VOLUME 24 · NUMBER 26 · SEPTEMBER 10 2006

JOURNAL OF CLINICAL ONCOLOGY

ORIGINAL REPORT

Pharmacogenomic Predictor of Sensitivity to Preoperative Chemotherapy With Paclitaxel and Fluorouracil, Doxorubicin, and Cyclophosphamide in Breast Cancer

Kenneth R. Hess, Keith Anderson, W. Fraser Symmans, Vicente Valero, Nuhad Ibrahim, Jaime A. Mejia, Daniel Booser, Richard L. Theriault, Aman U. Buzdar, Peter J. Dempsey, Roman Rouzier, Nour Sneige, Jeffrey S. Ross, Tatiana Vidaurre, Henry L. Gómez, Gabriel N. Hortobagyi, and Lajos Pusztai

From the Departments of Biostatistics and Applied Mathematics, Pathology, Breast Medical Oncology and Radiology, The University of Texas M.D. Anderson Cancer Center, Houston, TX; Breast Cancer Unit and Unité Propre de l'Enseignement Supérieur; Equipe d'Accueil 3535 of the Institut Gustave Roussy, Villejuif, France; Albany Medical College, Albany NY; Departamento de Medicina Instituto Nacional de Enfermedades Neoplásicas, Lima, Perú.

Submitted January 11, 2006; accepted May 1, 2006; published online ahead of print at www.jco.org on August 7, 2006.

Supported by grants from the National Cancer Institute (RO1-CA106290), the Breast Cancer Research Foundation, the Gilder Foundation, the Dee Simmons Fund, and the Nellie B. Connally Breast Cancer Research Fund.

Terms in blue are defined in the glossary, found at the end of this article and online at www.jco.org.

Authors' disclosures of potential conflicts of interest and author contributions are found at the end of this article.

Address reprint requests to Lajos Pusztai, MD, DPhil, Department of Breast Medical Oncology, The University of Texas M.D. Anderson Cancer Center, Unit 1354, PO Box 301439, Houston, TX 77230-1439; e-mail: lpusztai@mdanderson.org.

© 2006 by American Society of Clinical Oncology

0732-183X/06/2426-4236/$20.00

DOI: 10.1200/JCO.2006.05.6861

ABSTRACT

Purpose
We developed a multigene predictor of pathologic complete response (pCR) to preoperative weekly paclitaxel and fluorouracil-doxorubicin-cyclophosphamide (T/FAC) chemotherapy and assessed its predictive accuracy on independent cases.

Patients and Methods
One hundred thirty-three patients with stage I-III breast cancer were included. Pretreatment gene expression profiling was performed with oligonecleotide microarrays on fine-needle aspiration specimens. We developed predictors of pCR from 82 cases and assessed accuracy on 51 independent cases.

Results
Overall pCR rate was 26% in both cohorts. In the training set, 56 probes were identified as differentially expressed between pCR versus residual disease, at a false discovery rate of 1%. We examined the performance of 780 distinct classifiers (set of genes + prediction algorithm) in full cross-validation. Many predictors performed equally well. A nominally best 30-probe set Diagonal Linear Discriminant Analysis classifier was selected for independent validation. It showed significantly higher sensitivity (92% *v* 61%) than a clinical predictor including age, grade, and estrogen receptor status. The negative predictive value (96% *v* 86%) and area under the curve (0.877 *v* 0.811) were nominally better but not statistically significant. The combination of genomic and clinical information yielded a predictor not significantly different from the genomic predictor alone. In 31 samples, RNA was hybridized in replicate with resulting predictions that were 97% concordant.

Conclusion
A 30-probe set pharmacogenomic predictor predicted pCR to T/FAC chemotherapy with high sensitivity and negative predictive value. This test correctly identified all but one of the patients who achieved pCR (12 of 13 patients) and all but one of those who were predicted to have residual disease had residual cancer (27 of 28 patients).

J Clin Oncol 24:4236-4244. © 2006 by American Society of Clinical Oncology

INTRODUCTION

Despite the critical importance of selecting the most effective adjuvant/neoadjuvant chemotherapy for an individual, diagnostic tests to guide selection of the optimal regimen for a particular patient are lacking.[1-4] Estrogen receptor (ER) –negative status, high grade, and high proliferative activity are histologic characteristics that tend to indicate more chemotherapy-sensitive cancer.[5-7] However, these clinicopathologic variables predict general chemotherapy sensitivity and therefore, have little potential to guide selection of a specific regimen. Neoadjuvant (preoperative) chemotherapy provides an opportunity to directly assess tumor response to therapy. Furthermore, complete eradication of all invasive cancer from the breast and regional lymph nodes, pathologic complete response (pCR), is associated with excellent long-term cancer-free survival.[8,9] Our goal was to evaluate gene expression profiling as a potential tool to predict who may achieve pCR to sequential anthracycline paclitaxel preoperative chemotherapy. We selected a complex multidrug regimen for study because combination chemotherapy represents the current clinical standard for patients who require systemic cytotoxic treatment. Also, gene signatures that are predictive of response to individual drugs may not fully capture sensitivity

Downloaded from jco.ascopubs.org on October 20, 2008 . For personal use only. No other uses without permission.
Copyright © 2006 by the American Society of Clinical Oncology. All rights reserved.

to combination chemotherapy. In an earlier small study (n = 42) we reported that it is possible to perform gene expression profiling on fine-needle biopsies of newly diagnosed breast cancer and that a multigene predictor of pCR has been developed promising predictive performance.[10] The current study represents an extension of the earlier work. We used a larger sample size for predictor discovery (n = 82) and used a commercially available standard gene expression profiling technology. We systematically examined the performance of a large number of potential predictors in cross validation in the training data and selected one final predictor for independent validation in 51 new cases. In the current study, we also examined the reproducibility of prediction results in 31 replicate experiments.

PATIENTS AND METHODS

Patients and Samples

This biomarker discovery trial was conducted at the Nellie B. Connally Breast Center of the University of Texas M.D. Anderson Cancer Center (MDACC) in Houston, TX, and at the Instituto Nacional de Enfermedades Neoplásicas (INEN) in Lima, Peru. During this research, patients were asked to undergo pretreatment fine-needle aspiration (FNA) of the primary breast tumor or ipsilateral axillary metastasis before starting chemotherapy as part of an ongoing pharmacogenomic marker discovery program.[11] The aspiration was performed using a 23- or 25-gauge needle. Cells from two to three passes were collected in vials containing 1 mL RNA later solution (Ambion, Austin, TX) and stored at −80°C. FNA samples on average contain 80% neoplastic cells and contain little or no stromal cells or normal breast epithelium.[12] Gene expression data generated from FNAs capture the molecular characteristics of the invasive cancer, including the molecular class.[13] Approximately 70% of all aspirations yielded at least 1 μg total RNA required for the gene expression profiling. The main reason for failure to obtain sufficient RNA was acellular aspirations. One hundred thirty-one consecutively accrued patients with at least 1 μg RNA were included in this analysis. All patients received 24 weeks of sequential paclitaxel and fluorouracil-doxorubicin-cyclophosphamide preoperative chemotherapy. Metallic markers were placed under radiologic guidance in the shrinking tumor bed for any patient interested in breast-conserving surgery, whose tumor became less than 2 cm measured by ultrasonogram or mammogram during the course of treatment. At the completion of neoadjuvant chemotherapy, all patients had modified radical mastectomy or lumpectomy and sentinel lymph node biopsy or axillary node dissection as determined appropriate by the surgeon. Grossly visible residual cancer was measured and representative sections of the cross sectional area were submitted for histopathologic study. When there was not grossly visible residual cancer, the slices of the specimen were radiographed, and all areas of radiologically and/or architecturally abnormal tissue were entirely submitted for histopathologic study. pCR was defined as no residual invasive cancer in the breast or lymph nodes. Residual in situ carcinoma without invasive component was also considered a pCR. This study was approved by the institutional review boards of MDACC and INEN, and all patients signed an informed consent for voluntary participation. Clinical characteristics of the patients are presented in Table 1.

Table 1. Clinical Information and Demographics of the 133 Patients Included in the Study

	Training Set		Validation Set	
	No.	%	No.	%
Female	82	100	51	100
Age, years				
Median	52		50	
Range	29-79		28-73	
Race/ethnicity				
White	56	68	30	59
African American	11	13	3	6
Asian	7	9	2	4
Hispanic	6	7	16	31
Mixed	2	2	0	0
Histology				
Invasive ductal	73	89	51	100
Mixed ductal/lobular	6	7	0	
Invasive lobular	1	1	0	
Invasive mucinous	2	2	0	
TNM stage				
T1	7	9	6	12
T2	46	56	24	47
T3	15	18	7	14
T4	14	17	14	27
N0	28	34	12	23
N1	38	46	25	49
N2	8	10	6	12
N3	8	10	8	16
Nuclear grade (MBMN)				
1	2	2	0	0
2	23	37	24	47
3	35	61	27	53
ER positive*	35	43	35	69
ER negative	47	57	16	31
HER-2 positive†	25	30	8	16
HER-2 negative	57	70	42 (1 unknown)	82
Neoadjuvant therapy				
Weekly T × 12 + FAC × 4	69	84	46	90
3-weekly T × 4 + FAC × 4	13	16	5	10
Pathologic complete response	21	26	13	26
Residual disease	61	74	38	74

Abbreviations: MBMN, modified Black's nuclear grade; ER, estrogen receptor; T, paclitaxel; FAC, fluorouracil, doxorubicin, and cyclophosphamide.
*Cases where ≥ 10% of tumor cells stained positive for ER with immunohistochemistry (IHC) were considered positive.
†Cases that showed either 3+ IHC staining or had gene copy number >2.0 were considered HER-2 "positive."

RNA Extraction and Gene Expression Profiling

RNA was extracted from FNA samples using the RNAeasy Kit (Qiagen, Valencia, CA). The amount and quality of RNA were assessed with DU-640 UV Spectrophotometer (Beckman Coulter, Fullerton, CA), and they were considered adequate for further analysis if the optical density $_{260/280}$ ratio was ≥ 1.8 and the total RNA yield was ≥ 1μg. Of the 133 RNA specimens used in this study, 33 were also included in the previous pharmacogenomic analysis.[10] These 33 cases were profiled on both the Affymetrix U133A chip (Santa Clara, CA) and the proprietary cDNA array. The results of the cross platform comparison were published previously.[14] cRNA generation and second-strand cDNA synthesis were performed as described previously.[12-14] No second round amplification was performed. Thirty-one total RNA specimens were split, labeled, and hybridized in duplicates several months apart in the same and in a different laboratory to assess technical reproducibility of gene expression–based predictions.

Data Analysis

dCHIP V1.3 (http://www.dchip.org) software was used to generate probe level intensities and quality measures including median intensity, percent of probe set outliers, and percent of single probe outliers for each chip. This program normalizes all arrays to one standard array that represents a chip with median overall intensity. This reference chip and the normalization procedure is available online (http://www.bioinformatics.mdanderson.org/pubdata.html). Normalized gene expression values were transformed to the $\log_{10}$ scale for analysis. To identify differentially expressed genes between cases with pCR and those with residual disease (RD) genes, two-sample, unequal-variance *t* tests were performed and genes rank ordered by *P* values.

Downloaded from jco.ascopubs.org on October 20, 2008 . For personal use only. No other uses without permission.
Copyright © 2006 by the American Society of Clinical Oncology. All rights reserved.

Table 2. Top 31 Differentially Expressed Probe Sets by Unequal-Variance *t* Test (n = 82, false discovery rate ≤ 0.5%)

Rank by *t* Test	*t* Test	*t* Test *P* Value	Higher Expression in	Gene Symbol	Affymetrix Probe Set ID	Transcript ID	Gene Bank ID	Gene Name
1	6.6019	.00000001	No-pCR	*MAPT*	203929_s_at	Hs.101174.1	AI056359	Microtubule-associated protein τ
2	6.2428	.00000002	No-pCR	*MAPT*	203930_s_at	g8400712	NM_016835	Microtubule-associated protein τ
3	6.0259	.00000008	No-pCR	*BBS4*	212745_s_at	Hs.26471.0	AI813772	Bardet-Biedl syndrome 4
4	5.9166	.00000008	No-pCR	*MAPT*	203928_x_at	Hs.101174.1	AI870749	Microtubule-associated protein τ
5	6.0604	.00000010	No-pCR	*THRAP2*	212207_at	Hs.4084.0	BG426689	Thyroid hormone receptor associated protein 2
6	5.6999	.00000027	No-pCR	*MGC5370*	217542_at	Hs.168732.0	BE930512	Hypothetical protein MGC5370
7	5.6116	.00000028	No-pCR	*MAPT*	206401_s_at	g338684	J03778	Microtubule-associated protein τ
8	5.5118	.00000047	No-pCR	—	215304_at	Hs.159264.0	U79293	Human clone 23,948 mRNA sequence
9	5.4961	.00000047	No-pCR	*ZNF552*	219741_x_at	g13443019	NM_024762	Zinc finger protein 552
10	5.474	.00000051	No-pCR	*RAMP1*	204916_at	g5032018	NM_005855	Receptor (calcitonin) activity modifying protein 1
11	5.6112	.00000055	No-pCR	*BECN1*	208945_s_at	Hs.12272.0	AF139131	Beclin 1 (coiled-coil, myosin-like BCL2 interacting protein)
12	−5.9508	.00000056	pCR	*BTG3*	213134_x_at	Hs.77311.1	AI765445	BTG family, member 3
13	5.4395	.00000068	No-pCR	*SCUBE2*	219197_s_at	Hs.222399.0	NM_020974	Signal peptide, CUB domain, EGF-like 2
14	−5.8788	.00000070	pCR	*MELK*	204825_at	g7661973	NM_014791	Maternal embryonic leucine zipper kinase
15	−5.8634	.00000096	pCR	*BTG3*	205548_s_at	g5802989	NM_006806	BTG family, member 3
16	5.2823	.00000129	No-pCR	*AMFR*	202204_s_at	g5931954	NM_001144	Autocrine motility factor receptor
17	5.1606	.00000178	No-pCR	*CTNND2*	209617_s_at	g2661061	AF035302	Catenin (cadherin-associated protein), delta 2 (neural plakophilin-related arm-repeat protein)
18	5.349	.00000250	No-pCR	*GAMT*	205354_at	g7549759	NM_000156	Guanidinoacetate *N*-methyltransferase
19	5.0802	.00000282	No-pCR	*CA12*	204509_at	g8923149	NM_017689	Carbonic anhydrase XII
20	5.1848	.00000311	No-pCR	*FGFR1OP*	214124_x_at	Hs.108548.1	AL043487	FGFR1 oncogene partner
21	5.0679	.00000337	No-pCR	*KIAA1467*	213234_at	Hs.6189.0	AB040900	KIAA1467 protein
22	5.3029	.00000364	No-pCR	*MTRN*	219051_x_at	g13128999	NM_024042	Meteorin, glial cell differentiation regulator
23	5.2581	.00000374	No-pCR	*FLJ10916*	219044_at	g8922765	NM_018271	Hypothetical protein FLJ10916
24	−5.2639	.00000381	pCR	*E2F3*	203693_s_at	g12669913	NM_001949	E2F transcription factor 3
25	4.9635	.00000540	No-pCR	*ERBB4*	214053_at	Hs.390729	AW772192	V-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian)
26	5.0047	.00000559	No-pCR	*JMJD2B*	215616_s_at	Hs.301011.2	AB020683	Jumonji domain containing 2B
27	−4.9078	.00000618	pCR	*RRM2*	209773_s_at	g12804874	BC001886	Ribonucleotide reductase M2 polypeptide
28	4.824	.00000668	No-pCR	*FLJ12650*	219438_at	g13375663	NM_024522	Hypothetical protein FLJ12650
29	4.8992	.00000715	No-pCR	*GFRA1*	205696_s_at	g4885268	U97144	GDNF family receptor α 1
30	5.0633	.00000718	No-pCR	*IGFBP4*	201508_at	g10835020	NM_001552	Insulin-like growth factor binding protein 4
31	5.0448	.00000748	No-pCR	*KIF3A*	213623_at	Hs.43670.0	NM_007054	Kinesin family member 3A

Abbreviation: pCR, pathologic complete response.

Beta uniform mixture analysis (BUM) of the *P* values showed a nonuniform distribution and was used to estimate false discovery rates (FDRs).[15] We constructed multigene classifiers using combinations of the most informative genes and several different class prediction algorithms including support vector machines with linear, radial, and polynomial kernels (SVM), Diagonal Linear Discriminant Analysis (DLDA), and K-nearest neighbor (KNN) using Euclidean distance.[16] Monte Carlo cross validation (CV) was performed by repeated iteration (n = 100) of stratified random sampling to estimate the prediction performance of the different classifiers in the training data and to facilitate selection of a single classifier for independent validation. Stratification was performed to insure that the relative proportion of outcomes sampled in both cross-validation training and test sets was similar to the original proportions for the full training data. We performed complete CV including gene selection in each iteration to avoid selection bias.[17]

To assess whether the performance of a chosen predictor differed significantly from what chance alone could produce, a random-label permutation test was performed.[18] During this process, the outcome label of each case (ie, pCR *v* RD) was randomly reassigned. One thousand such data sets with randomly permutated labels were created, and predictors were generated with repeat selection of the top probes with each iteration. The observed prediction error rate for the data set is compared with the distribution of the error rates observed with the randomly permutated data sets to calculate a permutation *P* value.

Classifier performance on the validation data were assessed by using the area under the receiver operating characteristic (ROC) curve (AUC) and its compliment, the area above the curve (AAC; AAC = 1−AUC). The ROC curve is a graphical display of the false-positive rate and the true-positive rate from multiple classification rules.[19] The ROC curve arises when a continuous predictor value is calculated for each subject for a broad range of thresholds. A case is called test-positive (eg, predicted to have pCR) if the threshold is above a defined value. The total area under the ROC curve is a summary measure of the test's ability to correctly classify those with and without the outcome of interest. An AUC of 1 represents a perfect test; an AUC of 0.5 represents a test no better than random prediction.

We also evaluated the performance of a multivariate clinical predictor. The predictor utilized clinical variables only to predict pCR and was

Downloaded from jco.ascopubs.org on October 20, 2008 . For personal use only. No other uses without permission.
Copyright © 2006 by the American Society of Clinical Oncology. All rights reserved.

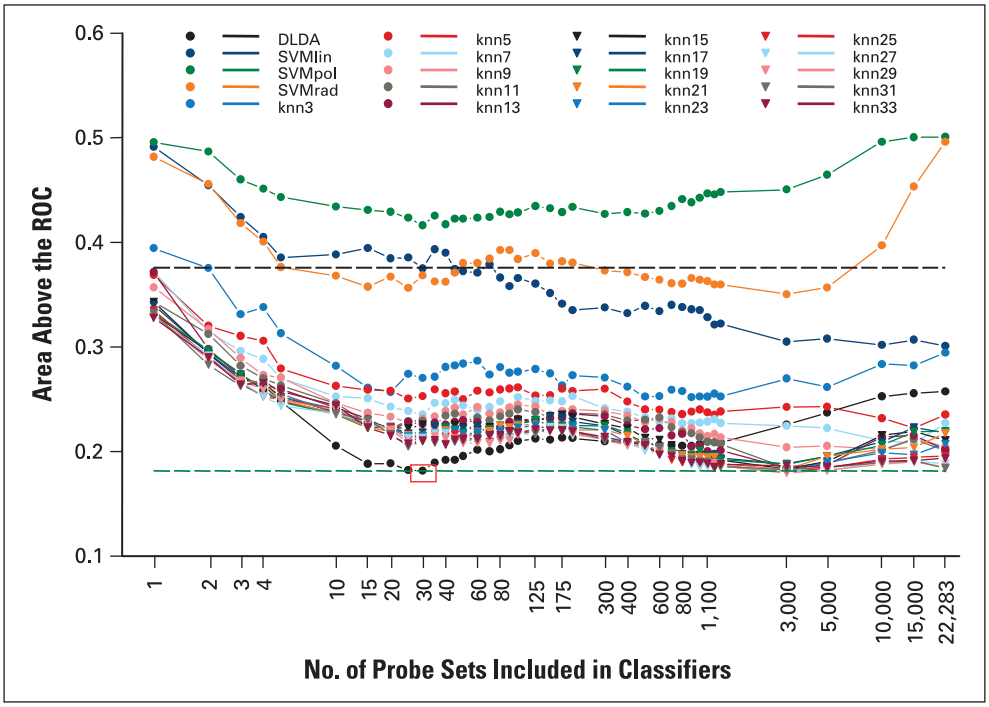

Fig 1. Mean area above the receiver operating characteristic (ROC) curves plotted against the number of top genes included in the classifiers. Complete 5-fold cross validation results (means over the 100 iterations) for 20 classifier algorithms including different numbers of probe sets (39 gene sets) are shown. Green and black horizontal dotted lines indicate the mean +/− 2SD for the nominally best Diagonal Linear Discriminant Analysis (DLDA) classifier with 30 probe sets that was selected for independent validation. polynomial kernels (SVM), and K-nearest neighbor

constructed similarly to the genomic predictor. Informative variables were identified through logistic regression performed on the training set, and a multivariate predictor was constructed using a DLDA machine learning algorithm identical to that used for the pharmacogenomic predictor. A three-variable–based (nuclear grade, age, and ER status) DLDA prediction model was tested on the independent validation set and its performance compared to the multigene predictor using ROC analysis. A combined clinical plus genomic model was also assessed that included age, dichotomized versions of ER status determined by routine immunohistochemistry ($\geq$ 10% of cells with positive nuclear staining versus $<$ 10%) and grade (grade 3 *v* grades 1 or 2) as variables in addition to the genes already included in the pharmacogenomic prediction models.

Predictor learning was also evaluated for selected pharmacogenomic models. One hundred and twenty cases from the training and validation sets were included in this analysis. Nine different training sample sizes ranging from 20 to100 by increments of 10 were created. The test set size was kept constant at 20. Stratified sampling was used to preserve the ratio of pCR and RD cases in each training and test sets. For each training set, feature selection was repeated to identify the top 30 informative genes. The area above the ROC curve and the misclassification error rate were calculated for 50 random sample sets generated for each nine training set sizes (n = 20, 30, 40, . . .100). The following learning curve model was fit to the resulting AAC and misclassification error rate (MER) values:

$$Y = a + (b^{*}TrainingSize^{c})$$

Gene expression data are available at http://bioinformatics.mdanderson.org/pubdata.html.

RESULTS

Pathologic Response Rate and Selection of Informative Genes

The overall pCR rate in the 133 patients was 26% (n = 34), which is consistent with results from a larger randomized study using the same preoperative therapy.[20] The first 82 cases were used as a training set (including 21 cases of pCR) to develop pharmacogenomic and clinical predictors. This training set size was determined by fitting learning curves to gene expression data (see below). The next 51 consecutive cases, including 13 cases of pCR, were used as independent test set to assess accuracy. In univariate analysis including clinical variables, age, nuclear grade, and ER status were significantly associated with pCR in the training set. In a logistic regression model including age, pretreatment T stage, N stage, nuclear grade, ER status, and HER2 status as predictors, only ER status (P = .0037) and age (P = .012) remained significant. To select informative genes for response prediction, we compared gene expression data from cases with pCR to those with RD in the training set. Setting the FDR to 5% resulted in 395 probe sets, 1% in 56 probe sets (corresponding to 49 genes) and 0.5% in 31 probe sets (27 genes). Table 2 presents the top 31 probe sets that were differentially expressed between the 2 response groups at an FDR of 0.5%. When similar analyses were performed separately for ER-positive (n = 35) and ER-negative cases (n = 47), BUM analysis of P values indicated the possibility of high false discovery rates even with small P values. Small sample size and relatively few events, particularly in the ER-positive group (three pCR), may have prevented the identification of differentially expressed genes with confidence within these subsets. Indeed, when predictors were constructed from genes separately associated with pCR in ER-negative and ER-positive cancers, these predictors were inferior to predictors that used genes selected from all cases.

Development of a Multigene Predictor of Pathologic Complete Response

We used gene expression data from the first 82 consecutively enrolled patients to develop a pharmacogenomic predictor. There is no consensus on what the best statistical method (if any) is to develop the most efficient class predictor from gene expression data. Therefore, we evaluated the performance of 780 different class predictors

Downloaded from jco.ascopubs.org on October 20, 2008 . For personal use only. No other uses without permission.
Copyright © 2006 by the American Society of Clinical Oncology. All rights reserved.

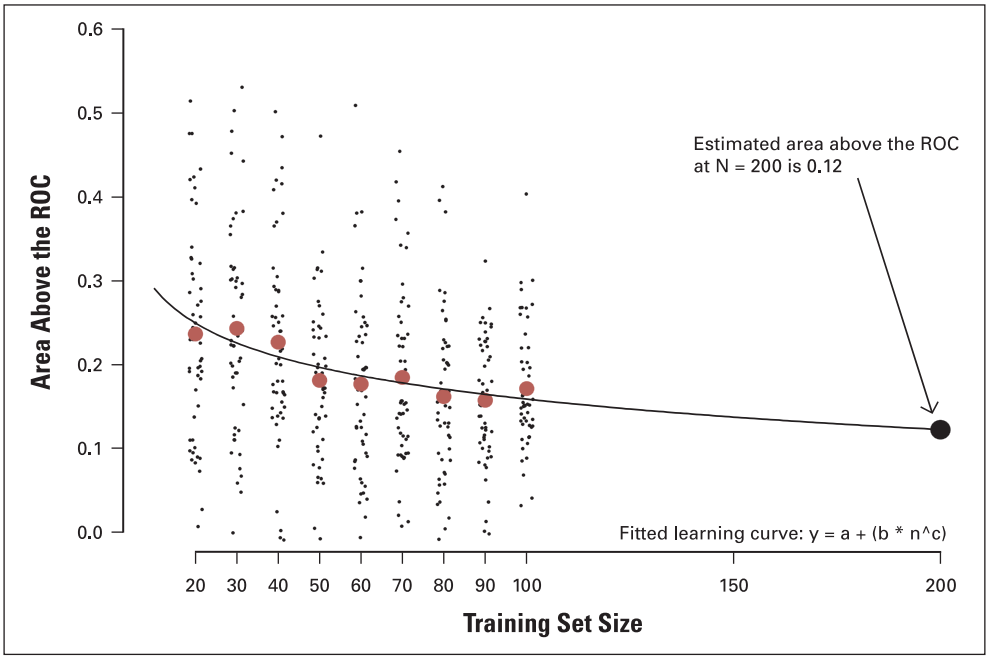

Fig 2. Learning curve for Diagonal Linear Discriminant Analysis–30 classifier. Mean area above the receiver operating characteristic (ROC) curves (AAC) developed from 20 to 100 cases in increments of 10 are shown. Fifty individual point estimates of AAC (small dots) and their means (large dots) are plotted for each of the training set sizes. The projected AAC is 0.12 if 200 cases.

that represented 20 different classification methods (DLDA, SVM with linear, polynomial and radial kernels, KNN with k = 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33) in combination with 39 distinct gene sets spanning the range 1 to 22,283 probes ranked by *P* values and spaced approximately equally on the log scale. Figure 1 shows the area above the ROC (AAC) results for each of these predictors derived from the means of 100 iterations of five-fold complete CV plotted against the number of genes included in the classifiers. The SVM classifiers did worse than the others in this particular data set. The performance of the DLDA and KNN classifiers improved with increasing numbers of genes leveling off at about 80 genes. Among the classifiers with fewer than 80 genes, DLDA did slightly better than KNN achieving the best performance at 26 genes corresponding to 30 probe sets that are shown on Table 2. In five-fold cross validation in the training set, this predictor showed mean area above the ROC curve (AAC) of 18% (95%CI: 0% to 37%), MER 27% (6% to 47%), sensitivity 75% (35% to 100%), specificity 73% (48% to 97%), and positive and negative predictive values (NPVs) of 50% (20% to 79%) and 90% (75% to 100%), respectively. To determine if the 30 probe set DLDA classifier performed significantly better than chance we did permutation testing in cross validation. None of the 1000 permuted data sets had performance as high as or higher than that calculated from the original class labels. This 30-probe set predictor (DLDA-30) was selected for independent validation. However, the mean AAC point estimates for most of the other classifiers fall within the 95% confidence interval of the DLDA-30 results. This indicates that picking the best classifier is a somewhat arbitrary process because many prediction methods and gene sets show statistically equal performance in complete cross validation.

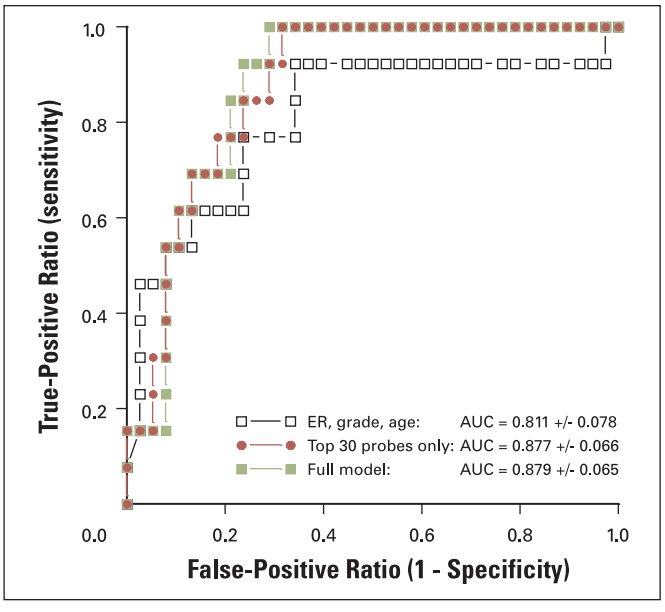

Fig 3. Receiver operating characteristic curves of three distinct pathologic complete response prediction models. The performance of the Diagonal Linear Discriminant Analysis–30 predictor and a predictor based on clinical variables and a combined clinical + pharmacogenomic prediction model are shown in the validation set (n = 51). ER, estrogen receptor; AUC, area under the curve.

Table 3. Performance Metrics of the Genomic and Clinical Predictors in the Validation Set (n = 51)

	Clinical Variables*		DLDA-30 Probe Sets	
Metric	Estimate	95% CI	Estimate	95% CI
Accuracy	0.78	0.65 to 0.89	0.76	0.62 to 0.87
Sensitivity	0.61	0.32 to 0.86	0.92	0.64 to 1.0
Specificity	0.84	0.69 to 0.94	0.71	0.54 to 0.85
PPV	0.57	0.29 to 0.82	0.52	0.31 to 0.73
NPV	0.86	0.71 to 0.95	0.96	0.82 to 1.0

Abbreviations: DLDA-30, Diagonal Linear Discriminant Analysis-30; PPV, positive predictive value; NPV, negative predictive value.
*Age, estrogen receptor status, and nuclear grade.

Downloaded from jco.ascopubs.org on October 20, 2008 . For personal use only. No other uses without permission.
Copyright © 2006 by the American Society of Clinical Oncology. All rights reserved.

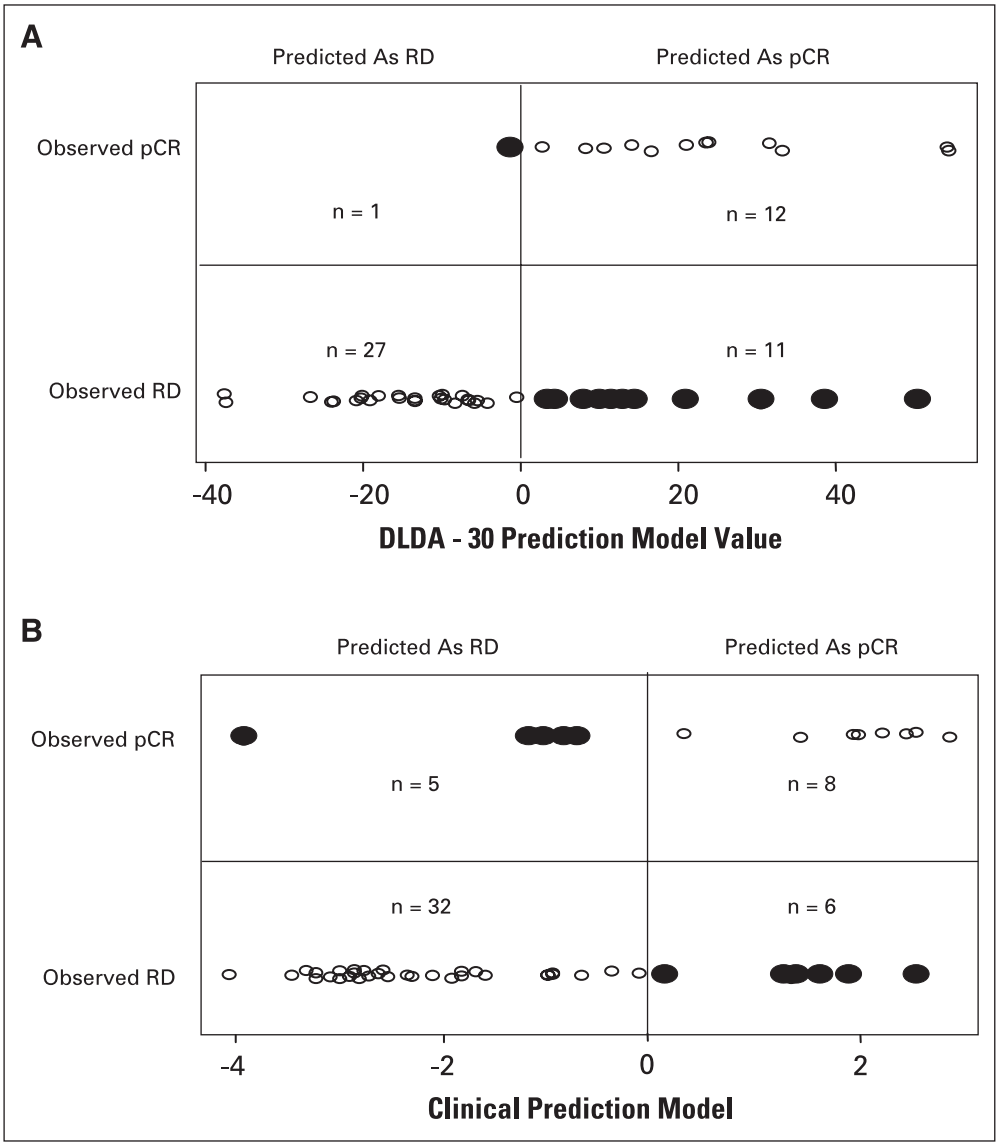

Fig 4. (A) Diagonal Linear Discriminant Analysis–30 (DLDA-30) probe set model prediction outcome (black dots indicate misclassified cases). (B) Clinical (estrogen receptor status, grade, age) model prediction (black dots indicate misclassified cases). RD, residual disease; pCR, pathologic complete response.

We also used the same 82 cases to develop a clinical variable based predictor of pCR. Younger age, ER-negative status, and high nuclear grade were associated with increased probability of pCR in univariate analysis. These three variables were combined with the same DLDA class prediction rule as used for the genomic data to create a clinical predictor. This model was trained on the 82 cases to determine the classification threshold and was applied to independent validation cases to compare its performance with that of the DLDA-30 genomic predictor.

To estimate the training set size that is necessary to develop a pharmacogenomic predictor that operates near to its plateau, we examined how the performance characteristics of DLDA-30 changed as the training set size increased. One would expect some improvement in prediction accuracy and narrowing of confidence intervals as the predictor is developed and trained on larger and larger sample sets. The steepness of this "learning curve" may help determine a reasonable sample size for predictor discovery.[21] Figure 2 shows the change in mean AAC as the classifier was developed from 20 to 100 cases in increments of 10. The results indicate a steady but modest improvement in performance as the training set size increases. Based on these observations a DLDA-30 predictor developed from 80 cases was projected to be only marginally inferior to a predictor developed from 200 cases. Essentially similar learning curves were obtained for KNN and SVM methods that all exhibited minimal improvement in projected performance beyond 60 to 100 training cases.

Performance of the DLDA-30 Pharmacogenomic Predictor in Independent Validation

We tested the prediction accuracy of the DLDA-30 predictor, the clinical variable-based predictor (DLDA including age, grade and ER-status) and a combined clinical-genomic predictor (DLDA with 33 variables including 30 probe sets + age, ER, and grade) on 51 independent individuals who were accrued consecutively after the discovery set was completed. Figure 3 shows the ROC curves for the distinct prediction models. The genomic (AUC = 0.877) and the combined clinical and genomic models (AUC = 0.879) showed

Downloaded from jco.ascopubs.org on October 20, 2008 . For personal use only. No other uses without permission.
Copyright © 2006 by the American Society of Clinical Oncology. All rights reserved.

better overall performance than the clinical-only model (AUC = 0.811). Formal statistical comparison of the three distinct ROC curves showed no significant difference between the curves. However, the models including genomic data had substantially higher sensitivity (92%; 95% CI, 0.64 to 0.99) compared with the clinical model (61%; 95% CI, 0.32 to 0.86). In the validation, the genomic predictor correctly identified 12 of 13 patients with pCR compared with eight correctly predicted by the combination of ER status, grade, and age. Also, 27 of 28 patients who were predicted to have residual disease by the pharmacogenomic test had residual cancer compared with of 32 of 37 when clinical variables were used. Table 3 and Figure 4 presents descriptive statistics for the clinical and pharmacogenomic predictors, respectively. Importantly, different cases were misclassified by the two models. This suggests that the combination of the clinical and genomic information may further improve performance. Indeed, the combined clinical plus genomic model was nominally best (AUC = 0.879). However, if it is at all possible to develop a combined model that shows unequivocal statistical superiority over a genomic or clinical predictor alone, it will require a substantially larger training set size.

Reproducibility of Pharmacogenomic Prediction in Replicate Experiments

We evaluated the technical reproducibility of the results in replicate experiments. Thirty-one of the 133 total RNA specimens were profiled twice at different points, and at two different laboratories but using the same platform and operating procedure. The replicates included 20 cases from the training set and 11 cases from the validation set. We applied the DLDA-30 predictor to both the original and replicate data and examined how often the same prediction was made. Thirty of the 31 replicate data sets yielded identical response prediction result (97% concordance). One case was a near miss (Fig 5). When the combined clinical and genomic model was used, all 31 replicates yielded the same prediction.

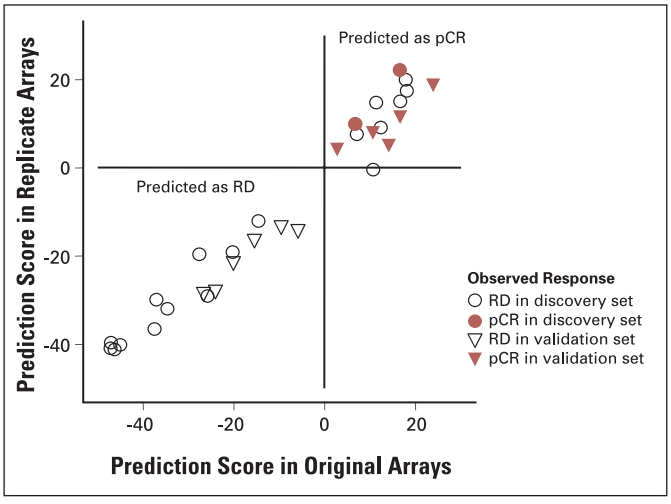

Fig 5. Prediction agreement in replicate experiments. Thirty-one RNA specimens were profiled twice to assess reproducibility of prediction results. All but one of the 31 replicates yielded identical prediction outcome for the Diagonal Linear Discriminant Analysis–30 predictor. Prediction scores < 0 predict for residual disease (RD) and > 0 predict for pathologic complete response (pCR).

DISCUSSION

We developed a multigene predictor of pCR to preoperative sequential weekly paclitaxel followed by FAC chemotherapy from fine-needle biopsies of breast cancer. pCR is a meaningful clinical end point to predict because these patients experience prolonged disease-free and overall survival compared with patients with lesser response. Good survival in these patients probably reflects benefit from chemotherapy since most clinical and gene expression variables that are associated with pCR (ie, high grade, ER-negative status, high Oncotype DX (Genomic Health Inc, Redwood City, CA) recurrence score) tend to predict worse prognosis in the absence of chemotherapy.[22] Gene expression data from 82 cases was used to identify genes associated with pCR. We constructed a large number of predictors and tested their performance in true cross validation in the training set. The classifier performances generally improved with increasing numbers of genes, but have reached a plateau at around 40 genes. This is not surprising since genes were ranked by *P* values of differential expression in univariate analysis; therefore, genes further down the list added less and less independent discriminating value. The majority of the 780 distinct predictors showed similar performance to the nominally best predictor. Other investigators have also observed this phenomenon.[23] This is due to the large number of genes that correlate at least to some extent with outcome and to the highly intercorrelated expression of individual genes. Many genes show correlation with outcome; however, the strength of correlation varies from training set to training set and therefore the rank order of genes is unstable. Since many genes are tightly coexpressed any one of the coexpressed genes could be selected for inclusion in a predictor and could yield equally good result. These observations partly explain why many of the prognostic and predictive signatures reported so far contain relatively few overlapping genes yet each gene set perform reasonably well in independent validation. Another important source of variation that contributes to the differences in predictor gene sets developed for the same purpose by different laboratories is the gene expression-profiling platform. Different platforms contain different genes and measure the expression of the same gene with different accuracy and dynamic range.[14]

We selected the nominally best predictor with the least number of genes for independent validation in 51 cases. This predictor, DLDA-30, includes 30 probe sets and uses DLDA to formulate outcome prediction rule. The predictor showed substantially higher sensitivity (92% *v* 61%) and slightly better NPV (96% *v* 86%) than a clinical variable (ER, grade, and age) –based model. The high sensitivity indicates that the predictor correctly identified almost all of the patients (92%) who actually achieved pCR. The positive predictive value (PPV) of the pharmacogenomic predictor was 52% (95% CI, 31% to 73%). Importantly, the lower bound of the 95% CI did not overlap with the 26% pCR rate observed with this regimen in unselected patients.[17] This indicates that the predictor can define a patient population who is more likely to achieve pCR than the general patient population. However, the 52% PPV also indicates that many who were predicted to have pCR had a lesser response. This type of error may be considered acceptable in the adjuvant treatment setting. The NPV of the test was also high (96%; 95% CI, 82% to 100%), which indicates that less than 5% of test-negative patients (ie, predicted to have residual disease) achieved pCR. These performance statistics are similar, with regards to the NPV, and better with regards to PPV, than those seen with ER immunohistochemistry or *HER2* gene amplification as predictive markers to endocrine or trastuzumab therapies,

Downloaded from jco.ascopubs.org on October 20, 2008 . For personal use only. No other uses without permission.
Copyright © 2006 by the American Society of Clinical Oncology. All rights reserved.

respectively. We also examined the reproducibility of the pharmacogenomic prediction results in 31 replicate experiments and observed a 97% concordance in prediction outcome for the DLDA-30 model and 100% concordance for the combined clinical and pharmacogenomic model. This indicates a very high level of technical reproducibility of pharmacogenomic predictions when the same RNA is used.

How do these results fit in with other recently reported molecular predictors of pCR? There is more than one method to predict probability of pCR. Patients with high recurrence score determined by the Oncotype Dx assay have greater probability to achieve pCR (to paclitaxel/FAC chemotherapy) than those with low recurrence score.[22] Patients with basal-like breast cancer determined by hierarchical clustering using the "intrinsic gene list" also have higher probability of pCR than other molecular classes of breast cancer.[13] Reassuringly, all of these methods tend to identify the same group of patients, those with ER-negative, high grade and highly proliferative tumors. However, they also seem to add some incremental predictive value to the known clinical characteristics. Appendix Tables 1 and 2 (online only) shows pCR rates in the training and validation sets as a function of ER-status and Appendix Figure 1 (online only) shows the correlation between Ki67 mRNA expression (probe sets 212022_s_at and 212023_s_at) and pCR. The discriminating value of these single gene variables is less than that of the combined model.

Predicting extreme chemotherapy sensitivity to paclitaxel/anthracycline therapy can be useful in some clinical situations. However, the greatest contribution will come from the molecular test that can discriminate between likelihoods of response to different chemotherapy regimens and can therefore guide selection of one treatment over another. To develop such predictor gene expression data from cohorts of patients who received different chemotherapy regimens will be needed. Pharmacogenomic research also has the potential to open new insights into the biology of breast cancer and treatment response. Indeed, some of the genes from our predictive signature seem to contribute to the mechanism of sensitivity to paclitaxel.[24]

In summary, we developed a 30-probe set DLDA classifier that predicts pathologic response to preoperative paclitaxel/FAC chemotherapy with higher sensitivity (92% *v* 61%) than a clinical variable–based predictor. In an independent validation set of 51 patients, this test correctly identified all but one of the patients who achieved pCR, and all but one of those who were predicted to have residual disease had residual cancer.

REFERENCES

1. Early Breast Cancer Trialists' Collaborative Group: Polychemotherapy for early breast cancer: An overview of the randomized trials. Lancet 352: 930-942, 1999

2. Carlson RW, Anderson BO, Bensinger W, et al: NCCN Practice guidelines for breast cancer. Oncology (Huntington) 14:33-49, 2000

3. National Institutes of Health Consensus Development Conference. http://consensus.nih.gov/2000/2000AdjuvantTherapyBreastCancer114html.htm

4. Goldhirsch A, Wood WC, Gelber RD, et al: Meeting Highlights: Updated International Expert Consensus Panel on the Primary Therapy of Early Breast Cancer. J Clin Oncol 21:3357-3366, 2003

5. Rouzier R, Pusztai L, Delaloge S, et al: Nomograms to predict pathologic complete response and metastasis-free survival after preoperative chemotherapy for breast cancer. J Clin Oncol 23:8331-8339, 2005

6. Ross JS, Linette GP, Stec J, et al: Breast cancer biomarkers and molecular medicine, Part I. Expert Rev Mol Diagn 3:573-585, 2003

7. Bast Jr RC, Ravdin P, Hayes D, et al: 2000 Update of recommendations for the use of tumor markers in breast and colorectal cancer: Clinical practice guidelines of the American Society of Clinical Oncology. J Clin Oncol 19:1865-1878, 2001

8. Fisher B, Bryant J, Wolmark N, et al: Effect of preoperative chemotherapy on the outcome of women with operable breast cancer. J Clin Oncol 16:2672-2685, 1998

9. Kuerer HM, Newman LA, Smith TL, et al: Clinical course of breast cancer patients with complete pathologic primary tumor and axillary lymph node response to doxorubicin-based neoadjuvant chemotherapy. J Clin Oncol 17:460-469, 1999

10. Ayers M, Symmans WF, Stec J, et al: Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel/FAC chemotherapy in breast cancer. J Clin Oncol 22:2284-2293, 2004

11. Pusztai L, Symmans WF, Hortobagyi GN: Development of pharmacogenomic markers to select preoperative chemotherapy for breast cancer. Breast Cancer 12:73-85, 2005

12. Symmans WF, Ayers M, Clark EA, et al: Total RNA yield and microarray gene expression profiles from fine needle aspiration and core needle biopsy samples of breast cancer. Cancer 97:2960-2971, 2003

13. Rouzier R, Perou CM, Symmans WF, et al: Different molecular subtypes of breast cancer respond differently to preoperative chemotherapy. Clin Cancer Res 11:5678-5685, 2005

14. Stec J, Wang J, Coombes K, Ayers M, et al: Comparison of the predictive accuracy of DNA array based multigene classifiers across cDNA arrays and Affymetrix GeneChips. J Mol Diagn 7:357-367, 2005

15. Pounds S, Morris SW: Estimating the occurrence of false positive and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. Bioinformatics 19:1236-1242, 2003

16. Hastie T, Tibshirani R, Friedman J: The Elements of Statistical Learning. Springer-Verlag, New York, 2001

17. Simon R: Roadmap for developing and validating therapeutically relevant genomic classifiers. J Clin Oncol 23:7332-7341, 2005

18. Hsing T, Attoor S, Dougherty E: Relation between permutation-test p-values and classifier error estimates. Machine Learning 52:11-30, 2003

19. Baker SG: The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. J Natl Cancer Inst 95:511-515, 2003

20. Green MC, Buzdar AU, Smith T, et al: Weekly paclitaxel improves pathologic complete remission in operable breast cancer when compared with paclitaxel once every 3 weeks. J Clin Oncol 23:5983-5992, 2005

21. Mukherjee S, Tamayo P, Rogers S, et al: Estimating dataset size requirements for classifying DNA microarray data. J. Comput Biol 10:119-142, 2003

22. Gianni L, Zambetti M, Clark K, et al: Gene expression profiles in paraffin-embedded core biopsy tissue predict response to chemotherapy in women with locally advanced breast cancer. J Clin Oncol 23:7265-7277, 2005

23. Ein-Dor L, Kela I, Getz G, et al: Outcome signature genes in breast cancer: Is there a unique set? Bioinformatics 21:171-178, 2005

24. Rouzier R, Rajan R, Hess KR, et al: Microtubule associated protein tau is a predictive marker and modulator of response to paclitaxel-containing preoperative chemotherapy in breast cancer. Proc Natl Acad Sci U S A 102:8315-8320, 2005

Acknowledgment

We would like to acknowledge the contributions of Stephen Tirrell, James Stec, Mark Ayers and Edwin Clark who contributed gene expression data and several useful comments to this project while employed at Millennium Pharmaceuticals Inc (Cambridge, MA).

Downloaded from jco.ascopubs.org on October 20, 2008 . For personal use only. No other uses without permission.
Copyright © 2006 by the American Society of Clinical Oncology. All rights reserved.

Appendix

The Appendix is included in the full-text version of this article, available online at www.jco.org. It is not included in the PDF version (via Adobe® Reader®).

Authors' Disclosures of Potential Conflicts of Interest

Although all authors completed the disclosure declaration, the following authors or their immediate family members indicated a financial interest. No conflict exists for drugs or devices used in a study if they are not being evaluated as part of the investigation. For a detailed description of the disclosure categories, or for more information about ASCO's conflict of interest policy, please refer to the Author Disclosure Declaration and the Disclosures of Potential Conflicts of Interest section in Information for Contributors.

Authors	Employment	Leadership	Consultant	Stock	Honoraria	Research Funds	Testimony	Other
W. Fraser Symmans		Member Scientific Advisory Board of Nuvera Biosciences Inc (A)	Affymetrix Inc (A)	Nuvera Biosciences Inc (A)				
Lajos Pusztai		Member of SAB of Nuvera Biosciences (A)	Affymetrix Inc (A)	Nuvera Biosciences Inc (A)				
Dollar Amount Codes (A) < $10,000 (B) $10,000-99,999 (C) ≥ $100,000 (N/R) Not Required								

Author Contributions

Conception and design: Kenneth R. Hess, W. Fraser Symmans, Gabriel N. Hortobagyi, Lajos Pusztai
Financial support: Gabriel N. Hortobagyi, Lajos Pusztai
Administrative support: Gabriel N. Hortobagyi, Lajos Pusztai
Provision of study materials or patients: W. Fraser Symmans, Vicente Valero, Nuhad Ibrahim, Daniel Booser, Richard L. Theriault, Aman U. Buzdar, Peter J. Dempsey, Nour Sneige, Tatiana Vidaurre, Henry L. Gomez, Gabriel N. Hortobagyi, Lajos Pusztai
Collection and assembly of data: W. Fraser Symmans, Jaime A. Mejia, Peter J. Dempsey, Roman Rouzier, Nour Sneige, Lajos Pusztai
Data analysis and interpretation: Kenneth R. Hess, Keith Anderson, W. Fraser Symmans, Roman Rouzier, Jeffrey S. Ross, Lajos Pusztai
Manuscript writing: Kenneth R. Hess, Lajos Pusztai
Final approval of manuscript: Kenneth R. Hess, Keith Anderson, W. Fraser Symmans, Vicente Valero, Nuhad Ibrahim, Jaime A. Mejia, Daniel Booser, Richard L. Theriault, Aman U. Buzdar, Peter J. Dempsey, Roman Rouzier, Nour Sneige, Jeffrey S. Ross, Tatiana Vidaurre, Henry L. Gomez, Gabriel N. Hortobagyi, Lajos Pusztai

GLOSSARY

ROC (receiver operating characteristic) curves: ROC curves plot the true positive rate (sensitivity) against the false-positive rate (1-specificity) for different cut-off levels of a test. The area under the curve is a measure of the accuracy of the test. An area of 1.0 represents a perfect test (all true positives), whereas an area of 0.5 represents a worthless test.

FDR (false discovery rate): FDR is a statistical method that is used to correct for multiple comparisons. FDR is the expected *proportion* of false positives (as opposed to the more traditional Type II error rate, which is the probability of any false positives).

Monte-Carlo cross validation: Cross validation (CV) is a process using a dataset to build a prediction algorithm and to estimate how well it will perform on new but similar data. A portion of the data is used to build an algorithm while the remainder is used to estimate the performance. Classical k-fold CV randomly divides the data into k portions using each portion in turn as a test set and the remaining data as training sets. One random partitioning of the data is used. The performance on new data is estimated as the average performance over the k test sets. In Monte-Carlo CV, the partitions are randomly selected hundreds or thousands of times.

BUM (beta-uniform mixture analysis): This model is used to quickly compute the FDR estimates for the analysis of microarray data by taking the distribution of individual p-values to follow a mixture of a Beta density function from the genes that differentially expressed and a uniform density function from the genes that are not differentially expressed.

Pharmacogenomic: The study of how a person's genome can affect their reaction to medications.

Class prediction algorithms: Computer programs used to combine known data on samples to predict unknown characteristics.

Downloaded from jco.ascopubs.org on October 20, 2008 . For personal use only. No other uses without permission.
Copyright © 2006 by the American Society of Clinical Oncology. All rights reserved.