Optimizing the Use of Gene Expression Profiling in Early-Stage Breast Cancer

Hyun-seok Kim, Christopher B. Umbricht, Peter B. Illei, Ashley Cimino-Mathews, Soonweng Cho, Nivedita Chowdhury, Maria Cristina Figueroa-Magalhaes, Catherine Pesce, Stacie C. Jeter, Charles Mylander, Martin Rosman, Lorraine Tafra, Bradley M. Turner, David G. Hicks, Tyler A. Jensen, Dylan V. Miller, Deborah K. Armstrong, Roisin M. Connolly, John H. Fetting, Robert S. Miller, Ben Ho Park, Vered Stearns, Kala Visvanathan, Antonio C. Wolff, and Leslie Cope

Author affiliations appear at the end of this article.

Published online ahead of print at www.jco.org on October 10, 2016.

Supported by research funding from Susan G. Komen Scholar Grant No. SAC110053 (A.C.W.), Susan G. Komen Foundation IIR Grant No. KG 110094 (C.B.U.), and National Cancer Institute Grant No. P30 CA006973 to the Johns Hopkins Sidney Kimmel Comprehensive Cancer Center.

H.K. and C.B.U. contributed equally to this article. A.C.W. and L.C. contributed as co-senior authors to this article.

Presented in part at the 50th Annual American Society of Clinical Oncology meeting, Chicago, IL, May 30-June 2, 2014.

Authors' disclosures of potential conflicts of interest are found in the article online at www.jco.org. Author contributions are found at the end of this article.

Corresponding author: Leslie Cope, PhD, The Johns Hopkins Sidney Kimmel Comprehensive Cancer Center, 550 N Broadway, Suite 1131, Baltimore, MD; e-mail: lcope1@jhmi.edu.

© 2016 by American Society of Clinical Oncology

0732-183X/16/3436w-4390w/\$20.00

DOI: 10.1200/JCO.2016.67.7195

ABSTRACT

Purpose

Gene expression profiling assays are frequently used to guide adjuvant chemotherapy decisions in hormone receptor–positive, lymph node–negative breast cancer. We hypothesized that the clinical value of these new tools would be more fully realized when appropriately integrated with high-quality clinicopathologic data. Hence, we developed a model that uses routine pathologic parameters to estimate Oncotype DX recurrence score (ODX RS) and independently tested its ability to predict ODX RS in clinical samples.

Patients and Methods

We retrospectively reviewed ordered ODX RS and pathology reports from five institutions ($n = 1,113$) between 2006 and 2013. We used locally performed histopathologic markers (estrogen receptor, progesterone receptor, Ki-67, human epidermal growth factor receptor 2, and Elston grade) to develop models that predict RS-based risk categories. Ordering patterns at one site were evaluated under an integrated decision-making model incorporating clinical treatment guidelines, immunohistochemistry markers, and ODX. Final locked models were independently tested ($n = 472$).

Results

Distribution of RS was similar across sites and to reported clinical practice experience and stable over time. Histopathologic markers alone determined risk category with $> 95\%$ confidence in $> 55\%$ (616 of 1,113) of cases. Application of the integrated decision model to one site indicated that the frequency of testing would not have changed overall, although ordering patterns would have changed substantially with less testing of estimated clinical risk–high or clinical risk–low cases and more testing of clinical risk–intermediate cases. In the validation set, the model correctly predicted risk category in 52.5% (248 of 472).

Conclusion

The proposed model accurately predicts high- and low-risk RS categories (> 25 or ≤ 25) in a majority of cases. Integrating histopathologic and molecular information into the decision-making process allows refocusing the use of new molecular tools to cases with uncertain risk.

J Clin Oncol 34:4390-4397. © 2016 by American Society of Clinical Oncology

INTRODUCTION

In the early 2000s, clinical trials questioned recommendations that most patients with early-stage breast cancer (ESBC) should receive adjuvant chemotherapy.^{1,2} Since then, several multigene assays and immunohistochemistry-based scores estimating recurrence risk were developed to improve clinical decision making.³⁻⁸ One assay, Oncotype DX (ODX; Genomic Health, Redwood City, CA), underwent prospective-retrospective

validation to identify patients with node-negative, estrogen receptor (ER) –positive disease most likely to benefit from adjuvant chemotherapy in ESBC.^{3,4,9,10} However, there is incomplete information on the incremental contribution of molecular assays to routinely available clinicopathologic data and how to integrate them to best inform clinical decision making.¹¹

Although prospective registry data show reduction in chemotherapy use, assays like ODX are often used to affirm pretest assessments that clinicians generate informally. Hassett et al¹² observed

that ODX testing was associated with higher odds of chemotherapy use in patients with small node-negative cancers and was associated with lower odds of chemotherapy use in node-positive or large node-negative ESBC. There has also been an increase in the proportion of recurrence score (RS) test results reported as intermediate risk, partly due to informal preselection by clinicians regarding when to request testing.¹²

Ongoing prospective trials, such as TAILORx (Trial Assigning Individualized Options for Treatment; NCT00310180) and RxPONDER (Rx for Positive Node, Endocrine Responsive Breast Cancer; NCT01272037), will help evaluate the utility of gene expression assays in node-negative and node-positive disease and identify optimal cutoffs among patients assigned to an intermediate (RS, 11 to 25) or low to intermediate ($RS \leq 25$) risk groups, respectively, regarding the role of adjuvant chemotherapy. MINDACT (Microarray in Node negative Disease may Avoid ChemoTherapy; NCT00433589) tested if ESBC with three or fewer positive lymph nodes and a high-risk clinical prognosis/low-risk molecular prognosis can be spared chemotherapy.^{12a} In the meantime, we aimed to address the simpler question of whether there are cases in which the clinician could opt to forgo a molecular assay because the outcome of the test might be predicted with a high degree of accuracy using routinely available measures.

Here, we reviewed all ODX test results over a 9-year period at several large medical institutions across the United States. Our goal was to develop and independently validate a model for predicting the ODX risk category using standard clinicopathologic parameters. Our overall hypothesis is that for many ER-positive tumors that are considered to have low- or high-risk features, the ODX risk category can be predicted with high confidence, making the molecular test redundant. This could lead to a two-step process where the prediction model is applied systematically, whereas the molecular test would be reserved for less clear-cut cases.

PATIENTS AND METHODS

Patient Selection and Pathology Variable Selection

Eligible cases for the initial model development cohort had an ODX test result ordered between June 2006 and September 2013 (closing date varied by institution) for an ER-positive, lymph node-negative, American Joint Committee on Cancer stage I or II breast cancer; were within 6 months from the initial diagnosis; and had results of locally performed ER, progesterone receptor (PR), Ki-67, human epidermal growth factor receptor 2 (HER2) assays, and Elston grade available (cases with ODX; see CONSORT diagram Appendix Fig A1, online only). Eligible cases for the independent validation cohort were sequential to the model development cohort within each institution through June 2015 and met same inclusion criteria.

Pathology reports were abstracted for clinical and pathologic information. Immunohistochemistry for ER, PR, HER2, and Ki-67 and in situ hybridization testing for HER2 were performed in the pathology laboratories of five participating institutions: Johns Hopkins Hospital (JHH), Anne Arundel Medical Center (AAMC) and Howard County General Hospital (HCGH; AAMC and HCGH are Johns Hopkins Affiliates [JHA]), the University of Rochester in New York, and Intermountain Healthcare in Utah (Data Supplement Table 1; Appendix, online only). ODX RS results were obtained from medical records at each institution and the Genomic Health ordering history of individual participating

physicians. The smaller HCGH cohort was combined with AAMC for analyses. To place ordering patterns into a larger context, we identified cases from the same period at JHH that satisfied our eligibility criteria but did not have the test ordered (cases without ODX) and used regression analysis to identify differences. The project was reviewed and approved by each institutional review board.

Model Development and Statistical Methods

Initial variable selection was carried out on the basis of univariate linear regression in JHH and JHA development samples. Variables were selected for significant association with RS ($P < .05$) in JHH samples, and the selections were validated in JHA samples using the same criteria. For this study, we adopted an RS cutoff of 25 as the criterion for high risk on the basis of the RS threshold for chemotherapy administration used in ongoing trials like TAILORx and RxPONDER after reanalysis of NSABP-B20.¹³

Both linear and random forest models were considered for modeling the relationship between the selected clinicopathologic variables and recurrence score.¹⁴⁻¹⁶ For the linear model, we fitted a multivariate model for ODX RS by minimizing the mean squared error in the usual manner but took a novel approach to reporting the result and characterizing uncertainty. Our primary goal in modeling the ODX RS was to identify cases for which the risk category could be predicted with a high degree of confidence. We hypothesized that full ODX testing could be preferentially used in cases where prediction was less reliable. Accordingly, to predict risk for a test case, we applied the model to predict RS and then selected as references all cases from our training database having a predicted RS close to that of the test case, reporting the proportion of reference cases that fell into each risk category. A case was predicted to be low risk if $\geq 95\%$ of reference cases have $RS \leq 25$ and predicted to be high risk if $\geq 95\%$ of references have $RS > 25$. Cases that could not be classified with 95% confidence were considered undetermined. Performance was calculated as the proportion of cases that could be classified as high or low risk. The level of confidence itself was a model parameter, nominally set to 0.95 for the analyses reported. The achieved level of confidence for each model (coverage probability) was calculated as the proportion of correct calls and was also reported (Appendix).

A similar approach was used for the random forest model. After model development, a test case was classified by applying the model, but reference cases were identified from the training database using a slightly modified approach. Whereas in the linear model, reference cases were matched to the test case according to predicted RS, here they were selected using the proximity measure native to the random forest, which measures sample similarity as the proportion of trees in which two samples end up in the same terminal leaf (Appendix).

Cross-validation was used to estimate the performance on the developmental data sets and select the best model. Cross-validation models were trained on samples from three centers and tested on the fourth, to reflect the situation we expect in practice, where our model may be applied to data that do not precisely match any of our training cohorts. Because the JHH and JHA training samples were used to select variables and tune model parameters, cross-validation results for those cohorts are subject to bias and overfitting. Both models were locked before testing on the independent data cases from the participating centers that were sequentially collected after the training data set.

Models were explored with scatterplots describing the relationship between input variables, observed ODX RS, and model predictions. The random forest model was also characterized using agglomerative hierarchical clustering via Ward's rule,¹⁷ with the proximity sample similarity measure defined by the random forest. Analysis of variance was used to compare the distributions of continuous variables across participating centers (ie, percent ER, PR, Ki-67, and ODX RS). Categorical variables (ie, HER2 status and Elston grade) were compared by χ^2 test. All data analyses were performed using R version 3.0.2.

Developing and Evaluating an Integrated Model

We also evaluated how ordering patterns of the ODX test at one site (JHH) would have changed if our prediction model were applied alongside consensus and evidence-based recommendations made by the National Comprehensive Cancer Network breast cancer clinical practice guidelines.¹² Patients at JHH matching our criteria were stratified into the three preliminary clinical risk groups (low, intermediate, and high). We then applied our prediction model to clinical intermediate risk group patients.¹²

RESULTS

Patients

The development cohort consisted of 1,113 cases, identified in the four data sets: JHH (n = 299), JHA (n = 308, 238 cases from AAMC and 70 cases from HCGH), Intermountain Healthcare in Utah (n = 301), and University of Rochester in New York (n = 205). Although the distributions of parameters such as age, tumor size, ER, PR, Ki-67, and grade varied across data sets (Data Supplement Table 2), differences in the ODX RS distributions across the data sets were not statistically significant, and overall distribution was similar to previous reports.^{12,18} The age and tumor characteristics of the JHH ODX patients changed little over the 6-year period (Data Supplement Figure 1).

Three of the same institutions contributed 472 subsequent cases to the independent validation cohort. Patient characteristics and RS distribution were mostly similar in the development and validation cohorts (Table 1).

Table 1. Clinical and Pathologic Characteristics

Characteristic	Development (n = 1,113)	Validation (n = 472)	P
Oncotype DX RS			
Low risk (RS < 18)	602 (54.1)	260 (55.0)	.94
Intermediate risk (RS, 18-30)	392 (35.2)	163 (34.5)	
High risk (RS ≥ 31)	119 (10.7)	49 (10.3)	
Age at diagnosis, years			.20
< 50	305 (27.4)	120 (25.5)	
50-59	683 (61.4)	317 (67.3)	
> 70	125 (11.2)	34 (7.2)	
Tumor size			< .001
T1a/T1b	282 (25.3)	81 (17.2)	
T1c	583 (52.4)	221 (47.0)	
T2	227 (20.4)	157 (33.4)	
T3	21 (1.9)	11 (2.3)	
ER, median (IQR)	90 (80-95)	95 (90-100)	< .001
PR, median (IQR)	79 (28-90)	75 (12.5-90)	.16
Ki-67, median (IQR)	17 (10-30)	16.5 (10-30)	.68
Elston grade			.23
Low	331 (29.7)	125 (26.7)	
Intermediate	606 (54.4)	269 (57.0)	
High	176 (15.8)	77 (16.3)	
HER2 status			.015
Negative	1,083 (97.3)	469 (99.4)	
Positive	30 (2.7)	3 (0.6)	

NOTE. Data presented as No. (%) unless otherwise noted. All patients were Stage I or II, node-negative. Comparison of key clinical and pathologic characteristics between development and external validation data sets. P values were calculated by t test for continuous variables and by χ^2 for factors.

Abbreviations: ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; IQR, interquartile range; PR, progesterone receptor; RS, recurrence score.

Variable Selection and Structure of Our Prediction Models

On univariate linear regression analyses, clinical parameters such as age and tumor size were not significantly correlated with ODX RS. Grade, HER2 status, and Ki-67 were all positively correlated, whereas ER and PR were inversely correlated (Data Supplement Table 3). Therefore, we included ER, PR, Ki-67, HER2, and grade in the final linear and random forest models.

The coefficients for each variable in the multivariate regression model are shown in Data Supplement Figure 2. PR, Ki-67, and grade were the most significant variables, although both ER and HER2 also offered independent value. The relationship between the predicted and observed ODX RS and relationships between individual predictors and predicted ODX are shown in Data Supplement Figure 3.

Although random forest models cannot be expressed in terms of simple coefficients, scatterplots and boxplots can help visualize the relationship between predicted and observed ODX RS (Fig 1A) and predictive variables (Data Supplement Figure 4). As expected, ER and PR are inversely correlated to predicted risk, whereas Ki67, HER2, and grade are directly correlated. Unsupervised cluster analysis was performed based on sample proximity measures from the random forest model to further visualize the structure of the model (Fig 1B), and we observed several distinct clusters on the basis of varying ER, PR, Ki-67, HER2, and grade, including one (highlighted in the colored square on right, Fig 1B) showing low ODX RS samples with moderate to strong ER and PR, low Ki-67, grade 1, and negative HER2.

Model Performance in Internal and External Validation

Our goal was to identify cases for which the ODX risk category could be predicted with high confidence using routine clinical and pathologic measures. This would allow additional molecular testing to be directed at the remaining cases, where the gene expression signature might yield new information. Accordingly, performance was measured as the proportion of cases that could be confidently called high or low risk before ODX. The target confidence level was set at 95%, and the observed coverage probability (ie, accuracy) was also reported.

Model performance is summarized in Table 2 and Fig 2. Internal validation was performed by cross-validation (Data Supplement Table 4), developing a model using data from three sites and testing it on the remaining site, for four iterations. The final, locked linear and random forest models, trained on the entire development cohort (n = 1,113), were independently tested on a subsequent/sequential data set (n = 472).

The random forest model outperformed the linear model in each iteration of the cross-validation, leading to our decision to recommend the random forest model for general use. In the validation set, 48.3% (228 of 472) of cases were predicted to be low risk using the random forest model, with an observed coverage probability of 96.9%. Likewise, 4.2% (20 of 472) of cases were predicted to be high risk (coverage probability, 96.5%). The linear model performed nearly as well on low-risk cases, confidently identifying 226 of 472 (47.9%) of all cases as low risk using the linear model (coverage probability, 98.7%), but identified fewer cases (14 of 472 or 3%) as high risk (coverage probability, 100%).

Fig 1. Structure of the random forest model. (A) Scatterplot comparing the predicted Oncotype DX recurrence score (y-axis) to the measured recurrence score (x-axis) in the model development (top) and validation (bottom) cohorts. The rank-based Spearman correlation coefficient is provided (Cor). (B) Hierarchical cluster analysis of database samples on the basis of the proximity distance measure from the random forest model. Input data are represented graphically to the left of the dendrogram, with darker colors representing higher values. Samples with Oncotype DX recurrence score (RS) > 25 are indicated in black, in contrast to those with RS ≤ 25, shown in white. Two representative clusters are highlighted. ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; PR, progesterone receptor.

Risk categories for cases with observed RS ≤ 25 or ≥ 31 were easier to predict than those with observed RS 26 to 30 (Data Supplement Table 5).

Input Values, Reference Sets, and Risk Predictions in the Development Set

Our training database of 1,113 cases comprised 480 unique combinations of input variables. The typical input case was matched to a median of 51 reference cases; only 3% were matched to < 10 references.

Extensive simulations in our model revealed that ER and PR were generally negatively associated and HER2 and Ki67 were generally positively associated with risk. In < 4% of cases, pathologic assessment of risk disagreed with quantitative model predictions, leading to a change in the predicted risk category in < 1% of cases.

Online Implementation of the Predictive Model

We organized our locked random forest model in a simple and intuitive online application^{17a} (Fig 3). However, before users apply

this model on individual cases, we strongly recommend that they evaluate the model on their own data using the “Test your data” tab on the site.

The interface allows the user to input values for ER, PR, Ki-67, grade, and HER2 status (Fig 3A). The standard output (Fig 3B) includes a 90% prediction interval for the RS, an estimate of the probability that the RS exceeds the high-risk threshold of 25, and a figure showing the expected distribution of recurrence scores for the case. Under the “Reference Case Set” tab, vital statistics are reported for the reference samples underlying the inference, including a graphical summary of the characteristics of the input variables (Fig 3C).

Evaluating an Integrated Model

As a final step, we evaluated how ODX ordering patterns might change if we integrated our predictions as part of the clinical decision-making process at one of our sites. On the basis of Hassett et al,¹² we modeled a decision-making process where patients were first stratified into three risk categories by stage and other clinical characteristics (Table 3). We assumed that immediate treatment

Table 2. Performance of Development and Validation Models

Model	Risk Category Classification Rates				
	Cross-Validation (n = 1,113)				Validation (n = 472)
	JHU (n = 299)	JHA (n = 308)	URNY (n = 205)	IHCU (n = 301)	
Random forest model, No. (%)					
Undetermined	130 (43)	105 (34)	76 (37)	173 (57)	224 (47)
Classified as high or low risk	169 (57)	203 (66)	129 (63)	128 (43)	248 (53)
95% CI, %	51 to 62	60 to 71	56 to 70	37 to 48	48 to 57
Coverage probability, %	95	97	98	96	97
Linear model, No. (%)					
Undetermined	145 (48)	126 (41)	94 (46)	217 (70)	236 (50)
Classified as high or low risk	154 (52)	182 (59)	111 (54)	84 (30)	236 (50)
95% CI, %	46 to 57	53 to 65	47 to 61	23 to 33	45 to 55
Coverage probability, %	97	98	98	98	99

NOTE. A case is predicted low if at least 95% of matched reference set cases had RS $\leq$ 25 and predicted high if at least 95% had RS $>$ 25. Cases that cannot be classified at this level of confidence are prediction undetermined. Coverage probability is presented as the percentage of correct predictions.

Abbreviations: IHCU, Intermountain Healthcare in Utah; JHA, Johns Hopkins Affiliates; JHU, Johns Hopkins University; URNY, RS, recurrence score; University of Rochester in New York.

decisions without ODX testing would be made for the clinical risk–low and clinical risk–high groups, whereas the clinical risk–intermediate group would be considered for ODX testing and could therefore be evaluated using our prediction algorithm. In all, 613 (65%) of 939 JHH patients matching study criteria (299 cases with ODX and 640 cases without ODX; Data Supplement Table 6) were defined as clinical intermediate risk group. Of these 613, 297

(48%) were then estimated to be in the undetermined risk using our predictive model and would therefore potentially be referred for ODX testing (Table 3).

Of interest, although these 297 JHH cases are similar in number to the 299 in the JHH ODX cohort, only 127 of the 299 JHH cases with ODX results would have been recommended for testing if our algorithm had been applied to them first. As such, our

Fig 2. Classification accuracy by predicted risk category. A case is predicted low if at least 95% of matched reference set cases had recurrence score (RS) $\leq$ 25 and predicted high if at least 95% had RS $>$ 25. Cases that cannot be classified at this level of confidence are prediction undetermined. Observed coverage probabilities, calculated as the percentage of correct predictions, meet or exceed the target confidence level of 95%. ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; PR, progesterone receptor.

Fig 3. Screenshot of the online recurrence estimator tool. (A) Input variables. The graphic interface uses sliders and radio buttons for easy input. (B) Box plot of recurrence score of reference cases. Standard output includes both box plot and density plots of reference cases selected to match the clinicopathologic characteristics of a case. Expected risk is summarized above the plot. (C) Ladder plot of characteristics of reference cases. Standard output includes a ladder plot showing the clinicopathologic characteristics of the reference sample set selected to match a case. The input parameters for the case are shown as a black dotted line. Grades low, intermediate, and high are plotted as 0, 50, and 100, respectively. Human epidermal growth factor receptor 2 (HER2) negative is plotted as 0, and positive is plotted as 100. ER, estrogen receptor; PR, progesterone receptor; RS, recurrence score. The application is available at <http://www.breastrecurrenceestimator.unc.jhmi.edu>. Before applying this model on individual cases, users should evaluate the model on their own data using the "Test your data" tab on the site.

algorithm seems to identify a group of patients for whom ODX ordering is more likely to contribute information not already available using routine measures.

DISCUSSION

Several commercially available molecular tests have demonstrable analytical and clinical validity in ER-positive ESBC,¹⁹ but clinicians remain uncertain about how to best use them. An indiscriminate approach of ordering them for all cases fails to integrate established routine pathology measures with known clinical utility.²⁰⁻²² Recent studies have used pathology measures to predict results from molecular assays,²³⁻³¹ some with specific recommendations regarding when a molecular assay is likely to independently contribute information,²⁹⁻³¹ but how to integrate this into clinical practice remains unclear.

Here, we describe an approach to identify cases where ordering ODX is likely to elicit information that is complementary to already

available standard clinicopathological predictors. Our model uses five common histopathologic variables: ER, PR, Ki-67, HER2, and Elston grade. Testing the locked model in an independent cohort confirmed its performance. It confidently allocated more than half (52.5%) of all cases in our validation data set to a high- or low-risk category (96.8% overall coverage probability), thereby potentially limiting ODX ordering to the remaining cases where the model was uninformative and the assay might elicit new information.

Our study has several strengths. We developed and independently tested a model using a diverse sample pool, representing large hospitals around the United States with pathology evaluation performed locally. Our approach to reporting risk predictions with reliable error estimation on the basis of a set of reference samples gives the clinician a uniquely intuitive summary of risk, and our model is available online for clinicians to use after their own validation. We developed it using both random forest and linear regression, but pragmatically prefer the random forest because it is more flexible at integrating diverse data sets and slightly outperforms the linear model.

Table 3. Evaluation of Model Integrating Predicted RS and NCCN categories

Clinical Risk	Cases With ODX	Cases Without ODX	Total
Low	50 (16.7)	219 (34.2)	269 (28.6)
Intermediate	245 (81.9)	368 (57.5)	613 (65.3)
Predicted low	108 (44.1)	173 (47.0)	281 (44.5)
Prediction undetermined	127 (51.8)	170 (46.2)	297 (47.1)
Predicted high	10 (4.1)	25 (6.8)	35 (5.5)
High	4 (1.3)	53 (8.3)	57 (6.1)
Total	299	640	939

NOTE. Data presented as No. (%) unless otherwise noted. Cases without ODX were collected over the same time period and otherwise met inclusion criteria. In this analysis, clinical intermediate risk cases were further stratified according to predicted risk category, and proportions shown for predicted risk are relative to the clinical intermediate category. The three clinical risk categories were previously defined by Hassett et al¹² following consensus and evidence-based recommendations made by the National Comprehensive Cancer Network breast cancer clinical practice guidelines. Clinical risk low: chemotherapy not recommended for invasive ductal/lobular cancers if stage T1aNO or stage T1bNO without unfavorable features (high nuclear or histologic grade, angiolymphatic invasion, or HER2 overexpression) or for tubular/colloid cancers in which the tumor is < 3 cm and nodes are not involved. Clinical risk intermediate: consider testing and chemotherapy for invasive ductal/lobular cancers if stage T1bNO with unfavorable features or if the tumor is > 1 cm, nodes not involved, and HER2 not overexpressed. Clinical risk high: recommend chemotherapy for tubular/lobular cancers if tumor ≥ 3 cm and for invasive ductal/lobular cancers if tumor > 1 cm, nodes not involved, and HER2 positive, or any node-positive cancers.

Abbreviations: HER2, human epidermal growth factor receptor 2; NCCN, National Comprehensive Cancer Network; ODX, Oncotype DX; RS, recurrence score.

Our study also has limitations. Current use of ODX in clinical practice reflects a personalized clinical decision-making process, and our exercise applies only to clinical situations where the clinician intended to and ordered the test. The selected RS threshold of 25 is commonly used to guide the off-protocol use of chemotherapy, but, should ongoing studies identify a more appropriate categorization of risk, the model could be retrained and revalidated accordingly. Measures like ER, PR, Ki-67, and grade were done locally without central standardization, but this real-life scenario in fact highlights the potential robustness of our model in view of the expected variability of these measures in clinical practice.³² Finally, results for uncommon combinations of input parameters were based on small numbers of reference cases and may rarely (< 1% of cases) produce counterintuitive estimates.

However, such cases can be easily recognized and interpreted, because our predictive model output includes a thorough summary of the strength of evidence.

In conclusion, in routine clinical practice scenarios where clinicians ordered ODX, our independently validated prediction model correctly classified the observed RS category (≤ 25 or > 25) with $\geq 95\%$ confidence in more than half of ESBC cases. Our approach could meaningfully change ordering patterns and improve the likelihood that molecular testing complements (not just replicates) standard clinicopathological information.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Disclosures provided by the authors are available with this article at www.jco.org.

AUTHOR CONTRIBUTIONS

Conception and design: Hyun-seok Kim, Christopher B. Umbricht, Antonio C. Wolff, Leslie Cope

Financial support: Christopher B. Umbricht, Antonio C. Wolff

Administrative support: Christopher B. Umbricht, Antonio C. Wolff, Leslie Cope

Provision of study materials or patients: Christopher B. Umbricht, Peter B. Illei, Stacie C. Jeter, Charles Mylander, Martin Rosman, Bradley M. Turner, David G. Hicks, Tyler A. Jensen, Dylan V. Miller, Deborah K. Armstrong, Roisin M. Connolly, John H. Fetting, Robert S. Miller, Ben Ho Park, Vered Stearns, Antonio C. Wolff

Collection and assembly of data: Hyun-seok Kim, Christopher B. Umbricht, Peter B. Illei, Ashley Cimino-Mathews, Nivedita Chowdhury, Maria Cristina Figueroa-Magalhaes, Catherine Pesce, Stacie C. Jeter, Charles Mylander, Martin Rosman, Lorraine Tafra, Bradley M. Turner, David G. Hicks, Tyler A. Jensen, Dylan V. Miller, Antonio C. Wolff, Leslie Cope

Data analysis and interpretation: Hyun-seok Kim, Christopher B. Umbricht, Soonweng Cho, Charles Mylander, Kala Visvanathan, Antonio C. Wolff, Leslie Cope

Manuscript writing: All authors

Final approval of manuscript: All authors

Accountable for all aspects of the work: All authors

REFERENCES

- International Breast Cancer Study Group (IBCSG): Endocrine responsiveness and tailoring adjuvant therapy for postmenopausal lymph node-negative breast cancer: A randomized trial. *J Natl Cancer Inst* 94:1054-1065, 2002
- Wolff AC, Abelson MD: Adjuvant chemotherapy for postmenopausal lymph node-negative breast cancer: It ain't necessarily so. *J Natl Cancer Inst* 94:1041-1043, 2002
- 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* 351:2817-2826, 2004
- Paik S, Tang G, Shak S, et al: Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *J Clin Oncol* 24:3726-3734, 2006
- Buyse M, Loi S, van't Veer L, et al: Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst* 98:1183-1192, 2006
- Parker JS, Mullins M, Cheang MC, et al: Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* 27:1160-1167, 2009
- Dowsett M, Sestak I, Lopez-Knowles E, et al: Comparison of PAM50 risk of recurrence score with oncoprint DX and IHC4 for predicting risk of distant recurrence after endocrine therapy. *J Clin Oncol* 31:2783-2790, 2013
- Sgroi DC, Sestak I, Cuzick J, et al: Prediction of late distant recurrence in patients with oestrogen-receptor-positive breast cancer: A prospective comparison of the breast-cancer index (BCI) assay, 21-gene recurrence score, and IHC4 in the Trans-ATAC study population. *Lancet Oncol* 14:1067-1076, 2013
- Tang G, Shak S, Paik S, et al: Comparison of the prognostic and predictive utilities of the 21-gene Recurrence Score assay and Adjuvant! for women with node-negative, ER-positive breast cancer: Results from NSABP B-14 and NSABP B-20. *Breast Cancer Res Treat* 127:133-142, 2011
- Genomic Health: Invasive node-negative breast cancer. <http://breast-cancer.oncotypedx.com/en-US/Managed-Care/Node-Negative.aspx>
- Marchionni L, Wilson LJ, Wolff AC, et al: Systematic review: Gene expression profiling assays in early-stage breast cancer. *Ann Intern Med* 148:358-369, 2008
- Hassett MJ, Silver SM, Hughes ME, et al: Adoption of gene expression profile testing and association with use of chemotherapy among women with breast cancer. *J Clin Oncol* 30:2218-2226, 2012
- Cardoso F, van't Veer LJ, Bogaerts J, et al: 70-gene signature as an aid to treatment decisions in

early-stage breast cancer. *N Engl J Med* 375: 717-729, 2016

13. Sparano JA, Paik S: Development of the 21-gene assay and its application in clinical practice and clinical trials. *J Clin Oncol* 26:721-728, 2008

14. Breiman L: Random forests. *Mach Learn* 45: 5-32, 2001

15. Liaw A, Wiener M: Classification and regression by randomForest. *R News* 2:18-22, 2002

16. Shi T, Horvath S: Unsupervised learning with random forest predictors. *J Comput Graph Stat* 15: 118-138, 2006

17. Maechler M, Rousseeuw P, Struyf A: Cluster: Cluster analysis basics and extensions. R package version 1.15.3, 2014

17a. Johns Hopkins University: Breast Cancer Recurrence Score Estimator. <http://www.breastrecurrenceestimator.onc.jhmi.edu>

18. Carlson JJ, Roth JA: The impact of the Oncotype Dx breast cancer assay in clinical practice: A systematic review and meta-analysis. *Breast Cancer Res Treat* 141:13-22, 2013 [Erratum: *Breast Cancer Res Treat* 146:233, 2014]

19. Harris LN, Ismaila N, McShane LM, et al: Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: American Society of Clinical Oncology clinical practice guideline. *J Clin Oncol* 34:1134-1150, 2016

20. Hammond ME, Hayes DF, Dowsett M, et al: American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *J Clin Oncol* 28:2784-2795, 2010

21. Wolff AC, Hammond ME, Hicks DG, et al: Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. *J Clin Oncol* 31:3997-4013, 2013

22. Wolff AC, Hammond ME, Hicks DG, et al: Reply to E.A. Rakha et al. *J Clin Oncol* 33:1302-1304, 2015

23. Ingoldsby H, Webber M, Wall D, et al: Prediction of Oncotype DX and TAILORx risk categories using histopathological and immunohistochemical markers by classification and regression tree (CART) analysis. *Breast* 22:879-886, 2013

24. Allison KH, Kandalafi PL, Sittani CM, et al: Routine pathologic parameters can predict Oncotype DX recurrence scores in subsets of ER positive patients: Who does not always need testing? *Breast Cancer Res Treat* 131:413-424, 2012

25. Mattes MD, Mann JM, Ashamalla H, et al: Routine histopathologic characteristics can predict oncotype DX recurrence score in subsets of breast cancer patients. *Cancer Invest* 31:604-606, 2013

26. Tang P, Wang J, Hicks DG, et al: A lower Allred score for progesterone receptor is strongly associated with a higher recurrence score of 21-gene assay in breast cancer. *Cancer Invest* 28: 978-982, 2010

27. Geradts J, Bean SM, Bentley RC, et al: The oncotype DX recurrence score is correlated with a composite index including routinely reported pathobiologic features. *Cancer Invest* 28:969-977, 2010

28. Flanagan MB, Dabbs DJ, Brufsky AM, et al: Histopathologic variables predict Oncotype DX recurrence score. *Mod Pathol* 21:1255-1261, 2008

29. Klein ME, Dabbs DJ, Shuai Y, et al: Prediction of the Oncotype DX recurrence score: Use of pathology-generated equations derived by linear regression analysis. *Mod Pathol* 26:658-664, 2013

30. Gage MM, Rosman M, Mylander WC, et al: A Validated model for identifying patients unlikely to benefit from the 21-gene recurrence score assay. *Clin Breast Cancer* 15:467-472, 2015

31. Turner BM, Skinner KA, Tang P, et al: Use of modified Magee equations and histologic criteria to predict the Oncotype DX recurrence score. *Mod Pathol* 28:921-931, 2015

32. Polley MY, Leung SC, Gao D, et al: An international study to increase concordance in Ki67 scoring. *Mod Pathol* 28:778-786, 2015

Affiliations

Hyun-seok Kim, Christopher B. Umbricht, Peter B. Illei, Ashley Cimino-Mathews, Soonweng Cho, Nivedita Chowdhury, Maria Cristina Figueroa-Magalhaes, Catherine Pesce, Stacie C. Jeter, Deborah K. Armstrong, Roisin M. Connolly, John H. Fetting, Ben Ho Park, Vered Stearns, Antonio C. Wolff, and Leslie Cope, The Johns Hopkins University School of Medicine; Kala Visvanathan, The Johns Hopkins University Bloomberg School of Public Health, Baltimore; Charles Mylander, Martin Rosman, and Lorraine Tafra, Anne Arundel Medical Center, Annapolis, MD; Bradley M. Turner, David G. Hicks, University of Rochester Medical Center, Rochester, NY; Robert S. Miller, American Society of Clinical Oncology, Alexandria, VA; and Tyler A. Jensen and Dylan V. Miller, Intermountain Healthcare, Salt Lake City, UT.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Optimizing the Use of Gene Expression Profiling in Early-Stage Breast Cancer

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or jco.ascopubs.org/site/ifc.

Hyun-seok Kim

No relationship to disclose

Christopher B. Umbricht

No relationship to disclose

Peter B. Illei

Consulting or Advisory Role: Genentech, Roche, Myriad Genetics, Celgene

Research Funding: Celgene

Ashley Cimino-Mathews

Consulting or Advisory Role: Genentech

Research Funding: Genentech

Soonweng Cho

Employment: Pfizer

Research Funding: Pfizer

Nivedita Chowdhury

No relationship to disclose

Maria Cristina Figueroa-Magalhaes

No relationship to disclose

Catherine Pesce

No relationship to disclose

Stacie C. Jeter

No relationship to disclose

Charles Mylander

No relationship to disclose

Martin Rosman

No relationship to disclose

Lorraine Tafra

Employment: EndoLogic (I)

Leadership: EndoLogic (I)

Stock or Other Ownership: EndoLogic

Consulting or Advisory Role: Medtronic, LexciiMedical

Research Funding: Genentech

Patents, Royalties, Other Intellectual Property: EndoLogic (I)

Travel, Accommodations, Expenses: EndoLogic (I)

Bradley M. Turner

No relationship to disclose

David G. Hicks

Honoraria: Genentech

Speakers' Bureau: Genentech

Travel, Accommodations, Expenses: Genentech

Tyler A. Jensen

No relationship to disclose

Dylan V. Miller

No relationship to disclose

Deborah K. Armstrong

Consulting or Advisory Role: Eisai (I), Morphotek

Research Funding: Exelixis (I), Clovis Oncology, AstraZeneca, Genentech, Astex Pharmaceuticals, Incyte, Eisai (I)

Roisin M. Connolly

Research Funding: Genentech, Novartis, Puma Biotechnology

John H. Fetting

No relationship to disclose

Robert S. Miller

No relationship to disclose

Ben Ho Park

Stock or Other Ownership: Loxo Oncology

Consulting or Advisory Role: Horizon Discovery

Research Funding: Genomic Health, Foundation Medicine

Vered Stearns

Research Funding: AbbVie, Merck, Pfizer, MedImmune, Novartis, Celgene, Puma Biotechnology

Kala Visvanathan

Patents, Royalties, Other Intellectual Property: Kala Visvanathan has been named as inventor on one or more issued patents or pending patent applications relating to methylation in breast cancer in the past 2 years, has assigned her rights to Johns Hopkins University, and participates in a royalty sharing agreement with Johns Hopkins University.

Antonio C. Wolff

Research Funding: Myriad Genetics, Pfizer

Patents, Royalties, Other Intellectual Property: Antonio Wolff has been named as inventor on one or more issued patents or pending patent applications relating to methylation in breast cancer in the past 2 years, has assigned his rights to Johns Hopkins University, and participates in a royalty sharing agreement with Johns Hopkins University.

Leslie Cope

Patents, Royalties, Other Intellectual Property: Leslie Cope has been named as inventor on one or more issued patents or pending patent applications relating to methylation in colon cancer in the past 2 years, has assigned his rights to Johns Hopkins University, and participates in a royalty sharing agreement with Johns Hopkins University.

Acknowledgment

We thank Elizabeth Garrett-Mayer, PhD, and Jeff Leek, PhD, for their careful manuscript reviews. We also thank Nyrrie Soukiazian for data annotation at the University of Rochester.

Appendix

Fig A1. Patient distribution flow diagram. AJCC, American Joint Committee on Cancer; ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; PR, progesterone receptor.

Supplemental Methods

Predicting recurrence scores. As shown in the Data Supplement, our estimate of the Oncotype DX recurrence score (RS) for a given test case is calculated from the observed RSs of a set of samples selected from our training database to have a risk profile similar to that of the test case. Rather than providing a direct estimate of the RS, the purpose of the (linear or random forest) model is to capture the relationship between the prediction variables and the RS so that sets of samples with similar levels of risk, as defined by the Oncotype DX RS, can be identified. In basing the inference on a well-defined set of cases and describing those cases thoroughly, we believe that our approach offers the user a uniquely transparent and highly interpretable model.

Linear model: Identifying risk-equivalent reference cases. Calculations were performed using the generalized linear model function in R. (<https://www.R-project.org/>) Initial variable selection was carried out using univariate linear regression models in Johns Hopkins Hospital (JHH) training samples, and the results were checked on Johns Hopkins Affiliations (JHA) training samples before testing on University of Rochester in New York (URNY) and Intermountain Healthcare in Utah (IHCU) training samples as well as the separate test cohort. The selected variables: estrogen receptor, progesterone receptor, Ki-67, human epidermal growth factor receptor 2 (HER2), and grade were then used to fit a multivariate linear model.

After fitting the multivariate linear model to the entire training set and estimating the RS for each sample, the next step was to match each case to a set of risk-equivalent reference cases. For this, we took the simple approach of using all cases for which the estimated RS was close to that of the current test case. To select the precise distance to be used to identify similar cases, we defined the neighborhood for selecting reference samples as the minimum distance for which the resulting reference sets achieved at targeted 95% confidence level. The optimization was carried out over a grid of values ranging 0.5 to 10 by increments of 0.25, and the value 1.25 was selected as optimal. Following the procedure used for initial variable selection, the parameter estimation was carried out using cross-validation (Hastie et al: New York, NY, Springer Verlag, 2009) on JHH training samples to select the optimal value and verifying that the selected value preserved the targeted level of 95% confidence by testing in JHA training samples. This approach ensured that URNY and IHCU samples remained available for true, independent cross-validation.

Random forest model: Identifying risk-equivalent reference cases. Calculations were performed using the Random Forest package in R software, arbitrarily implementing 50,000 trees with a maximum of 50 nodes per tree, using the five significant variables from the univariate analysis. To minimize computation time in the online implementation, the reference set was precalculated for each set of possible input parameters.

For the random forest model (Breiman: Machine Learning 45:5-32, 2001), the similarity between samples was calculated using proximity (Lin et al: J Am Stat Assoc 101:578-590, 2006), a measure of similarity native to the random forest. The proximity between any two cases is the proportion of trees in which the two are classified together in the same terminal leaf. Two cases with identical pathology values would traverse each tree together and end up with a proximity of 1, whereas a pair of slightly different cases might have a proximity of 0.5, meaning that they follow identical paths through half of the trees and diverge in the other half.

To select the precise distance to be used to identify similar cases, we defined the neighborhood for selecting reference samples as the minimum distance for which the resulting reference sets achieved at targeted 95% confidence level. The optimization was carried out over a grid of values ranging from 0 to 1 by increments of 0.025, and the value 0.225 was selected as optimal. Following the procedure used for initial variable selection, the parameter estimation was carried out using cross-validation on JHH training samples to select the optimal value and verifying that the selected value preserved the targeted level of 95% confidence by testing in JHA training samples. This approach ensured that URNY and IHCU samples remained available for true, independent cross-validation.

Immunohistochemical Analysis

All pathology departments involved in this study followed identical procedures for reporting on immunohistochemical (IHC) measures. Cancer cell estrogen receptor and progesterone receptor labeling are manually scored and reported as percentage (0 to 100%) and intensity (weak, moderate, and strong) of nuclear labeling, such that any nuclear labeling > 1% is considered positive (Hammond et al: J Clin Oncol 28:2784-2795, 2010). Cancer cell Ki-67 labeling by IHC is manually scored and reported as percentage (0 to 100%) nuclear labeling (Dowsett et al: J Natl Cancer Inst 103:1656-1664, 2011). Testing for HER2 follows the algorithm of first performing HER2 IHC, followed by reflex fluorescent in situ hybridization testing of all IHC equivocal (2+) cases. The carcinoma cell HER2 labeling by IHC is manually scored and reported as 0, 1+, 2+, or 3+ as per the contemporary American Society of Clinical Oncology/College of American Pathologists guidelines at the time of the case (Wolff et al: J Clin Oncol 25:118-145, 2007; Wolff et al: J Clin Oncol 31:3997-4013, 2013). A carcinoma with an HER2 IHC score of 3+ or HER2:chromosome 17 fluorescent in situ hybridization ratio of > 2.2 is considered positive.