T, Solbach C, et al. Identification of high risk breast-cancer patients by gene-expression profiling. *Lancet* 2002; **359:** 131–32.

14 Huang E, Cheng SH, Dressman H, et al. Gene expression predictors of breast cancer outcomes. *Lancet* 2003; **361:** 1590–96.

15 Sotiriou C, Neo S-Y, McShane LM, et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci USA* 2003; **100:** 10393–98.

16 Woelfle U, Cloos J, Sauter G, et al. Molecular signature associated with bone marrow micrometastasis in human breast cancer. *Cancer Res* 2003; **63:** 5679–84.

17 Ma X-J, Salunga R, Tuggle JT, et al. Gene expression profiles of human breast cancer progression. *Proc Natl Acad Sci USA* 2003; **100:** 5974–79.

18 Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. *Nat Genet* 2003; **33:** 1–6.

19 Chang JC, Wooten EC, Tsimelzon A, et al. Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet* 2003; **362:** 362–69.

20 Sotiriou C, Powles TJ, Dowsett M, et al. Gene expression profiles derived from fine needle aspiration correlate with response to systemic chemotherapy in breast cancer. *Breast Cancer Res* 2003; **4:** R3.

21 Hedenfalk I, Duggan D, Chen Y, et al. Gene-expression profiles in hereditary breast cancer. *N Engl J Med* 2001; **344:** 539–48.

22 Early Breast Cancer Trialists' Collaborative Group. Effects of radiotherapy and surgery in early breast cancer: an overview of the randomized trials. *N Engl J Med* 1995; **333:** 1444–55.

23 Foekens JA, Portengen H, van Putten WLJ, et al. Prognostic value of receptors for insulin-like growth factor 1, somatostatin, and epidermal growth factor in human breast cancer. *Cancer Res* 1989; **49:** 7002–09.

24 Foekens JA, Portengen H, van Putten WLJ, et al. Prognostic value of estrogen and progesterone receptors measured by enzyme immunoassays in human breast tumor cytosols. *Cancer Res* 1989; **49:** 5823–28.

25 Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. *Nat Genet* 1999; **21:** 20–24.

26 Kaplan EL, Maier P. Non-parametric estimation of incomplete observations. *J Am Stat Assoc* 1958; **53:** 457–81.

27 Efron B. Censored data and the bootstrap. *J Am Stat Assoc* 1981; **76:** 312–19.

28 Keyomarsi K, Tucker SL, Buchholz TA, et al. Cyclin E and survival in patients with breast cancer. *N Engl J Med* 2002; **347**: 1566–75.

29 Herrera-Gayol A, Jothy S. Adhesion proteins in the biology of breast cancer: contribution of CD44. *Exp Mol Pathol* 1999; **66**: 149–56.

30 Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. *Nat Rev Cancer* 2004; **4**: 309–14.