A Multimarker Model to Predict Outcome in Tamoxifen-Treated Breast Cancer Patients

Steven P. Linke,[1] Troy M. Bremer,[1] Christopher D. Herold,[1] Guido Sauter,[2] and Cornelius Diamond[1]

Abstract

Purpose: This study was designed to produce a model to predict outcome in tamoxifen-treated breast cancer patients based on clinicopathologic features and multiple molecular markers.

Experimental Design: This was a retrospective study of 324 stage I to III female breast cancer patients treated with tamoxifen for whom standard clinicopathologic data and tumor tissue microarrays were available. Nine molecular markers were studied by semiquantitative immunohistochemistry and/or fluorescence *in situ* hybridization. Cox proportional hazards analysis was used to determine the contributions of each variable to disease-specific and overall survival, and machine learning was used to produce a model to predict patient outcome.

Results: On a univariate basis, the following features were significantly associated with worse survival: high pathologic tumor or nodal class, histologic grade, epidermal growth factor receptor, ERBB2, MYC, or TP53; absent estrogen receptor (ER) or progesterone receptor; and low BCL2. CCND1 and CDKN1B did not reach statistical significance. On a multivariate basis, nodal class, ER, and MYC were statistically significant as independent factors for survival. However, the benefit of ER-positive status was moderated by BCL2, ERBB2, and progesterone receptor. BCL2 and TP53 also interacted as an independent risk factor. A kernel partial least squares polynomial model was developed with an area under the receiver operating characteristic curve of 0.90.

Conclusions: Our data show the predictive value of BCL2, ERBB2, MYC, and TP53 in addition to the standard hormone receptors and clinicopathologic features, and they show the importance of conditional interpretation of certain molecular markers. Our multimarker predictive model performed significantly better than standard guidelines.

Breast cancer is the most common malignancy in Western women, second only to lung cancer as the most common cause of cancer death. Options for the treatment of breast cancers are complex and varied, including surgery, radiation, endocrine therapy, cytotoxic chemotherapy, and antibody therapy (e.g., trastuzumab). Early-stage patients typically have their tumors surgically resected. Depending on the specific clinicopathologic features, this can range from a lumpectomy to a total mastectomy. Radiation therapy to the breast area with a boost to the tumor bed and, in some cases, to surrounding lymph node regions reduces the chance of recurrence, particularly when breast-conserving surgery is done (1).

Roughly 75% of breast cancers are positive for the hormone-based estrogen receptor (ER) and/or progesterone receptor (PGR; ref. 2). Binding of estrogen to ER causes its phosphorylation and dimerization followed by transcription of a variety of genes, including secreted growth and angiogenic factors, as well as PGR (3). Because the ER pathway can contribute to tumor growth, most hormone receptor–positive patients are treated with an inhibitory endocrine therapy either as an adjuvant to radiation and/or surgery in early-stage disease or as the primary treatment in more advanced disease. The most common endocrine therapy has been the selective ER modulator tamoxifen. It has been in use for >20 years and demonstrably prolongs survival (4).

In breast tissue, tamoxifen competes with estrogen for binding to ER, thereby reducing proliferation through inhibition of the ER pathway. However, ~40% of hormone receptor–positive patients fail to respond to tamoxifen (5–7). Tamoxifen resistance can arise through several direct mechanisms, such as defects in the ER pathway, cross-talk between the ER pathway and other growth factor pathways (6, 8, 9), and deregulated expression of ER coregulators (8, 10). Cancers may also be resistant to tamoxifen through indirect mechanisms, such as activation of aggressive pathways that enable growth, invasion, and/or metastasis.

Early identification of breast cancer patients with tamoxifen-resistant cancers is important, as the patients may be more responsive to alternative or additional therapies, such as other endocrine therapies, targeted agents, or cytotoxic drugs. There are several current standards that assign patients to risk categories based solely on hormone receptor status and/or clinicopathologic features. These include the NIH guidelines

Authors' Affiliations: [1]Prediction Sciences, La Jolla, California and [2]Department of Pathology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Received 7/20/05; revised 10/20/05; accepted 12/7/05.

Grant support: National Institute of Standards and Technology, Advanced Technology Program award 20024937.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Requests for reprints: Steven P. Linke, Prediction Sciences, 9404 Genesee Avenue, Suite 210, La Jolla, CA 92037. Phone: 858-404-0454; Fax: 858-777-3614; E-mail: slinke@predict.net.

doi:10.1158/1078-0432.CCR-05-1562

(11), the St. Gallen guidelines (12), and the Nottingham Prognostic Index (NPI; ref. 13). These standards provide three general levels of guidance regarding the recommendation of cytotoxic chemotherapy: 1, low risk, chemotherapy not recommended; 2, intermediate risk, chemotherapy encouraged but optional; 3, high risk, chemotherapy strongly recommended. Unfortunately, these standards are relatively inaccurate at classifying patients, and they produce ambiguous results for a large percentage.

In the past several years, several genomic studies have reported subsets of predictive/prognostic genes based on differential expression (e.g., refs. 14–16). Although such molecular assays show promise, numerous questions have been raised recently about the experimental and statistical methodologies used in the studies, creating concerns about their reproducibility and true superiority over the current standards (e.g., refs. 17–22).

In this study, we analyzed data on the standard hormone receptors (ER and PGR) as well as the growth factor receptors epidermal growth factor receptor (EGFR) and ERBB2. In addition, we studied the tumor suppressors CDKN1B and TP53, the antiapoptotic factor BCL2, the proliferation marker CCND1, and the MYC oncogene. Our final model included the standard markers ER, PGR, and ERBB2 as well as BCL2, MYC, and TP53. Although some studies have indicated that the latter markers have prognostic significance, no consensus has been reached on their utility. Here, we show the importance of conditional interpretation of certain markers on others due to their interdependency. In combination with the application of advanced informatics, this enabled the development of a predictive model that performed significantly better than the current standards.

Materials and Methods

Patients. A total of 324 females with stage I to III invasive breast carcinomas were analyzed. Based on a retrospective evaluation of patient data files, the patients received hormone therapy but no adjuvant cytotoxic chemotherapy. The patients were treated and evaluated at the University Hospital in Basel (Switzerland), the Women's Hospital Rheinfelden (Germany), and the Kreiskrankenhaus Lörrach (Germany) between 1985 and 2000. Follow-up data were collected by contacting the patients' attending physicians when necessary. Tamoxifen was used in nearly all of the cases (5 years was standard, but detailed information is not available), although a negligible number of patients (<2%) could have received a different hormone therapy. A subset of the patients received neoadjuvant cytotoxic chemotherapy (<1%) and/or adjuvant radiotherapy, but these were not statistically significant factors in survival. All tumors were reviewed by one pathologist (G.S.) for grade and histologic subtype. Staging variables were collected from the patient files. The patient set and some of the molecular features in this study were described previously (23, 24). Patient identities had been anonymized, and the Ethics Committee of the Basel University Clinics had approved the use of the specimens and data for research.

Immunohistochemistry. Mouse monoclonal antibodies (clone; epitope, if applicable; dilution) against ER (1D5; NH_2 terminus; 1:1,000), PGR (1A6; A/B region; 1:600), BCL2 (124; 1:1), CDKN1B (SX53G8; 1:1,000), EGFR (EGFR.113; extracellular domain; 1:20), and TP53 (DO-7; NH_2 terminus; 1:1) were used for immunohistochemical analysis. All antibodies were obtained from DAKO (Glostrup, Denmark) except PGR and EGFR, which were obtained from Novocastra (Newcastle-upon-Tyne, United Kingdom). The HercepTest kit (DAKO) was used for ERBB2.

Tissue microarrays constructed from formalin-fixed, paraffin-embedded primary tumor samples (25) were stained with a standard immunoperoxidase immunohistochemistry protocol (24). Tumors with known positivity were used as positive controls, and the primary antibodies were eliminated for negative controls. Most of the markers were scored for both intensity (on a scale of 0-3) and the estimated percentage of positively staining cells in ~10% increments. Final scores on a 0 to 3 scale (0, none; 1, weak; 2, moderate; 3, strong) were determined from a combination of these attributes for these markers. For the statistical analyses, tumors were considered positive for ER or PGR when staining was evident in ≥10% of cells. BCL2 was considered "high" when the final score was 3 (strong). Tumors were considered positive for EGFR when the final score was 1 to 3 (weak to strong). ERBB2 staining was scored only on the intensity scale of 0 to 3, and it was considered "positive" when the score was >0. This was based on our previous observation that 1+ ERBB2 staining on tissue microarrays is associated with HER-2 amplification in a high fraction of cases in our laboratory (26).

Fluorescence in situ hybridization. CCND1, ERBB2, and MYC gene amplifications were determined as described previously (23). Briefly, the tissue microarrays were proteolyzed, deparaffinized, dehydrated, and denatured. They were then subjected to standard dual-label fluorescence *in situ* hybridization with Spectrum Orange–labeled gene-specific probes and Spectrum Green–labeled centromere probe controls from chromosomes 11 (CCND1), 17 (ERBB2), and 8 (MYC) from Vysis, Inc. (Downers Grove, IL). The nuclei were counterstained with 4′,6-diamidino-2-phenylindole in antifade solution and examined by indirect fluorescence microscopy. A gene was considered amplified if the ratio of its signal number to that of the corresponding centromere was ≥2.

Cox proportional hazards analysis. Survival measures were defined as the proportions of patients who were still alive for a defined number of months after diagnosis. For overall survival, death from any cause was included. For disease-specific survival, patients who died due to a cause other than cancer were censored. All variables were studied with Cox proportional hazards analysis.

Univariate analyses were conducted on clinicopathologic features and molecular markers, and conditional analyses were conducted on marker pairs with plausible biological interactions. In some cases, the features were also examined as categorical data: pT1-2 versus pT3-4, pN0-1 versus pN2-3, TP53%, and age. For TP53% and age, a limited number of thresholds were examined based on the data distribution and values reported in previous publications to find a good threshold for categorical analysis of hazard ratios (HR). To control the false discovery rate for multiple tests, q-values were calculated from the corresponding Ps from all tests, including tests not shown in the final tables (27–29). Setting the maximum false discovery rate at <5% ($q < 0.05$), a corresponding P of 0.047 was determined, and only those features with $P < 0.047$ were considered significant.

The multivariate model was constructed from the significant results of the univariate and conditional analyses. The features were assessed with the Wald statistic, and nonsignificant ones were sequentially removed to produce the final model.

Analyses were conducted with the computer software MATLAB version R13 (The Mathworks, Inc., Natick, MA) and R (30) with the Survival package (31) and the q-value package (27–29).

Machine learning-based modeling and analysis. A machine learning-based approach was used to combine the variables identified in the multivariate Cox proportional hazards model. Because of the form of the variables in the Cox model, all of the data were coded in a binary format. To classify patients, a model was trained using the kernel partial least squares (KPLS) method with a third-order polynomial kernel and the maximum number of latent factors equal to the number of variables. This allowed up to third-order interactions between features in Mercer kernel space with a limited number of model variables (32). The model was trained in MATLAB R13 using elements from SPIDER version 1.5 (33).

We trained the multimarker KPLS model and evaluated the results based on 5-year overall survival. Patients were excluded if they were censored within 5 years due to loss of follow-up or were negative for both ER and PGR or if any of the constructed features were indeterminate due to missing data. Given its relatively low incidence, patients with unknown MYC amplification status were assumed to be MYC normal. A total of 123 patients (89 with and 34 without 5-year overall survival) met these requirements (the modeling set).

We evaluated the utility of the multimarker model using leave-one-out cross-validation. For each patient in the modeling set, a model was trained using all of the data, except for the subject of interest. After parameterization was completed, the model was applied to the patient data that had not been used during model training. This process was repeated until a validated prognosis score was obtained for each patient in the modeling set.

Receiver operating characteristic (ROC) curves were generated for the multimarker leave-one-out cross-validation results, the NPI, and the 2005 St. Gallen guidelines from the 123-patient modeling set based on 5-year overall survival. Continuous raw scores were used for both the multimarker model and the NPI, whereas three discrete risk categories were used for the St. Gallen guidelines. Bootstrapping (10,000×) was used to obtain a robust, smooth estimate of the mean ROC curve, and the area under the mean ROC curve was calculated for each prognostic index (34).

For Kaplan-Meier survival analysis of the multimarker model, 201 patients were analyzed. A final multimarker KPLS model was constructed using all 123 patients in the modeling set, and this final model was used to score 78 additional patients who previously have been censored due to loss of follow-up. For the patients in the modeling set, the leave-one-out cross-validated scores were used. Patients with multimarker scores greater than a threshold corresponding to 81% sensitivity for <5-year overall survival were assigned to a "bad" prognosis group, whereas the others were assigned to a "good" prognosis group.

Results

Clinicopathologic characteristics. The clinicopathologic features of the full set of 324 stage I to III breast cancer patients are shown in Table 1. Mean age at diagnosis was 64.3 years. Pathologic tumor class (pT) was known for all patients and is dependent on the size or invasiveness of the primary tumor. Pathologic nodal class (pN) was known for 81% of the patients and is dependent on the number of positive lymph nodes. Stage was determined by combining the pT and pN variables using the 2002 modified American Joint Committee on Cancer staging system (35). Histologic grade was determined by the Elston-modified Bloom/Richardson method (36). Consistent with other studies, the histology of the carcinomas was predominantly ductal or lobular.

Univariate analysis of clinicopathologic features and molecular marker data. In univariate Cox proportional hazards analysis of the clinicopathologic characteristics, increasing values for pT, the square root of the number of positive nodes, pN, and stage were significantly associated with shorter survival (Table 2). Patients with pT3-4, pN2-3, or histologic grade 3 were at particularly high risk of death relative to the lower classes in each case.

The use of tissue microarrays for the molecular analyses allowed them to be done simultaneously on many patients. Thus, the staining and scoring are consistent internally. For the immunohistochemistry and fluorescence *in situ* hybridization markers, values were available for 88% to 94% and 73% to 79% of the patients, respectively.

Table 1. Clinicopathologic characteristics

Variable	*n* (%)
Age (y)	
<50	16 (5)
50-59	64 (20)
60-69	92 (28)
70-79	99 (31)
≥80	53 (16)
pT (mm)	
1 (≤20)	105 (32)
2 (21-50)	155 (48)
3 (>50)	16 (5)
4 (*)	48 (15)
pN (no. positive nodes)	
0 (0)	108 (33)
1 (1-3)	96 (30)
2 (4-9)	42 (13)
3 (≥10)	17 (5)
Unknown	61 (19)
Stage	
I	42 (13)
II	142 (44)
III	79 (24)
Unknown (I-III)	61 (19)
Histologic type	
Ductal	243 (75)
Lobular	46 (14)
Tubular	8 (2)
Medullary	7 (2)
Mucinous	7 (2)
Other	13 (4)
Histologic grade (Elston-modified Bloom/Richardson method)	
1	95 (29)
2	139 (43)
3	90 (28)
Radiotherapy	
Yes	79 (24)
No	103 (32)
Unknown	142 (44)

*Any size primary tumor with chest wall or skin invasion.

In univariate Cox proportional hazards analysis (Table 2), the lack of ER or PGR or the presence of EGFR, ERBB2, or amplified MYC were all significantly associated with shorter survival. Low BCL2 (all scores below the maximum of 3) also was significantly associated with shorter survival. TP53 was significantly associated with worse outcome when the staining intensity was moderate to strong (TP53 high intensity). A high percentage (≥70%) of cells staining positively for TP53 was not significant for overall survival and was only marginally significant for disease-specific survival in this independent univariate analysis. However, it became more significant when considering interactions with BCL2. The HRs for disease-specific survival and overall survival were not statistically different in all cases when the covariate remained statistically significant for overall survival, although only the *P*s are shown

for overall survival. CCND1 and CDKN1B were not significant in this analysis, and the number of EGFR-positive patients was insufficient to assess its contribution.

Conditional analysis of molecular markers to determine interactions. Next, analyses were conducted on marker pairs with plausible biological interactions Several of the molecular markers exhibited dependencies relative to other markers, which increased their predictive values. For example, disease-specific survival and overall survival of ER-positive patients who were PGR negative and who had low BCL2 scores were not statistically different than ER-negative patients. However, if either PGR was present or the BCL2 score was high, the ER-positive patients had significantly better outcome. Independent of this observation, ER-positive patients who were both PGR positive and BCL2 high experienced even better outcome (Table 3; Fig. 1A). Similarly, ERBB2 positivity was significantly associated with worse outcome in ER-positive patients but not in ER-negative patients, although the number of ER-negative/ERBB2-positive patients was relatively small (Table 3; Fig. 1B).

There were also strong interactions between BCL2 and TP53. In the subset of patients with low TP53 staining, low BCL2 staining was significantly associated with worse disease-specific survival; in the subset of patients with high BCL2 staining, high TP53 staining was significantly associated with worse disease-specific survival. However, when one of these markers of poorer outcome (low BCL2 or high TP53) was evident, the status of the other marker did not significantly further affect outcome (data not shown). Combining these results, the presence of either low BCL2 or high TP53 or both was significantly associated with worse disease-specific survival and overall survival (Table 3; Fig. 1C).

Not surprisingly, values for TP53 intensity (i.e., the percentage of cells staining positively for TP53) and TP53 score correlated with each other in individual patients. Although any amount of TP53 staining typically is indicative of the presence of a mutant form, we observed a sudden and significant decrease in survival in patients with the highest intensity/overall score compared with those with weak or moderate values (e.g., 5-year disease-specific survival was 82-86% when TP53 intensity was 0-2 and only 53% when the intensity was 3). Based on analysis of all TP53 staining variables, we determined that 70% of cells staining positively was the most useful cutoff.

Multivariate model. A multivariate Cox proportional hazards model was constructed based on the univariate and conditional analyses. pN, age, MYC, ER (including interactions

Table 2. Univariate Cox proportional hazards analysis for 5-year survival

Variable	*n*	Disease-specific survival			Overall survival
		%	HR (95% CI)	*P*	*P*
pT (1-4)*	319	—	1.64 (1.31-2.06)	1.21E−05	7.85E−09
pT1-2	256	85			
pT3-4	63	64	3.14 (1.86-5.31)	1.90E−05	4.63E−08
√+Nodes*	260	—	1.48 (1.25-1.76)	7.40E−06	4.26E−06
pN (0-3)*	260	—	2.15 (1.62-2.84)	2.30E−08	2.58E−08
pN0-1	203	86			
pN2-3	57	59	4.62 (2.68-7.97)	3.80E−08	2.70E−10
Stage (I-III)*	260	—	2.94 (1.81-4.78)	1.30E−05	8.94E−08
Grade I-II	183	86			
Grade III	77	66	3.73 (2.15-6.47)	5.05E−07	6.36E−09
ER+	257	86			
ER−	41	59	2.94 (1.66-5.21)	0.00021	1.70E−05
PGR+	134	89			
PGR−	154	72	2.85 (1.58-5.15)	5.30E−04	0.0049
EGFR−	277	83			
EGFR+	10	32	4.7 (2.00-11.0)	9.12E−05	6.00E−04
ERBB2−	244	83			
ERBB2+	54	66	2.38 (1.35-4.19)	0.0027	0.0016
MYC normal	215	83			
MYC amplification	28	58	2.37 (1.21-4.65)	0.0096	0.0187
BCL2 high	236	84			
BCL2 low	56	66	2.47 (1.44-4.23)	0.001	1.94E−04
TP53 low intensity	236	85			
TP53 high intensity	66	67	2.36 (1.41-3.94)	7.46E−04	0.0396
TP53 <70%	263	83			
TP53 ≥70%	39	66	1.95 (1.08-3.50)	0.023	**0.0964**

Abbreviations: √+Nodes, square root of the number of positive lymph nodes; %, percentage of 5-year disease-specific survival; *P*, significance based on log-rank test statistic; boldfaced *P*, not significant ($P \geq 0.047$, $q \geq 0.05$).
*Variables treated as continuous data.

Table 3. Conditional Cox proportional hazards analysis of molecular markers for 5-year survival

Variable	*n*	Disease-specific survival					Overall survival	
		%	SE (%)	HR (95% CI)	*P* (Wald)	*P* (log rank)	*P* (Wald)	*P* (log rank)
ER−	31	51	10.0			1.76E−07		8.22E−08
When ER+								
PGR− and BCL2 low	20	69	11.0	1.11 (ND)	0.79		0.53	
PGR+ XOR BCL2 high	103	83	4.4	0.22 (0.12-0.51)	1.70E−04		1.70E−05	
PGR+ and BCL2 high	108	88	3.7	0.18 (0.09-0.39)	1.20E−05		3.40E−06	
When ER+								
ERBB2−	207	88	2.7			0.00128		0.019
ERBB2+	39	71	8.3	3.00 (1.49-6.07)	0.0022		0.022	
When ER−								
ERBB2−	25	60	10.5			**0.87**		**0.29**
ERBB2+	12	51	16.3	1.09 (ND)	**0.87**		**0.29**	
BCL2 high and TP53 <70%	240	84	2.7			5.35E−04		0.015
BCL2 low or TP53 ≥70%	48	61	7.3	2.54 (1.47-4.38)	8.30E−04		0.016	

Abbreviations: SE, SE for percentage of 5-year disease-specific survival; *P* (Wald), significance of individual covariates using Wald statistic; *P* (log rank), significance of overall submodel using log-rank statistic; XOR (exclusive OR), either one or the other but not both; ND, not determined; boldfaced *P* values, not significant ($P \geq 0.047$, $q \geq 0.05$).

with PGR and BCL2), BCL2 (including interaction with TP53), and ERBB2 (including interaction with ER) remained independent for both disease-specific survival and overall survival (Table 4). The overall *P*s (log-rank statistic) of the multivariate model were highly significant at 3.22E−12 and 3.58E−09, respectively, for disease-specific survival and overall survival.

Multimarker predictive model based on machine learning. Because of the expected degree of molecular marker interactions, we used a nonlinear regression modeling technique instead of standard logistic regression or simple discriminant analysis. Using the features and the cutoff values contained in the multivariate model, a predictive model based on 5-year overall survival was produced for hormone receptor–positive patients by a KPLS third-order polynomial with leave-one-out cross-validation. One of the best ways to compare the predictive accuracy of different models is through ROC analysis. ROC curves were plotted for the multimarker model as well as the NPI and the new 2005 St. Gallen consensus guidelines. Our multimarker model performed significantly better than these current standards (Fig. 2A; data not shown). The area under the ROC curve was 0.90 for our model, whereas it was only 0.70 for both the NPI and the 2005 St. Gallen guidelines. The 2003 St. Gallen guidelines produced an area under the ROC curve of just 0.62, and the NIH guidelines performed even more poorly (data not shown).

Direct comparisons between models can also be made at specific operating points on the ROC curves. Using the NPI high-risk threshold (score = 5.4) as a cutoff to assign patients to good and bad prognosis groups, the NPI correctly identified 81% of the patients who survived for at least 5 years by placing them in the low-risk or intermediate-risk category (specificity). However, 51% of the patients who were predicted to survive ended up dying (false negative rate), which could have led to their undertreatment. In contrast, at this same 81% specificity, the multimarker model had a false-negative rate of only 14% (Fig. 2A).

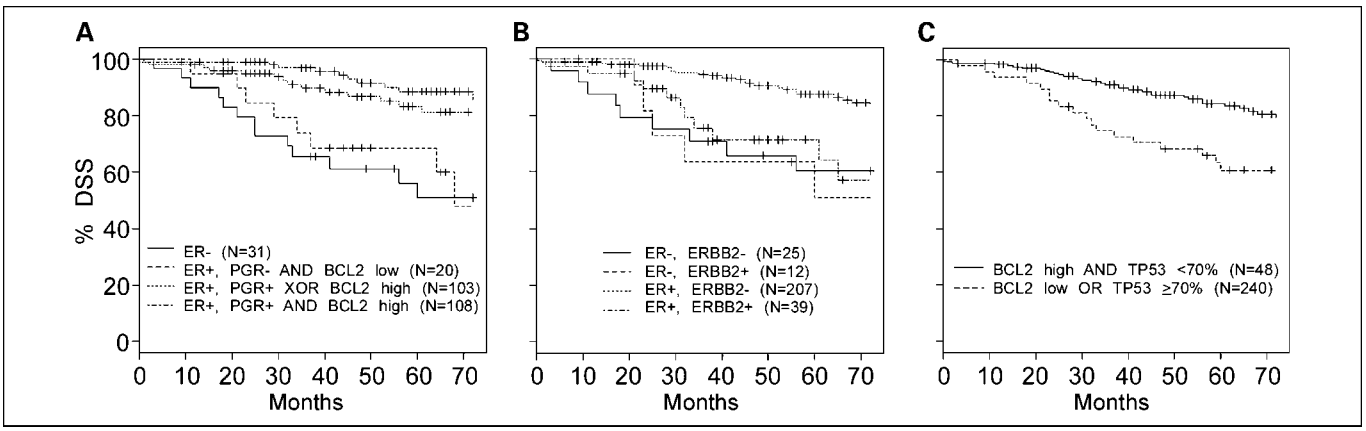

Fig. 1. Effects of molecular marker interactions on disease-specific survival (*DSS*; Kaplan-Meier plot). *A,* ER, PGR, and BCL2 interactions. *B,* ER and ERBB2 interactions. *C,* BCL2 and TP53 interactions. BCL2 low, score = 0-2; BCL2 high, score = 3; TP53 low, percentage of positively staining cells <70%; TP53 high, percentage of positively staining cells ≥70%; *XOR* (exclusive OR), either one or the other but not both; +, censored patients.

Table 4. Multivariate Cox proportional hazards analysis for 5-year survival

Variable	Disease-specific survival		Overall survival	
	HR (95% CI)	*P*	HR (95% CI)	*P*
Age ≥ 85	8.31 (1.73-39.9)	8.10E−03	16.48 (5.54-48.98)	4.60E−07
pN2-3	3.91 (1.94-7.90)	0.0001	3.35 (1.78-6.29)	0.0002
MYC amplification	3.47 (1.55-7.80)	0.0026	2.69 (1.27-5.62)	0.0096
When ER+				
PGR− and BCL2 low	0.99 (ND)	**0.98**	0.96 (ND)	**0.95**
PGR+ XOR BCL2 high	0.25 (0.088-0.72)	0.010	0.25 (0.089-0.68)	0.0067
PGR+ and BCL2 high	0.15 (0.049-0.46)	0.0008	0.20 (0.073-0.55)	0.0019
BCL2 low or TP53 ≥70%	2.53 (1.13-5.63)	0.023	1.55 (ND)	**0.24**
When ERBB2+				
ER+	3.67 (1.55-8.67)	0.0031	3.39 (1.52-7.54)	0.0028
ER−	0.50 (ND)	**0.31**	1.15 (ND)	**0.85**

NOTE: The listed variable is the positive case of the covariate, the presence of which results in the indicated HR. *P*, significance of overall submodel using log-rank statistic; XOR (exclusive OR), either one or the other but not both; boldfaced *P* values, not significant ($P \geq 0.05$).

Using the NPI low-risk threshold (score = 3.4) as a cutoff to assign patients to good and bad prognosis groups, the NPI correctly identified 81% of the patients who died within 5 years by placing them in the intermediate-risk or high-risk category (sensitivity). However, 67% of the patients who were predicted to die ended up surviving (false positives), which could have led to their overtreatment with more aggressive therapy. In contrast, at this same 81% sensitivity, the multimarker model had a false-positive rate of only 13% (Fig. 2A). Using the 81% sensitivity cutoff, Kaplan-Meier survival analysis also revealed the superiority of the multimarker model over the NPI in accurately separating patients into good and bad prognosis categories (Fig. 2B and C).

Discussion

In this study, we analyzed data on a large number of previously characterized molecular markers on a set of uniformly treated patients. Analysis of tissue microarrays allowed uniform staining and scoring, enabling accurate patient comparisons. Although tumor tissue tends to be heterogenous and the amount of tissue on tissue microarrays is limited, the results in our tissue microarrays were highly concordant with full sections and did not compromise the predictive value of the markers (24).

Our analysis revealed several marker dependencies and interactions. For example, although BCL2, ERBB2, and PGR were significant univariate factors, they provided better predictive value when considering their dependencies on ER status and/or interactions with each other. The same was true for the interaction between BCL2 and TP53. In addition, preliminary evidence suggests that different cutoffs for TP53 staining may be relevant in different patient subsets based on the status of other molecular markers. Advanced, nonlinear machine learning methodologies were used to integrate clinicopathologic features with the molecular markers into a predictive model that performed significantly better than standard guidelines.

The predictive value of ER in tamoxifen response is well established. PGR is an estrogen-regulated gene product (37). Thus, the presence of PGR may be a surrogate indicator of a functional estrogen response pathway, particularly in cases where ER is present at functional levels that are too low to detect (false negative). Consistent with several previous studies (e.g., ref. 38), we found that PGR was a predictive factor for tamoxifen treatment in univariate analysis (Table 2).

PGR negativity may arise when ER is detected but is a nonfunctional mutant or variant (false positive; ref. 38) due to signaling through alternative growth pathways, such as EGFR/ERBB2 or insulin growth factor-I receptor (39, 40). In the subpopulation of ER-positive patients, PGR negativity increased the risk of tamoxifen treatment resistance (HR, 2.1; $P = 0.02$). Thus, although PGR shows univariate statistical significance, its role in predicting tamoxifen treatment response, along with BCL2, may be better elucidated in the context of ER status (Table 4).

EGFR and ERBB2 are growth factor receptor tyrosine kinases that initiate cell survival and proliferation signaling cascades. They are elevated in ~15% and 25% of breast cancers, respectively. In the presence of the appropriate peptide growth factors, activation of these pathways may overcome the growth inhibitory effects of tamoxifen on the ER pathway. In addition, there is substantial cross-talk between the ER pathway and the ERBB2 and EGFR growth factor pathways (6, 7, 9). For example, there is evidence that various downstream members in these pathways (e.g., extracellular signal-regulated kinases 1 and 2 and AKT) can directly activate ER. Reciprocally, there is evidence that ER can directly activate members of the ERBB2 and EGFR pathways. Interestingly, binding of ER by either estrogen or tamoxifen may be sufficient for this activation. In fact, a preclinical study indicates that tamoxifen can actually stimulate cell proliferation in ERBB2-positive breast cancer cells, shifting tamoxifen from an antagonist to an agonist role (9). Consistent with this finding, ERBB2-positive patients given tamoxifen can experience even higher rates of recurrence than untreated patients (41).

In agreement with these previous studies, we found that EGFR or ERBB2 positivity were predictive of tamoxifen

resistance. Unfortunately, there were an insufficient number of EGFR-positive patients to include it in the multivariate model. Interestingly, although ER and ERBB2 levels tend to be inversely related in breast cancers, ERBB2 positivity was a statistically significant risk factor in the ER-positive patient subpopulation. This suggests that ERBB2 worked mainly by reducing the effective inhibition of the ER pathway by tamoxifen perhaps through growth factor pathway cross-talk.

Several studies indicate that a low BCL2 level is associated with worse outcome in tamoxifen-treated breast cancers (e.g., refs. 42–44). This is counterintuitive, as BCL2 is an antiapoptotic factor that might be expected to inhibit drug-induced apoptosis in the tumor cells. However, there is evidence that, similar to PGR, the BCL2 gene itself is ER regulated. Thus, high BCL2 may be indicative of an intact ER pathway that is driving tumor growth and should be sensitive to endocrine therapy (45). Furthermore, in those tumors in which it is highly expressed, BCL2 may be the leading antiapoptotic factor, so down-regulation with tamoxifen may be an effective inducer of apoptosis. In addition, in the multivariate analysis, low BCL2 predicted worse outcome independent of ER status, indicating that BCL2 participates in both ER pathway–dependent and ER pathway–independent mechanisms.

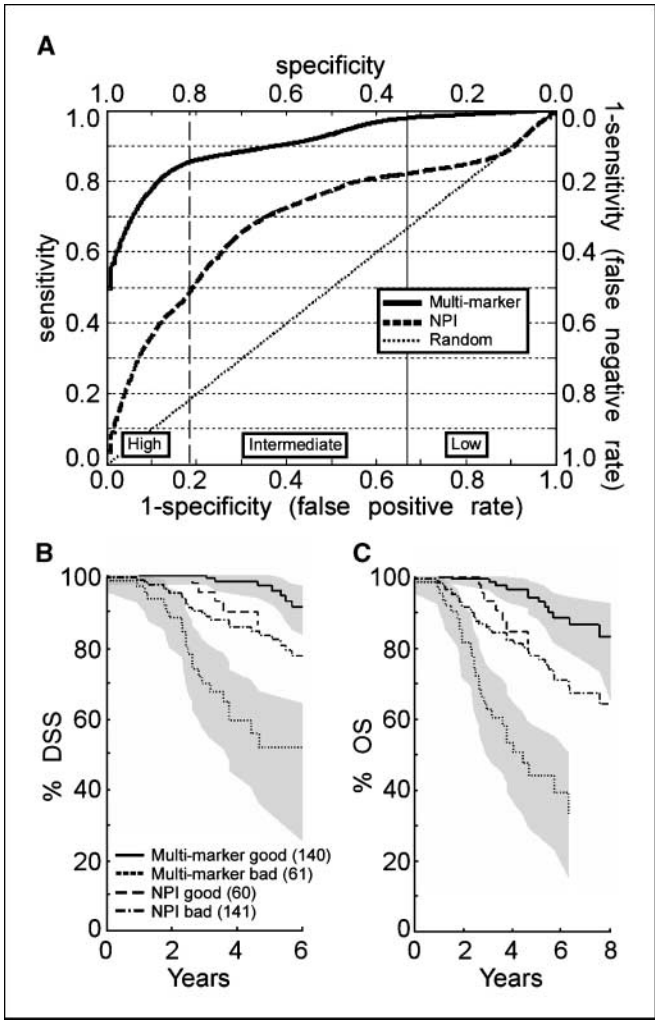

Fig. 2. Comparison of the performance of the multimarker model with the NPI. *A,* ROC curves showing prognostic accuracies based on 5-year overall survival (*OS*) in the 123-patient modeling set. Solid vertical lines and dashed lines, NPI score thresholds between low/intermediate (3.4) and intermediate/high (5.4) risk categories, respectively. *B* and *C,* Kaplan-Meier plots of disease-specific survival and overall survival, respectively. For the NPI, because chemotherapy is typically considered for intermediate-risk and high-risk patients, they were pooled into the bad prognosis category, and the low-risk patients represented the good prognosis category. For the multimarker model, a sensitivity of 81% for <5-year overall survival was chosen as the cutoff point to categorize patients into the good and bad prognosis categories based on the operating point on the NPI ROC curve that corresponded to the boundary between the NPI low and intermediate risk groups. Shaded regions, 95% CI around the multimarker model curves.

Mutations in the tumor suppressor TP53, most of which lead to elevated basal levels of the protein, are observed in ~30% of breast cancers. TP53 can be activated by stresses, such as DNA damage, leading to its regulation of genes that induce growth arrest or apoptosis through either transcription-dependent or transcription-independent mechanisms. Numerous studies show that mutant TP53 is associated with resistance to endocrine therapies, including tamoxifen (e.g., refs. 42, 46).

In this study, high TP53 predicted worse outcome independent of ER. Interestingly, this was related to the same observation made with low BCL2. Patients with either high TP53 or low BCL2 were at similar risk, but having both of the markers in this state did not further increase risk. Other studies have implemented subpopulation grouping based on BCL2 and TP53 with varying results. For example, one study reported that TP53 status was only significant in the BCL2-positive subset (44). Other reports have shown utility in separating ER-positive and ER-negative patients by TP53 and/or BCL2 status to determine subgroups with different prognoses (42, 44, 47). We saw similar, although not identical, results in our patient set. However, we determined that directly combining BCL2 and TP53 and separately assessing BCL2 in the context of ER status achieved better predictive power than these alternative methods. One confounding factor was that our study included only tamoxifen-treated patients, whereas the others also included patients who received various cytotoxic chemotherapy regimens.

Although mutations that lead to loss of TP53 function are well characterized, there is also evidence that some TP53 mutants exert gain-of-function effects. Such mutants have altered transcriptional activities and/or protein-binding targets, favoring growth and/or apoptosis resistance. Our observation that low to moderate TP53 does not significantly affect outcome compared with high TP53 raises the possibility that high levels of TP53 gain-of-function mutants can contribute to resistance independent of the ER pathway.

When amplified, MYC can inappropriately stimulate cell division through its functions in metabolism, replication, differentiation, and apoptosis (48). Approximately 11% of analyzable patients exhibited MYC amplification in this study, which is consistent with previous findings (48). Although MYC amplification is reportedly associated with ER negativity (23), it was a strong predictor of poor outcome independent of all other variables, including nodal and hormone receptor status. This may be related to the cellular functions described above, but it could also be a more general indicator of elevated genomic instability (23).

Current standard guidelines to classify patients into risk categories for recurrence include the NIH, St. Gallen, and NPI. Although these guidelines were not developed to specifically predict resistance to tamoxifen, they are used here in the same

manner they would be used by an oncologist in hormone receptor–positive patients to help determine whether these candidates for tamoxifen monotherapy are at sufficiently high risk of recurrence to justify more aggressive therapy, such as cytotoxic chemotherapy.

The NIH and older St. Gallen guidelines categorize a very large number of patients in the "intermediate-risk" or "high-risk" categories. Although this results in a very low "false negative for bad outcome" rate, it leads to the overtreatment of a sizable proportion of patients. The NPI improves performance by using an algorithm based on multivariate analysis of clinicopathologic factors from retrospective studies of breast cancer patients, and the new 2005 St. Gallen guidelines improve performance by incorporating the molecular marker ERBB2. However, they still incorrectly categorize a relatively large number of patients in the intermediate and high categories and fail to identify an important subset of higher-risk patients.

In attempts to produce superior predictive/prognostic models, several gene expression profiles are under development (e.g., refs. 14–16). However, a variety of serious questions have been raised about the experimental and statistical methodologies used in many of these studies (e.g., refs. 17–22). For example, data overfitting is a common problem, in which thousands or even tens of thousands of genes are analyzed in a relatively small number of patients. In many cases, the validation sets are not entirely independent of the training sets or they are too small to establish reliable 95% confidence intervals (95% CI) for prediction accuracy. It is also interesting to note that application of more sophisticated algorithms to the standard clinicopathologic data, such as the NPI, or the use of an artificial neural network may essentially match the performance of gene expression signatures in the same patient set (21).

Beyond the statistical issues, gene expression assays can only measure transcript levels, which do not always correlate with functional protein levels, and they cannot detect protein mislocalization. In addition, the assays are relatively complicated and costly, often requiring sophisticated and/or proprietary technology and multiple steps, including methods to try to reduce the contribution of adjacent nontumor tissue and to account for RNA degradation. Thus, several significant challenges still face this promising technology.

In this study, we used a machine learning methodology, allowing us to account for multiple marker interactions in a limited patient set for the development of a prognostic model. We employed a false discovery rate method using q-values to limit the number of false-positive identifications to compensate for multiple comparisons testing. In addition, we prevented overfitting and added robustness to the modeling process by employing leave-one-out cross-validation. We have directly shown that the model is superior to not only the cruder NIH and old St. Gallen guidelines but also the more sophisticated NPI and 2005 St. Gallen guidelines. We speculate that it also performs as well as or better than gene expression profile signatures without many of the associated drawbacks. Furthermore, the model integrates clinicopathologic features with demonstrated prognostic utility and molecular markers with established roles in drug response and general tumor aggressiveness, which can be collected with well-characterized and cost-effective immunohistochemistry and fluorescence *in situ* hybridization assays.

To produce the survival curves (Fig. 2B and C), a specific threshold value was chosen to categorize patients with good or bad prognosis with tamoxifen treatment alone. However, the multimarker model can be used to produce a risk of recurrence percentage as a continuous function of the score, so patients and their oncologists could be provided with a more specific risk rating. Alternatively, two thresholds could be chosen, one at specificity x to recommend against more aggressive therapy and the associated side effects and another at sensitivity y to recommend more aggressive therapy. Although there would be an intermediate group of patients with ambiguous scores using this strategy, the set of patients would be significantly smaller than it is with the current prognostic standards in which a majority or plurality of patients end up in the intermediate category.

Although this study has been fully cross-validated, it will be useful to independently validate it on a completely new set of patients. In addition, it will be useful to extend the study to additional markers with putative roles in tamoxifen resistance, such as ER coregulators (8), and tumor aggressiveness, such as factors reflecting proliferation, invasion, anchorage-independent growth, and/or angiogenesis. These markers could further increase the sensitivity and specificity of the predictive model. Given the mechanism of action of tamoxifen and the specific molecular markers included in this study, our model likely will contribute significantly to predictive tests for other endocrine therapies, such as aromatase inhibitors, which are becoming the standard treatment in postmenopausal women. Considering the very different costs and side effects associated with tamoxifen and aromatase inhibitor treatments, it would be interesting to compare their efficacy based on molecular modeling. Our goal is to produce a test that will enable the timely administration of the most biologically and cost-effective treatments, contributing positively to both life expectancy and quality of life for breast cancer patients.

References

1. Breast cancer (PDQ): treatment [cited 2005]. Available from: http://www.cancer.gov/cancertopics/pdq/treatment/breast/healthprofessional.
2. Osborne CK. Steroid hormone receptors in breast cancer management. Breast Cancer Res Treat 1998; 51:227–38.
3. Osborne CK, Shou J, Massarweh S, Schiff R. Crosstalk between estrogen receptor and growth factor receptor pathways as a cause for endocrine therapy resistance in breast cancer. Clin Cancer Res 2005;11: 865–70s.
4. Tamoxifen for early breast cancer: an overview of the randomised trials. Early Breast Cancer Trialists' Collaborative Group. Lancet 1998;351:1451–67.
5. Clarke R, Liu MC, Bouker KB, et al. Antiestrogen resistance in breast cancer and the role of estrogen receptor signaling. Oncogene 2003;22:7316–39.
6. Nicholson RI, Gee JM, Knowlden J, et al. The biology of antihormone failure in breast cancer. Breast Cancer Res Treat 2003;80 Suppl 1:S29–34; discussion S5.
7. Osborne CK, Schiff R. Growth factor receptor cross-talk with estrogen receptor as a mechanism for tamoxifen resistance in breast cancer. Breast 2003; 12:362–7.
8. Osborne CK, Bardou V, Hopp TA, et al. Role of the estrogen receptor coactivator AIB1 (SRC-3) and HER-2/*neu* in tamoxifen resistance in breast cancer. J Natl Cancer Inst 2003;95:353–61.
9. Shou J, Massarweh S, Osborne CK, et al. Mechanisms of tamoxifen resistance: increased estrogen receptor-HER2/*neu* cross-talk in ER/HER2-positive breast cancer. J Natl Cancer Inst 2004;96:926–35.
10. Smith CL, Nawaz Z, O'Malley BW. Coactivator and

corepressor regulation of the agonist/antagonist activity of the mixed antiestrogen, 4-hydroxytamoxifen. Mol Endocrinol 1997;11:657–66.

11. 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;93:979–89.

12. Goldhirsch A, Glick JH, Gelber RD, Coates AS, Thurlimann B, Senn HJ. Meeting highlights: international expert consensus on the primary therapy of early breast cancer 2005. Ann Oncol 2005;16:1569–83.

13. Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham Prognostic Index in primary breast cancer. Breast Cancer Res Treat 1992;22:207–19.

14. van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–6.

15. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004;351:2817–26.

16. Ma XJ, Wang Z, Ryan PD, et al. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. Cancer Cell 2004;5:607–16.

17. Lombardo JF. Molecular prediction of recurrence of breast cancer [author reply]. N Engl J Med 2005;353:1300.

18. Reid JF, Lusa L, De Cecco L, et al. Limits of predictive models using microarray data for breast cancer clinical treatment outcome. J Natl Cancer Inst 2005;97:927–30.

19. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics 2005;21:171–8.

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

21. Eden P, Ritz C, Rose C, Ferno M, Peterson C. "Good Old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. Eur J Cancer 2004;40:1837–41.

22. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003;95:14–8.

23. Al-Kuraya K, Schraml P, Torhorst J, et al. Prognostic relevance of gene amplifications and coamplifications in breast cancer. Cancer Res 2004;64:8534–40.

24. Torhorst J, Bucher C, Kononen J, et al. Tissue microarrays for rapid linking of molecular changes to clinical endpoints. Am J Pathol 2001;159:2249–56.

25. Sauter G, Simon R, Hillan K. Tissue microarrays in drug discovery. Nat Rev Drug Discov 2003;2:962–72.

26. Simon R, Nocito A, Hubscher T, et al. Patterns of HER-2/*neu* amplification and overexpression in primary and metastatic breast cancer. J Natl Cancer Inst 2001;93:1141–6.

27. Storey JD. A direct approach to false discovery rates. J Royal Stat Soc Ser B 2002;64:479–98.

28. Storey JD. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. J Royal Stat Soc Ser B 2004;66:187–205.

29. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 2003;100:9440–5.

30. R. 2.0.1 ed. Vienna (Austria): R Development Core Team.

31. R Survival. 2.15 ed. Vienna (Austria): R Development Core Team.

32. Rosipal R, Trejo LJ. Kernel partial least squares regression in reproducing kernel Hilbert space. J Machine Learn 2001;2:97–123.

33. Weston J, Elisseeff A, BakIr G, Sinz F. SPIDER: object-orientated machine learning library; 2004. Available from: http://www.kyb.tuebingen.mpg.de/bs/people/spider/.

34. Effron T, Tibshirani R. An Introduction to the bootstrap. New York: Chapman & Hall; 1993.

35. Singletary SE, Allred C, Ashley P, et al. Revision of the American Joint Committee on Cancer staging system for breast cancer. J Clin Oncol 2002;20:3628–36.

36. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology 1991;19:403–10.

37. Horwitz KB, McGuire WL. Estrogen control of progesterone receptor in human breast cancer. Correlation with nuclear processing of estrogen receptor. J Biol Chem 1978;253:2223–8.

38. Bardou VJ, Arpino G, Elledge RM, Osborne CK, Clark GM. Progesterone receptor status significantly improves outcome prediction over estrogen receptor status alone for adjuvant endocrine therapy in two large breast cancer databases. J Clin Oncol 2003;21:1973–9.

39. Cui X, Zhang P, Deng W, et al. Insulin-like growth factor-I inhibits progesterone receptor expression in breast cancer cells via the phosphatidylinositol 3-kinase/Akt/mammalian target of rapamycin pathway: progesterone receptor as a potential indicator of growth factor activity in breast cancer. Mol Endocrinol 2003;17:575–88.

40. Dowsett M, Harper-Wynne C, Boeddinghaus I, et al. HER-2 amplification impedes the antiproliferative effects of hormone therapy in estrogen receptor-positive primary breast cancer. Cancer Res 2001;61:8452–8.

41. Carlomagno C, Perrone F, Gallo C, et al. c-erb B2 overexpression decreases the benefit of adjuvant tamoxifen in early-stage breast cancer without axillary lymph node metastases. J Clin Oncol 1996;14:2702–8.

42. Silvestrini R, Benini E, Veneroni S, et al. p53 and bcl-2 expression correlates with clinical outcome in a series of node-positive breast cancer patients. J Clin Oncol 1996;14:1604–10.

43. Elledge RM, Green S, Howes L, et al. bcl-2, p53, and response to tamoxifen in estrogen receptor-positive metastatic breast cancer: a Southwest Oncology Group study. J Clin Oncol 1997;15:1916–22.

44. Gasparini G, Barbareschi M, Doglioni C, et al. Expression of bcl-2 protein predicts efficacy of adjuvant treatments in operable node-positive breast cancer. Clin Cancer Res 1995;1:189–98.

45. Perillo B, Sasso A, Abbondanza C, Palumbo G. 17β-Estradiol inhibits apoptosis in MCF-7 cells, inducing bcl-2 expression via two estrogen-responsive elements present in the coding sequence. Mol Cell Biol 2000;20:2890–901.

46. Berns EM, Klijn JG, van Putten WL, et al. p53 protein accumulation predicts poor response to tamoxifen therapy of patients with recurrent breast cancer. J Clin Oncol 1998;16:121–7.

47. Mauri FA, Maisonneuve P, Caffo O, et al. Prognostic value of estrogen receptor status can be improved by combined evaluation of p53, Bcl2 and PgR expression: an immunohistochemical study on breast carcinoma with long-term follow-up. Int J Oncol 1999;15:1137–47.

48. Deming SL, Nass SJ, Dickson RB, Trock BJ. C-myc amplification in breast cancer: a meta-analysis of its occurrence and prognostic relevance. Br J Cancer 2000;83:1688–95.

Clinical Cancer Research

A multimarker model to predict outcome in tamoxifen-treated breast cancer patients.

Steven P Linke, Troy M Bremer, Christopher D Herold, et al.

Clin Cancer Res 2006;12:1175-1183.