Markers of Progression in Early Stage Invasive Breast Cancer: a Predictive Immunohistochemical Panel Algorithm for Distant Recurrence Risk Stratification

Aleskandarany MA[1,2], Soria D[3,4], Green AR[1], Nolan C[1], Maria Diez-Rodriguez, Ellis IO[1], Rakha EA[1,2]

[1]Division of Cancer and Stem Cells, School of Medicine, University of Nottingham, Nottingham, UK, [2]Pathology Department, Menofia Faculty of Medicine, Egypt, , [3]School of Computer Science, University of Nottingham, Nottingham, UK, [4]Advanced Data Analysis Centre, University of Nottingham, Nottingham, UK,

Corresponding Author:

Mohammed Aleskandarany, *MD, PhD*,

Molecular Pathology Research Unit,

Division of Cancer and Stem Cells,

Nottingham City Hospitals,

The University of Nottingham,

Nottingham NG5 1PB

T: +44 (0) 115 8231859

F: +44 (0) 115 9627768

Email: mohammed.aleskandarany@nottingham.ac.uk

Key words: Breast cancer, immunohistochemistry, metastasis, risk stratification algorithm

Abstract

Accurate distant metastasis (DM) prediction is critical for risk stratification and effective treatment decisions in breast cancer (BC). Many prognostic markers/models based on tissue marker studies are continually emerging using conventional statistical approaches analysing complex/dimensional data association with DM/poor prognosis. However, few of them have fulfilled satisfactory evidences for clinical application. This study aimed at building DM risk assessment algorithm for BC patients.

A well-characterised series of early invasive primary operable BC (n=1902), with immunohistochemical (IHC) expression of a panel of biomarkers (n=31) formed the material of this study. Decision tree algorithm was computed using WEKA software, utilising quantitative biomarkers' expression and the absence/presence of distant metastases.

Fifteen biomarkers were significantly associated with DM, with six temporal subgroups characterised based on time-to-development of DM ranging from < 1 year to > 15 years of follow-up. Of these 15 biomarkers, 10 had a significant expression pattern where Ki67LI, HER2, p53, N-cadherin, P-cadherin, PIK3CA and TOMM34 showed significantly higher expressions with earlier development of DM. In contrast, higher expressions of ER, PR, and BCL2, were associated with delayed occurrence of DM. DM prediction algorithm was built utilising cases informative for the 15 significant markers. Four risk groups of patients were characterised. Three markers; p53, HER2 and BCL2 predicted the probability of DM, based on software-generated cut-offs, with a precision rate of 81.1% for positive predictive value and 77.3%, for the negative predictive value.

This algorithm reiterates the reported prognostic values of these three markers and underscores their central biologic role in BC progression. Further independent validation of this pruned panel of biomarkers is therefore warranted.

INTRODUCTION

Distant recurrence is the major cause of cancer related deaths in breast cancer (BC) patients [1]. For cancer cells to successfully colonise a secondary site, they have to fulfil specific prerequisites to overcome the vast stresses throughout the metastatic cascade [2]. Collectively, the success of the metastatic process results from integration and contribution of complex molecular pathways controlling cellular proliferation, survival, metabolism, invasion and migration [3].

Accuracy in BC prognosis/prediction, particularly distant recurrence risk assessment, is critical for accurate patients' stratification and effective treatment decision making. Many prognostic markers and models based on tissue marker studies are continually emerging; however, very few of them have fulfilled satisfactory evidence for clinical application. Poor study design and misleading statistical analyses have been proposed as to explain discrepancies in research studies generating relevant clinically useful prognostic markers [4].

Mining large datasets regarding the expression patterns of a large number of biomarkers and clinical variables requires stringent statistical approaches to derive robust conclusions. Decision tree is an approach followed to automatically learn, through machine learning, to recognise complex patterns and illustrate relations between observed variables to make intelligent decisions. Therefore, decision tree algorithms could help improving our basic understanding of cancer development and progression [5] which can be used to assist the classification of breast cancer cases by sorting them based on feature or attribute values (e.g. tissue marker expression). Each node in a decision tree represents a feature in a case to be classified, and each branch represents a value that the node can assume. Cases are classified starting at the root node and sorted based on their feature values [6].

Perhaps the most common algorithm in literature for building decision trees is the C4.5 developed by Ross Quinlan. C4.5 is a statistical classifier system which uses the concept of information gain to make a tree of classificatory decisions with respect to a previously chosen target classification (e.g. distant recurrence). Each attribute of the data can be used to make a decision that splits the data into smaller subsets [7]. These outputs are then expressed as models, in the form of decision trees or sets of if-then rules, which can be used to classify new cases, with an emphasis on making the models understandable as well as accurate. In general, it is often possible to prune a decision tree to obtain a simpler and more accurate tree [6, 8].

The aims of this study are to explore biomarkers of greatest impact on distant metastasis development in BC patients and their combinatorial behavioural expression patterns, and to build a decision tree algorithm for predictive tissue markers of distant metastasis which could be used, following validation, in newly diagnosed BC cases.

MATERIALS AND METHODS

Patients and tumours

This study was based on a well-characterised cohort of early stage (I-III) primary operable invasive BC (n=1902) from patients enrolled into the Nottingham Tenovus Primary Breast Carcinoma Series between 1987 and 1998, and managed in accordance to a uniform protocol and has been comprehensively studied with a broad range of markers [9, 10]. During the follow-up time within this series, distant recurrence had developed in 578/1902 cases (30 %). The median time to distant metastasis (DM) was 128 months (range 4- 247 months).

This study included 31 biomarkers of clinical and biological relevance to BC tumourigenesis and progression [9, 11]. These were: hormone receptors [estrogen receptor (ER), progesterone receptor (PR)], epidermal growth factor receptor family members [HER1 (EGFR), HER2, HER3, HER4], cytokeratins [basal CKs; CK5/6 and CK14, and luminal CKs; CK7/8, 18 and 19], tumour suppressor and cell cycle regulator proteins [p53 and P27] anti-apoptotic BCL2, a proliferation marker (Ki-67/MIB1 clone), cadherin family [E-cadherin, N-cadherin and P-cadherin], markers of key molecular pathways [TGFβ1, PIK3CA, pAkt-S473], transcription factors [phospho-STAT3 and TWIST2], markers reported to be associated with invasiveness and tumour aggressiveness [CTEN, CD44 and CD24] [12, 13], in addition to five markers/proteins encoded by five transcripts/genes significantly expressed between metastatic and non-metastatic breast cancer: (TOMM34, ZFN22, KRT23, ST8SIA6, and chromogranin-A]. These latter markers resulted from ANN of analysis of cDNA expression data of 128 primary invasive frozen BC samples from the Nottingham Tenovus Primary Breast Carcinoma series previously studied using gene expression profiling [14]. This approach stratified the transcripts on their ability to classify samples based on the occurrence of DM (n = 35) compared with those without DM (n = 93), as previously

described [15, 16]. These five proteins are encoded by genes/transcripts' data analysed by artificial neuronal network. These genes were amongst the top 40 differentially expressed genes between metastatic and non-metastatic cases. This research was approved by Nottingham Research Ethics Committee 2 under the title of ''Development of a molecular genetic classification of breast cancer''.

Immunohistochemistry (IHC)

Four-µm sections were cut from paraffin processed block of previously prepared TMAs and mounted on Superfrost slides (Surgipath). MIB1 expression was determined using full face FFPE breast tissue sections as previously described [17]. Tissue sections were deparaffinised in xylene (Genta Medica, York, UK), rehydrated in descending series of ethanol, 10 seconds each. Heat induced retrieval of antigen epitopes was carried out using microwave treatment of slides in 10 mM sodium citrate buffer (pH 6.0) for 20 minutes. Slides were then incubated primary antibody in optimal working dilution (Table 1). Secondary detection system was NovoLink™ Polymer Detection System (Leica, RE7150-K). Reaction was visualised using freshly prepared filtered solution of 3-3' Diaminobenzidine tetrahydrochloride (DAB, Dako, K3468). Counterstaining was performed with Mayer's haematoxylin (DAKO, AR106) for 6 minutes. Sections were dehydrated in alcohol, cleared in xylene, and coverslipped using DPX mounting medium (BDH, Poole, UK).

Assessment of IHC staining

Slides were scanned as high resolution digital images (0.45µm/pixel) using a NanoZoomer slide scanner (Hamamtsu Photonics, Welwyn Garden City, UK) and accessed using a web based interface (Distiller, SlidePath Ltd, Dublin, Ireland). TMA cores were scored at 20x magnification using a minimum of 24" high resolution screen (1920x1080).

Scoring of IHC staining of markers was performed using the modified H-score method [18], except MIB1LI and BCL2 which were scored as the percentage of expression. All sections were scored without prior knowledge of the patients' pathologic or outcome data.

Statistical analysis

To establish a set of rules to determine to which group; presence or absence of distant metastasis, a patient is more likely to be assigned using its variables' values; WEKA software was used to compute the decision tree algorithm C4.5. Wilcoxon test, a non-parametric version of T-test, was used to specify those markers appearing to behave differently in the two groups of patients. Results were validated using univariate Cox regression analysis. A p value < 0.05 (two-tailed) was considered significant. Box plots were organised to visualise the differential distribution of each marker between those cases with metastatic disease from those without.

RESULTS

Box plots for the distribution of all the 31 markers' expression within the studied series with relevance to the presence or absence of distant metastases is summarised in Figure 1A. According these plots, the variable distribution of markers within both patients' subsets can be inferred. For instance, the H-score of PR expression for 95% of cases with DM ranges from "0-150", median 5, compared with "0-200", median 90 in cases with no DM.

To test for those markers which were the drivers of the two groups, Wilcoxon test and univariate Cox regression analysis were performed. This resulted in a panel consisted of 15 biomarkers being significantly associated with distant recurrence ($p < 0.001$). These markers were: Ki67/MIB1LI, ER, PR, HER2, EGFR, p53, BCL2, N-cadherin, P-cadherin, PIK3CA, pSTAT3 nuclear expression, TOMM34, ZFN22, CD44 and Ck5/6. Table 2 displays these markers and the functional group under which they could be classified. Figure 1B displays box plots showing differential expression levels of this panel within metastatic versus non metastatic groups of the studied series. As these plots display, variable distribution of different markers between the two groups could be appreciated. For instance, the H-score of BCL2 expression for 95% of cases with distant recurrence lies between "0-75", median 50, compared with "0-100", median 70 in cases with no distant recurrence.

Biomarker Expression Pattern and Time to Distant Metastasis

Based on time to development of DM in cases where distant recurrence occurred (n=578 patients), six temporal subgroups were characterised, which ranged from less than one year of follow up to more than 15 years. These were: 1) < 1 year (n=50), 2) between 1-2 years (n=102), 3) > 2 up to 5 years (n= 201), 4) > 5 up to 10 years (n= 155), 5) > 10 up to 15 years (n=56), and 6) > than 15 years (n=14). Box blots were constructed to depict the expression pattern of the 15 biomarker panel significantly associated with DM within these temporal

subgroups, Figure 2. According to box plots of these metastatic subgroups, 10 markers had a significant expression pattern with respect to time of developing DM. However, the remaining five markers did not show this temporal relation with occurrence of DM.

Markers with Significant Expression Trend within Groups of Time to Distant Recurrence

Within this group of markers, Ki67/MIB1LI, HER2, p53, N-cadherin, P-cadherin, PIK3CA and TOMM34 showed characteristic pattern of differential expression between the six subgroups, where higher expression values were associated with significantly earlier development of distant recurrence, and vice versa. For PIK3CA, the expression values were very high for those cases that developed earlier DM (mean H-sore = 210 in those developed DM in less than one year compared with 130 in those developed DM > 10 years up to 15 years). In contrast, higher expression values of ER, PR, and BCL2, were observed to be associated with delayed occurrence of DM.

Markers with no observed expression trend within the time to distant recurrence period:

The remaining markers, however, (i.e. EGFR, phospho-STAT3, ZFN22, CD44 and CK5/6), did not show an evident behavioural/differential trend in their expression with relevance to the time to development of distant metastasis in the six metastatic subgroups.

Decision Tree-Calculated Metastasis Prediction Algorithm

To build a distant recurrence risk assessment algorithm for breast cancer patients, decision trees were computed. For the purpose of robustness, only cases with complete values for all biomarkers were used to compute the decision tree. The number of informative cases available for the 15 significant markers was 176 cases, which constituted a test set for build a decision tree algorithm for metastatic recurrence prediction. DM had developed in 64 (36.3%) cases, with the remaining 112 (63.7%) cases remained DM free throughout the period of follow-up. The input data for WEKA software was the expression data of these 15 markers,

entered as continuous data "H-score or % expression", and the metastatic status (Yes/No), with minimum number of cases in each branch to be equal to or more than 4. The resulting tree is as illustrated in Figure 3.

According to the tree, four groups of patients were characterised, based on the expression of three markers; p53, HER2 and BCL2, which were able to predict the probability of DM in the test set. Cut-off points for these three markers were automatically set by the software. Table 3 displays these groups and the numbers of correctly classified and misclassified cases in each of group.

This ability of this algorithm to classify patients on the basis of probability of DM (i.e. DM Yes and DM No) was significantly associated with tumour size, tumour grade, number of positive axillary lymph nodes, and BC molecular subtype as assessed by IHC [19]. Although associations with axillary nodal stage and lymphovascular invasion (LVI) did not reach statistical significance, more proportions of cases with ≥ 4 positive axillary nodes experienced DM more than those with node negative disease 30% versus 13.5%, respectively). Supplementary Table 1 displays the results of these associations. Multivariate Cox regression analysis showed that this algorithm is significantly associated with breast cancer specific survival (BCSS) and metastasis free interval [P = 0.001, Hazard ratio (HR) = 3.139, 95% interval (CI) = 1.640-6.011, and HR = 2.856, 95% CI = 1.538-5.305, respectively] independently of grade, size, stage, molecular subtype and number of positive axillary lymph nodes.

The precision rate of this algorithm in predicting distant metastasis was evaluated using the positive predictive value (PPV) and the negative predictive value (NPV) [20]. For this algorithm, the PPV = 26/32 = 81.1%, and the NPV= 150/194 = 77.3%.

DISCUSSION

Within the studied series, DM had developed in 30% of cases during the period of follow-up and the outcome of this group was markedly reduced compared to patients who did not develop DM (5-year survival rate was 58% compared to 98% respectively). This significant decline in patients' survival is, to a large extent, attributable to the biological differences in tumours with metastatic potential from those without. Therefore, the molecular factors driving growth and differentiation pathways in tumours with metatastes were scrutinised to explore their relative contribution in their non-metastatic tumour counterparts.

For this purpose, the expression pattern of 31 biomarkers with close relevance to breast cancer biology and progression was studied with relevance to the occurrence of DM. Based on their distributions as continuous variables, many of the studied markers displayed variable expression within both patients' subsets with and without DM, respectively. However, a panel formed of fifteen markers was significantly associated with the occurrence of DM.

Functional categories within this biomarker panel revealed that they belonged to molecular pathways responsible for carcinogenesis and cancer progression including: hormonal receptors (ER, PR), epidermal growth factor receptor family members (EGFR, HER2), tumour growth fraction as assessed by MIBL1, tumour suppressor p53, anti-apoptotic BCL2, cell adhesion molecules (N-cadherin, p-cadherin), signalling pathways and transcription factors (PIK3CA, pSTAT3, and ZNF22), basal Ck14, TOMM34 and CD44.

In meta-analysis of publicly available breast cancer gene expression profiling (GEP) studies, Wirapati and colleagues showed that the key biological drivers in nine prognostic signatures were proliferation related genes, in addition to ER signalling and HER2 amplification [21]. The current IHC marker panel shared ER, PR, HER2 and proliferative fraction as major

drivers of progression with GEP studies [21]. Additionally, p53 and BCL2 were among the major contributors of DM in the IHC panel. Inactive TP53, as assessed by positive p53 protein expression, disturbs the functional braking and emergency cell cycle arrest in genetically damaged cells leading ultimately to cell cycle progression [22]. On the other hand, BCL2 is a cellular pro-survival molecule that protects transformed cells from apoptotic cell death. Therefore, from functional point of view, p53 and BCL2 which were more expressed in metastatic cases, lead to enhanced cell proliferation, through cell cycle progression and cell immortalisation [23].

The expression of basal/myoepithelial markers has been previously reported to contribute to identification of a subset of breast cancer characterised by poor outcome; the basal-like subtype [24]. In the current IHC panel, two markers could be assigned into this category of markers; EGFR and Ck14. These results support recommendations of using these two makers, beside ER, PR and HER2 negativity, as additional surrogates in characterising breast cancers with basal phenotype [25]. Moreover, the roles played by PIK3CA in BC progression through its downstream effectors, especially through driving an epithelial-to-mesenchymal transition program (EMT) with up regulation of N-cadherin and P-cadherin has been reported in the same series [26, 27], as well as in others [28].

Metastatic recurrence developed in these patients with primary BC over a time schedule ranged from four months to more than 15 years. Ten markers of the IHC biomarker panel displayed differential expression patterns within the temporal groups of time to distant metastasis. These include: ER, PR, HER2, MIB1LI, p53, BCL2, N-cadherin, P-cadherin, PIK3CA and TOMM34. However, the remaining five markers did not show this sort of trend. Markers of poor prognostic impact [MIB1LI, HER2, p53, N-cadherin, P-cadherin, PIK3CA and TOMM34] were more expressed in cases where DM had manifested earlier; while less

expression was observed with prolonged metastasis-free intervals. The reverse was true for biomarkers of good prognostic impact [ER, PR, and BCL2]. These findings could be interpreted in view of the concept of tumour dormancy, in which disseminated cancer cells leaving the primary tumour, stay dormant for variable periods of times in another anatomical niche that could extend into many years. During this dormancy stage, cells may remain quiescent or form clinically undetectable micrometastases. Entering of these dormant cells or micrometastatic nodules into an overt progressive growth phase leads to the commencement of clinically detectable metastasis. The length of dormancy periods has been determined by balancing cell proliferation and apoptosis [29, 30]. More insight into the differentially expressed markers between early and late metastatic groups reveals, once again, that major drivers of breast cancer progression, especially proliferation and apoptosis regulator, not only play major roles in emergence of recurrences but also in the time of their arousal.

Using the 15 IHC biomarker panel and decision tree, a probabilistic algorithm was computed to be applied for risk assessment of DM in breast cancer patients. According to the resulting algorithm, four risk groups of patients were characterised. Expression levels of p53, HER2, and BCL2 at automatically-generated specific cut-off points were able to predict the probability of distant recurrence in the studied set with satisfactory precision rate; 81.1% PPV and 77.3% NPV. Therefore, a tumour suppressor, an oncogene, and an anti-apoptotic marker could be reliably used in DM risk prediction. In the root node of the decision tree, p53 expression was the first determinant, with distant recurrence expected in cases of high p53 (H-score > 60) with HER2 or p53 high, low HER2 and low BCL2. However, low p53 alone or high p53, low HER2 and high BCL2 were associated with lower risk of DM. The ability of this probabilistic model in stratifying patients into DM risk groups was significantly associated with tumour size, grade, number of positive axillary lymph nodes, and BC molecular subtype. Moreover, the model was significantly associated with both BCSS and

metastasis free interval, independently of other factors. This algorithm reiterates the documented prognostic values of these three markers and underscoring the central biologic role played by each of these three markers in breast cancer progression [31-33]. According to these results, this small or pruned panel of biomarkers could be used with acceptable success in distant recurrence prediction. However, its performance needs to be validated in an independent breast cancer patient series especially on a prospective basis.

Conclusions: Metastatic recurrence in the studied series appears to result from contribution of a molecular biomarker panel controlling the major nodes in carcinogenic and progression pathways including hormonal receptors, growth factors, tumour suppressor, apoptotic regulator, cell adhesion apparatus and transcription factors. A predictive algorithm formed of p53, BCL2 and HER2 IHC expression was able to successfully predict the probability of distant recurrence, which requires independent validation. These findings affirm that metastasis is an inherent early cancer trait that could be predicted from the primary tumour biomarker expression profile.

Conflict of interest: The authors have no conflicts of interest to declare.

Figures

Figure 1: Box plots for the studied markers in cases with no distant recurrence versus those with distant recurrence. The box for a certain marker represents 95% of cases. The horizontal bold line inside the box is the median value of the marker. A) Distribution of all markers (n=31) within the studied series. B: Box plots of biomarker panel showing significant expression between metastatic versus non-metastatic patients' subsets.

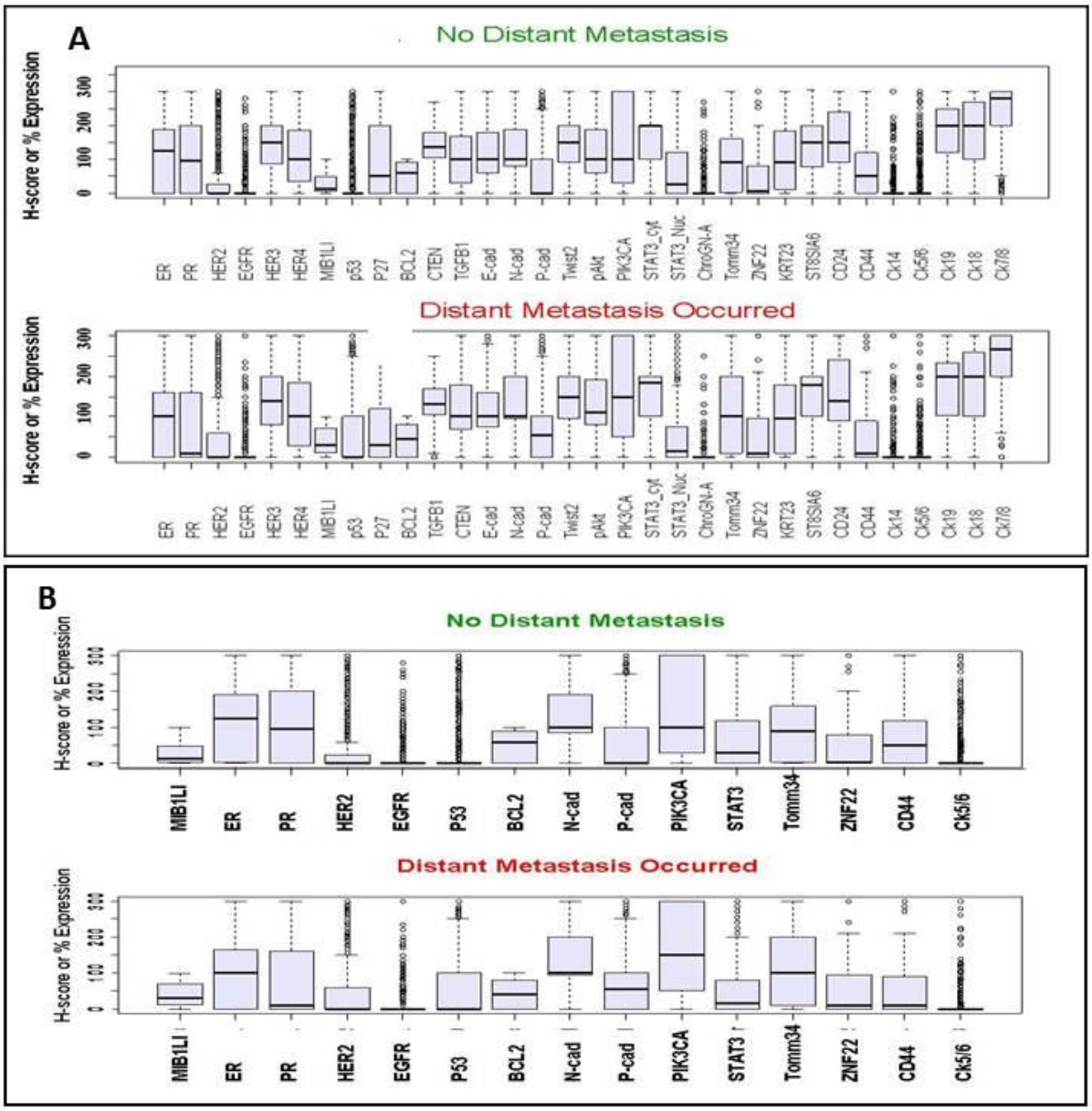

Figure 2: Box plot showing expression pattern of metastatic biomarker panel within the six temporal groups within the metastatic group based on time to DM. A: < 1 year (n=50), B: from 1-2 yrs (n=102), C: > 2 up to 5 yrs (n=201), D: > 5 up to 10 yrs (n=155), E: > 10 up to 15 yrs (n=56), and F: > 15 yrs (n=14). Y axis represents the H-score or percent expression of markers on X axis.

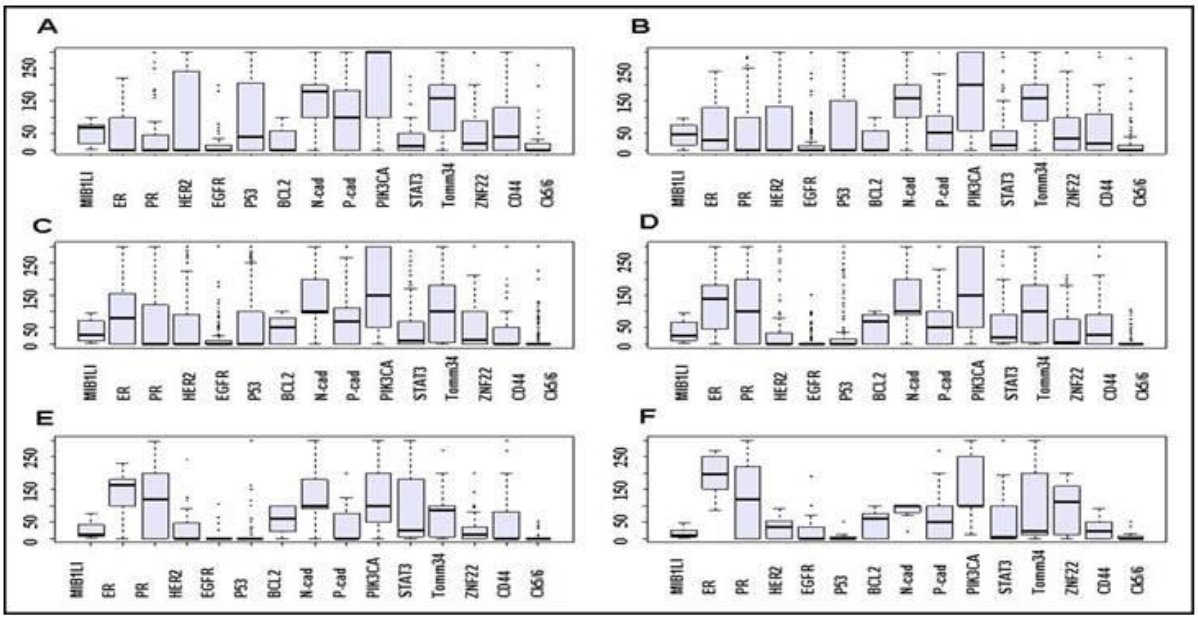

Figure 3: Decision tree algorithm for predicting distant recurrence. Circles represent the markers in the algorithm (p53, HER2, and BCL2). Rectangles represent feature value tested (DM); Zero = No DM, One = Yes DM and numbers represent subsets of patients correctly classified and misclassified, respectively. Branches emerging from each marker are levels (H-score for p53 and HER2, and percentage of BCL2) of expression below or above which a specific case is to be classified into either zero or one.

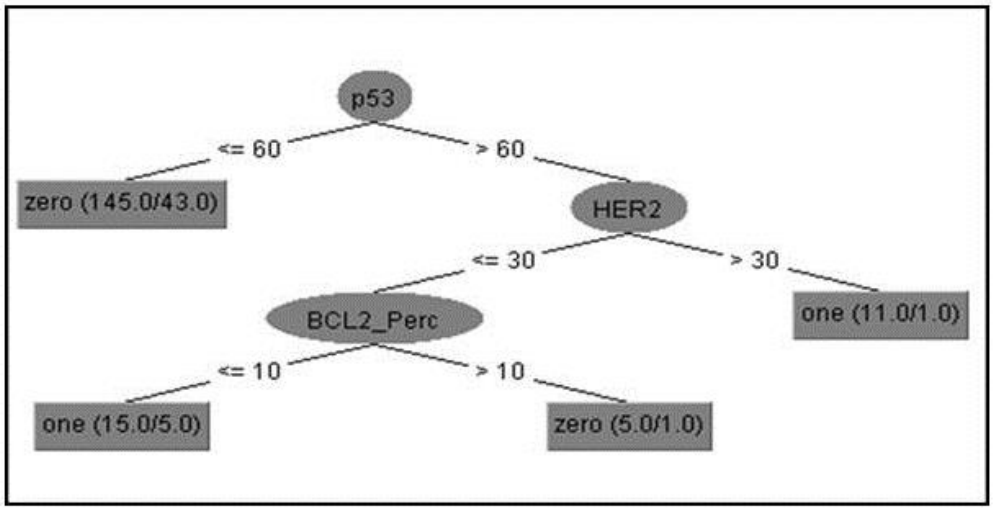

Table1: Dilution and source/clone for the antibodies used in this study.

	Marker	Clone/ Source	Dilution
1	ER	[clone SP1], Dako Corporation	1 : 150
2	PR	[clone PgR 636], Dako Corporation	1:100
3	HER2	[cerbB-2], Dako Corporation	1:250
4	EGFR	[clone EGFR.113], Novocastra	1:10
5	EGFR3	[clone RTJ1], Novocastra	1:20
6	EGFR4		
7	Ki67	[clone MIB-1], Dako Corporation	1 : 50
8	P53	[DO7], Leica Biosystems, Newcastle, UK	1 : 50
9	Bcl-2	[clone 124], Dako Corporation	1:400
10	P27	[Clone SX53G8], Dako Corporation	1:25
11	PIK3CA	Sigma-HPA009985	1:50
12	pAKT-S474	Neomarkers-RB-10369-P1	1:150
13	TGFß1	Abcam-Ab27969	1:400
14	STAT3*	Abcam-Ab76315	1:150
15	TWIST2*	Abcam-Ab57997	3μg/ml
16	E-cadherin	[clone HECD-1], Zymed Laboratories	1:100
17	N-cadherin	Sigma-C3865	4 μg/ml
18	P-cadherin	[clone 56], BD Biosciences	1:200
19	CTEN*	Abcam, ab57940	1:75
20	CD24**	SWA11 mouse antibody	1: 500
21	CD44	Cell Signalling-156-3c11	1:100
22	Ck18	[clone DC10], Dako Corporation	1:50
23	Ck19	[clone BCK 108], Dako Corporation	1:100
24	Ck7/8	[clone CAM 5.2], Becton Dickinson	1:2
25	CK5/6	[D5/16134], Boehringer Biochemica	1 : 100
26	CK14	Anti-human CK14, LL002, leicabiosystems, Newcastle, UK	1 : 100
27	Tomm34	Sigma-HPA018845	1: 100
28	KRT 23	Abnova-H00025984	4μg/ml
29	ST8SIA6	Sigma-HPA011635	1:75
30	ZFN22	Sigma-HPA016736	1: 100
31	Chromagranin	clone DAK-A3	1:100

* Microwave heat induced retrieval of antigens' epitopes was performed in citrate buffer at pH 6.0 for all the studied markers except: STAT3, CTEN, and TWIST2 for which EDTA solution at pH 8.0 were used.

**: Gift from Professor P. Altevogt, German Cancer Research Centre, Heidelberg, Germany.

Table 2: Functional categories of biomarker panel with the occurrence of distant recurrence and their descriptive measures. All markers were scored using H-score (i.e. ranged from 0-300), except MIB1LI and BCL2 which were scored as percentages (i.e. ranged from 0-100)

Tissue marker	**Distant Metastasis**		**Significance** *p* **value**
	No	**Yes**	
Hormone receptors	**Mean (Median)**		
1- ER	115 (125)	93 (100)	< 0.001
2- PR	108 (95)	78 (10)	< 0.001
EGFR family members			
3- EGFR	11 (0)	17 (0)	0.002
4- HER2	30 (0)	52 (0)	< 0.001
Proliferation markers			
5- Ki-67/MIB1LI	28 (14)	39 (30)	< 0.001
Tumour suppressor genes			
6- p53	35 (0)	55 (0)	< 0.001
Anti-apoptotic			
7- BCL2	50 (60)	40 (40)	< 0.001
Key Molecular Pathways			
8- PIK3CA	135 (100)	160 (150)	< 0.001
Cadherin Family members			
9- N-cadherin	120 (100)	130 (100)	0.018
10- P-cadherin	55 (0)	70 (55)	< 0.001
Transcription Factors			
11-Phospho-STAT3	70 (30)	55 (15)	0.003
Markers of proposed stem cell lineage			
12- CD44	80 (50)	60 (10)	< 0.001
Basal CKs			
13- Ck5/6	10 (0)	15 (0)	0.006
Gene Microarray Genes			
14- Tomm34	90 (90)	115 (100)	< 0.001
15- ZNF22	40 (5)	50 (10)	0.002

Table 3: Groups of probability of distant metastasis resulting from decision tree and numbers of patients within each group. P53 and HER2 expression were assessed as H-scores, while BCL2 was assessed as % of expression.

Group	Correctly classified	Misclassified
Distant metastasis - Yes		
1- p53 > 60 and HER2 > 30	11	1
2- p53> 60 and HER2 <=30 and BCL2 <=10	15	5
Distant metastasis - No		
3- p53 <= 60	145	43
4- p53> 60 and HER2 <= 30 and BCL2 > 10	5	1
Total	176	50

Supplementary Table 1: Statistical associations between DM predication algorithm and BC prognostic parameters.

Variables	Distant Metastasis		*p*-value (χ^2)
	Yes (%)	**No (%)**	
Tumour size (cm)			
≤ 2.0	88 (92.6)	7 (7.4)	0.003
> 2.0	62 (76.5)	19 (23.5)	(9.988)
Tumour grade			
1	17 (100)	0 (0)	<0.001
2	56 (98.2)	1 (1.8)	(18.300)
3	77 (75.5)	25 (24.5)	
Axillary nodal stage			
1	83 (86.5)	13 (13.5)	0.119
2	53 (88.3)	7 (11.7)	(4.259)
3	14 (70.0)	6 (30.0)	
Number of positive axillary lymph nodes	1.05*	2.54*	0.015 (2.683)**
Lymphovascular invasion			
Negative	99 (86.1)	16 (13.9)	0.661
Definite	51 (83.6)	10 (16.4)	(0.195)
Molecular subtype			
Luminal	107 (93.9)	7 (6.1)	<0.001
HER2 positive	19 (73.1)	7 (26.9)	(19.649)
Triple negative (Basal and non-Basal)	24 (66.7)	12 (33.3)	

*: The mean number of positive axillary lymph nodes, **: based on *t*-test.

References

1. Rabbani F, Koppie TM, Charytonowicz E, Drobnjak M, Bochner BH, Cordon-Cardo C (2007) Prognostic significance of p27(Kip1) expression in bladder cancer. *BJU Int*.
2. Nguyen DX, Massague J (2007) Genetic determinants of cancer metastasis. *Nat Rev Genet*, 8(5):341-352.
3. Pantel K, Brakenhoff RH (2004) Dissecting the metastatic cascade. *Nat Rev Cancer*, 4(6):448-456.
4. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM (2005) REporting recommendations for tumor MARKer prognostic studies (REMARK). *Nat Clin Pract Oncol*, 2(8):416-422.
5. Cruz JA, Wishart DS (2007) Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, 11;2:55-77.
6. Kotsiantis SB (2007) Supervised machine learning: A review of classification techniques. *Informatica*, 31:249-268.
7. Salzberg S (1994) C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Mach Learn*, 16(3):235-240.
8. Quinlan JR (1993) C4.5: Programs for Machine Learning. Los Altos, California.
9. Abd El-Rehim DM, Ball G, Pinder SE, Rakha E, Paish C, Robertson JF, Macmillan D, Blamey RW, Ellis IO (2005) High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses. *International journal of cancer Journal international du cancer*, 116(3):340-350.
10. Rakha EA, Elsheikh SE, Aleskandarany MA, Habashi HO, Green AR, Powe DG, El-Sayed ME, Benhasouna A, Brunet J-S, Akslen LA *et al* (2009) Triple-Negative Breast Cancer: Distinguishing between Basal and Nonbasal Subtypes. *Clinical Cancer Research*, 15(7):2302-2310.
11. Abd El-Rehim DM, Pinder SE, Paish CE, Bell JA, Rampaul RS, Blamey RW, Robertson JF, Nicholson RI, Ellis IO (2004) Expression and co-expression of the members of the epidermal growth factor receptor (EGFR) family in invasive breast carcinoma. *British journal of cancer*, 91(8):1532-1542.
12. Albasri A, Seth R, Jackson D, Benhasouna A, Crook S, Nateri AS, Chapman R, Ilyas M (2009) C-terminal Tensin-like (CTEN) is an oncogene which alters cell motility possibly through repression of E-cadherin in colorectal cancer. *J Pathol*, 218(1):57-65.
13. Ling LJ, Wang S, Liu XA, Shen EC, Ding Q, Lu C, Xu J, Cao QH, Zhu HQ, Wang F (2008) A novel mouse model of human breast cancer stem-like cells with high CD44+CD24-/lower phenotype metastasis to human bone. *Chin Med J (Engl)*, 121(20):1980-1986.
14. Naderi A, Teschendorff AE, Barbosa-Morais NL, Pinder SE, Green AR, Powe DG, Robertson JFR, Aparicio S, Ellis IO, Brenton JD *et al* (2006) A gene-expression signature to predict survival in breast cancer across independent data sets. *Oncogene*, 26(10):1507-1516.
15. Zhang H, Rakha EA, Ball GR, Spiteri I, Aleskandarany M, Paish EC, Powe DG, Macmillan RD, Caldas C, Ellis IO *et al* (2010) The proteins FABP7 and OATP2 are associated with the basal phenotype and patient outcome in human breast cancer. *Breast Cancer Res Treat*, 121(1):41-51.
16. Naderi A, Teschendorff AE, Barbosa-Morais NL, Pinder SE, Green AR, Powe DG, Robertson JF, Aparicio S, Ellis IO, Brenton JD *et al* (2007) A gene-expression signature to predict survival in breast cancer across independent data sets. *Oncogene*, 26(10):1507-1516.
17. Aleskandarany MA, Rakha EA, Macmillan RD, Powe DG, Ellis IO, Green AR (2011) MIB1/Ki-67 labelling index can classify grade 2 breast cancer into two clinically distinct subgroups. *Breast cancer research and treatment*, 127(3):591-599.

18. McCarty KS, Jr., Miller LS, Cox EB, Konrath J, McCarty KS, Sr. (1985) Estrogen receptor analyses. Correlation of biochemical and immunohistochemical methods using monoclonal antireceptor antibodies. *Arch Pathol Lab Med*, 109(8):716-721.
19. Rakha EA, Aleskandarany M, El-Sayed ME, Blamey RW, Elston CW, Ellis IO, Lee AH (2009) The prognostic significance of inflammation and medullary histological type in invasive carcinoma of the breast. *Eur J Cancer*.
20. Altman DG, Bland JM (1994) Statistics Notes: Diagnostic tests 2: predictive values. *BMJ*, 309(6947):102.
21. Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, Desmedt C, Ignatiadis M, Sengstag T, Schutz F *et al* (2008) Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res*, 10(4):R65.
22. Liu Y, Kulesz-Martin MF (2006) Sliding into home: facilitated p53 search for targets by the basic DNA binding domain. *Cell Death Differ*, 13(6):881-884.
23. Goodsell DS (2002) The molecular perspective: Bcl-2 and apoptosis. *Oncologist*, 7(3):259-260.
24. Rakha EA, Putti TC, El-Rehim DMA, Paish C, Green AR, Powe DG, Lee AH, Robertson JF, Ellis IO (2006) Morphological and immunophenotypic analysis of breast carcinomas with basal and myoepithelial differentiation. *Journal of Pathology*, 208(4):495-506.
25. Cheang MC, Voduc D, Bajdik C, Leung S, McKinney S, Chia SK, Perou CM, Nielsen TO (2008) Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. *Clin Cancer Res*, 14(5):1368-1376.
26. Aleskandarany MA, Negm OH, Green AR, Ahmed MA, Nolan CC, Tighe PJ, Ellis IO, Rakha EA (2014) Epithelial mesenchymal transition in early invasive breast cancer: an immunohistochemical and reverse phase protein array study. *Breast Cancer Res Treat*, 145(2):339-348.
27. Aleskandarany MA, Rakha EA, Ahmed MA, Powe DG, Paish EC, Macmillan RD, Ellis IO, Green AR (2010) PIK3CA expression in invasive breast cancer: a biomarker of poor prognosis. *Breast cancer research and treatment*, 122(1):45-53.
28. Zardavas D, Phillips WA, Loi S (2014) PIK3CA mutations in breast cancer: reconciling findings from preclinical and clinical data. *Breast cancer research : BCR*, 16(1):201.
29. Holmgren L, O'Reilly MS, Folkman J (1995) Dormancy of micrometastases: balanced proliferation and apoptosis in the presence of angiogenesis suppression. *Nat Med*, 1(2):149-153.
30. Townson JL, Chambers AF (2006) Dormancy of solitary metastatic cells. *Cell Cycle*, 5(16):1744-1750.
31. Abdel-Fatah TM, Powe DG, Agboola J, Adamowicz-Brice M, Blamey RW, Lopez-Garcia MA, Green AR, Reis-Filho JS, Ellis IO (2010) The biological, clinical and prognostic implications of p53 transcriptional pathways in breast cancers. *J Pathol*, 220(4):419-434.
32. Cianfrocca M, Goldstein LJ (2004) Prognostic and predictive factors in early-stage breast cancer. *Oncologist*, 9(6):606-616.
33. Callagy GM, Webber MJ, Pharoah PD, Caldas C (2008) Meta-analysis confirms BCL2 is an independent prognostic marker in breast cancer. *BMC cancer*, 8:153.