JOURNAL OF CLINICAL ONCOLOGY ORIGINAL REPORT

Prognostic Value of a Combined Estrogen Receptor, Progesterone Receptor, Ki-67, and Human Epidermal Growth Factor Receptor 2 Immunohistochemical Score and Comparison With the Genomic Health Recurrence Score in Early Breast Cancer

Jack Cuzick, Mitch Dowsett, Silvia Pineda, Christopher Wale, Janine Salter, Emma Quinn, Lila Zabaglo, Elizabeth Mallon, Andrew R. Green, Ian O. Ellis, Anthony Howell, Aman U. Buzdar, and John F. Forbes

See accompanying article doi: 10.1200/JCO.2011.34.7963 and editorial doi: 10.1200/JCO.2011.37.5824

Jack Cuzick, Silvia Pineda, and Christopher Wale, Centre for Cancer Prevention, Wolfson Institute of Preventive Medicine, Queen Mary University of London; Mitch Dowsett, Janine Salter, Emma Quinn, and Lila Zabaglo, Royal Marsden Hospital, London; Elizabeth Mallon, Western Infirmary, North Glasgow University Hospital, Glasgow; Andrew R. Green and Ian O. Ellis, School of Molecular Medical Sciences, University of Nottingham and Nottingham University Hospitals National Health Services Trust, City Hospital, Nottingham City Hospital, Nottingham; Anthony Howell, Christie Hospital, Manchester Breast Centre and Breakthrough Breast Cancer Research Unit, Manchester, United Kingdom; Aman U. Buzdar, The University of Texas MD Anderson Cancer Center, Houston, TX; and John F. Forbes, University of Newcastle, School of Medical Practice and Population Health, Callaghan, New South Wales, Australia.

Submitted August 2, 2010; accepted May 31, 2011; published online ahead of print at www.jco.org on October 11, 2011.

Written on behalf of the Arimidex, Tamoxifen, Alone or in Combination/Long-Term Anastrozole Versus Tamoxifen Treatment Effects (ATAC/LATTE) trialists.

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

Clinical Trials repository link available on JCO.org.

Corresponding author: Jack Cuzick, PhD, Centre for Cancer Prevention, Wolfson Institute of Preventive Medicine, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, United Kingdom; e-mail: j.cuzick@qmul.ac.uk.

0732-183X/11/2999-1/$20.00

DOI: 10.1200/JCO.2010.31.2835

ABSTRACT

Purpose
We recently reported that the mRNA-based, 21-gene Genomic Health recurrence score (GHI-RS) provided additional prognostic information regarding distant recurrence beyond that obtained from classical clinicopathologic factors (age, nodal status, tumor size, grade, endocrine treatment) in women with early breast cancer, confirming earlier reports. The aim of this article is to determine how much of this information is contained in standard immunohistochemical (IHC) markers.

Patients and Methods
The primary cohort comprised 1,125 estrogen receptor–positive (ER-positive) patients from the Arimidex, Tamoxifen, Alone or in Combination (ATAC) trial who did not receive adjuvant chemotherapy, had the GHI-RS computed, and had adequate tissue for the four IHC measurements: ER, progesterone receptor (PgR), human epidermal growth factor receptor 2 (HER2), and Ki-67. Distant recurrence was the primary end point, and proportional hazards models were used with sample splitting to control for overfitting. A prognostic model that used classical variables and the four IHC markers (IHC4 score) was created and assessed in a separate cohort of 786 patients.

Results
All four IHC markers provided independent prognostic information in the presence of classical variables. In sample-splitting analyses, the information in the IHC4 score was found to be similar to that in the GHI-RS, and little additional prognostic value was seen in the combined use of both scores. The prognostic value of the IHC4 score was further validated in the second separate cohort.

Conclusion
This study suggests that the amount of prognostic information contained in four widely performed IHC assays is similar to that in the GHI-RS. Additional studies are needed to determine the general applicability of the IHC4 score.

J Clin Oncol 29. © 2011 by American Society of Clinical Oncology

INTRODUCTION

With the advent of screening, a large number of small estrogen receptor–positive (ER-positive) tumors are being diagnosed with a generally good prognosis. However, current markers do not allow accurate prediction of the likelihood of recurrence, and improvements are needed to clearly identify which women are at sufficiently low risk to be able to safely avoid the use of chemotherapy and its accompanying adverse effects.[1] We recently reported that the Genomic Health recurrence score (GHI-RS), also known as Onco*type* Dx, provided additional prognostic information regarding distant recurrence beyond that obtained from classical clinicopathologic factors (age, nodal status, tumor size, grade, randomized treatment) in 1,231 patients with ER-positive primary breast cancer who did not receive chemotherapy in the Arimidex, Tamoxifen, Alone or in Combination (ATAC) trial and were randomly assigned to either anastrozole or tamoxifen.[2] The prognostic value of the GHI-RS was also

compared with the prognoses made by Adjuvant! Online, which is based on classical clinical variables, and was found to be largely independent of this score ($R^2 = 0.05$). For node-negative patients, the two scores contributed an almost equal amount of information to the final predictive model. In node-positive patients, the number of positive nodes was the dominant variable, but the GHI-RS still added a significant amount of clinically relevant information, especially when only one to three nodes were positive and other factors did not indicate high risk. Thus, there is an opportunity to integrate clinical and molecular markers to create a combined score with substantially greater prognostic value than either approach alone.

It is recognized that certain well-established immunohistochemical (IHC) markers also provide useful prognostic and, in some cases, predictive information in addition to these classical clinical factors.[3,4] However, no direct comparisons or evaluations have been made with the GHI-RS to date. Here we develop a prognostic score based on four widely measured IHC markers (IHC4)—ER, progesterone receptor (PgR), human epidermal growth factor receptor 2 (HER2; including fluorescent in situ hybridization in the 2+ group), and Ki-67—by using tumor biopsy tissues collected from patients in the ATAC trial and determine the extent to which the four markers provide additional prognostic information not captured by classical clinicopathologic variables. Sample-splitting methods are then used to compare the added information in this score with that added by the predefined GHI-RS. Finally, a prognostic model is given that integrates the IHC4 score with classical variables to obtain an overall prediction model.

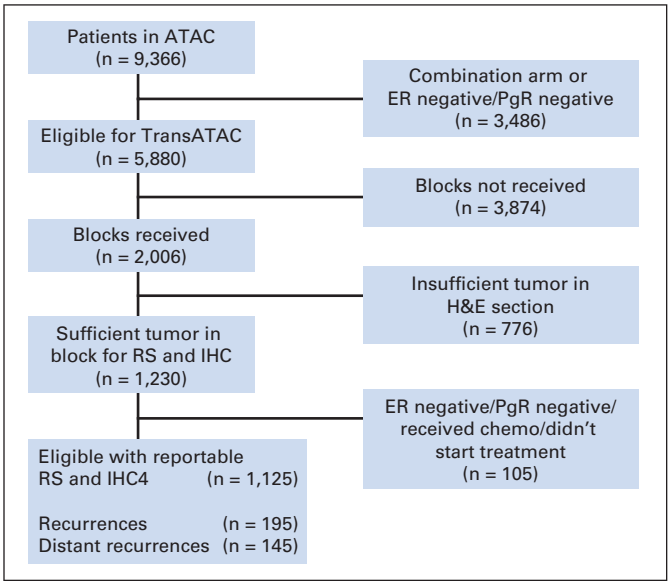

Fig 1. CONSORT diagram indicating which patients from the Arimidex, Tamoxifen, Alone or in Combination (ATAC) trial were evaluated by four immunohistochemical markers (IHC4; estrogen receptor [ER], progesterone receptor [PgR], human epidermal growth factor receptor 2, and Ki-67) and Genomic Health recurrence score (GHI-RS). Chemo, chemotherapy; H&E, hematoxylin and eosin; RS, recurrence score; TransATAC, the translational science substudy within the ATAC study.

PATIENTS AND METHODS

Formalin-fixed paraffin-embedded tumor tissue from women from the tamoxifen and anastrozole arms of the ATAC trial reported to be ER positive and/or PgR positive for whom the GHI-RS had been computed[2] and for whom sufficient additional tissue was available to do the IHC analyses were included. We excluded women who received chemotherapy before trial entry and those found to be both ER negative and PgR negative on central review of the new sections, leaving 1,125 eligible patients with measurements for all parameters (Fig 1). The few patients with unknown nodal status (n = 44) were taken to be node negative, as in previous analyses. Further validation of the IHC4 score was performed by using a cohort of 786 women treated in Nottingham from 1990 to 1998.[5,6] All of these patients were ER positive (H-score > 10; H-score is defined as the percentage of cells staining weakly plus two times the percentage of cells staining moderately plus three times the percentage of cells staining strongly) and received either tamoxifen or no endocrine treatment.

Laboratory Methods

Tissue microarrays (TMAs) were constructed by using a manual tissue arrayer (MTA-1; Beecher Instruments, Sun Prairie, WI) with 600-μm tissue cores. Hematoxylin and eosin–stained slides were reviewed by a pathologist and/or an experienced technician, and three representative areas that contained invasive tumor cells were selected. Areas of invasive tumor away from in situ or benign tissue components were marked on both the slides and corresponding paraffin blocks for TMA construction. Three cores were extracted from each donor block and were assembled into three recipient blocks.

ER, PgR, HER2, and Ki-67 IHC and Scoring

IHC for ER, PgR, and HER2 was conducted as previously described[4] and IHC for Ki-67 was described in detail elsewhere.[7] ER and Ki-67 analyses were performed on 4-μm sections from the triplicate TMA blocks, and PgR and HER2 analyses were performed on single 4-μm whole sections from the donor blocks used in the TMA construction. ER was quantified by using the H-score and was considered positive if greater than 1. The variable ER_{10} was obtained by dividing the H-score by 30 to obtain a variable with a range of 0 to 10. PgR was scored as a percentage of cells staining positive with a positive cutoff of 10%. PgR_{10} was obtained by dividing this percentage by 10 to obtain a variable with a range of 0 to 10. HER2 was scored according to the manufacturer's recommendation: 3+ was positive and equivocal 2+ samples underwent fluorescent in situ hybridization analysis and were considered positive only if the ratio was more than 2. Ki-67 scores were recorded as the percentage of positively staining malignant cells.

Similar methods and scoring algorithms were used for the Nottingham cohort, except that the MiB1 antibody was used on whole sections for Ki-67, and TMAs were used for ER, PgR, and HER2. For more details, see Appendix 1 (online only).

Statistical Methods

The four variables used to derive the IHC4 score were determined prospectively before examining the data. These were evaluated in both univariate and multivariate proportional hazards models by using age ($< 65, \geq 65$ years), nodal status (0, 1-3, 4+), tumor size (≤ 1 cm, > 1 to ≤ 2 cm, > 2 to ≤ 3 cm, > 3 cm) centrally read grade (poor, intermediate, well differentiated), and randomized treatment (anastrozole *v* tamoxifen). The primary end point was time to distant recurrence (TTDR). Follow-up was based on the 100-month median follow-up database.[8] The contribution of each of the four variables was evaluated by the change in likelihood ratio χ^2 (LR-χ^2; 1 *df*) in three ways: by univariate analyses, as an addition to a model containing only the clinical variables, and as a decrease in LR-χ^2 when the variable was removed from the full model containing the clinical variables and all four IHC variables. Finally, a model was created that used all the data to give the best IHC4 score when combined with the classical variables, leading to an overall prognostic score. To allow for the small amount of overfitting of the IHC4 score, a shrinkage correction was used[9] for the final score. To compare the IHC4 score against the GHI-RS, sample splitting was used in which the IHC4 score was generated by using half the data, and then this score was compared with the GHI-RS in the remaining half of the data. For the Nottingham cohort, exactly the same combined score was used except that about half the patients in that cohort received no endocrine treatment, and an additional indicator variable was added to account for this. Manual reading of Ki-67 was also used, leading to

on October 27, 2011 for 134.100.99.149

Table 1. Prognostic Value of Individual Components of the IHC4 Score As Assessed by Change in the χ^2_1 Value Based on the Likelihood Ratio Statistic

Variable	Patient Group	Univariate		Addition to Clinical Model		Removal From Clinical Model Plus IHC4 Model	
		χ^2_1	P	χ^2_1	P	χ^2_1	P
ER_{10}	All	11.53	.0007	8.40	. 004	3.31	.07
	Node negative	15.82	.0001	13.17	.0003	6.41	.011
PgR_{10}	All	21.84	$< 10^{-5}$	18.77	$< 10^{-4}$	10.20	.0014
	Node negative	13.75	< .0002	8.88	.003	1.36	.24
HER2	All	23.13	$< 10^{-5}$	16.15	.0001	6.97	.008
	Node negative	24.87	$< 10^{-6}$	18.49	$< 10^{-4}$	8.77	.003
Ki-67	All	31.85	$< 10^{-7}$	9.29	.002	7.00	.008
	Node negative	26.18	$< 10^{-6}$	10.51	.001	6.62	.01

NOTE. The magnitude of the χ^2_1 value gives a quantitative estimate of the amount of information provided by the variables.
Abbreviations: ER_{10}, estrogen receptor H-score [H-score is defined as the percentage of cells staining weakly plus two times the percentage of cells staining moderately plus three times the percentage of cells staining strongly divided by 30 to obtain a variable with a range of 0 to 10]; HER2, human epidermal growth factor receptor 2; IHC4, score for four immunohistochemical markers (ER, PgR, HER2, and Ki-67); PgR_{10}, progesterone receptor scored as a percentage of cells staining positive with a positive cutoff of 10%, with the score divided by 10 to obtain a variable with a range of 0 to 10; χ^2_1, χ^2 on one degree of freedom.

higher readings, and a rescaling factor of 2.5 (based on the ratio of median levels) was used to normalize the readings. Full details are provided in Appendix 2 (online only).

RESULTS

Values for the GHI-RS and the four IHC markers were available for 1,125 women of whom 793 were node-negative (Fig 1). There were 195 recurrences of which 145 were distant recurrences. In node-negative women, there were 101 recurrences of which 67 were distant recurrences.

Contribution of Individual Markers

We began by determining the value of each of the four IHC markers in three ways: by univariate analyses, as an addition to a model containing the classical variables, and when added to a model containing the classical variables and the other three IHC markers (Table 1). This was done for all women and separately for node-negative women only. It can be seen that each of the four variables added a significant amount of information; Ki-67 was the most powerful in univariate analyses but not in multivariate analyses because of its correlation with grade. For the multivariate models, PgR was most prognostic overall but less so in node-negative patients in whom ER, HER2, and Ki-67 had similar values.

Creation of IHC4 Score

Next we determined the most informative combination of the four IHC variables for distant recurrence in the presence of classical variables. These variables showed only modest correlation except for the expected correlations between PgR with HER2 and grade with HER2 and Ki-67 (Appendix Table A1, online only). The overall contribution of the IHC measurements was highly significant (χ^2 4 $df = 39.1$; $P < .0001$), and the shrinkage-adjusted IHC4 score was computed as

$$\text{IHC4} = 94.7 \times \{-0.100\ \text{ER}_{10} - 0.079\ \text{PgR}_{10} + 0.586\ \text{HER2} + 0.240 \ln(1 + 10 \times \text{Ki67})\}.$$

The optimal underlying clinical score (as part of an overall model) was found to be

$$\text{clinical score} = 100 \times \{0.417\text{N}_{1\text{-}3} + 1.566\text{N}_{4+} + 0.930(0.497\text{T}_{1\text{-}2} + 0.882\text{T}_{2\text{-}3} + 1.838\text{T}_{>3} + 0.559\text{Gr}_2 + 0.970\text{Gr}_3 + 0.130\text{Age}_{\geq 65} - 0.149\text{Ana})\},$$

where N_j, T_j, Gr_j, and Age_j denote categories of nodal status, tumor size, grade, and age, respectively, and Ana denotes treatment with anastrozole as opposed to tamoxifen. The overall score was the sum of these two scores. A shrinkage factor of $0.947 = (35.1/39.1)^{1/2}$ for the IHC4 score and $0.930 = (45.1/52.1)^{1/2}$ for the non-nodal part of clinical score was applied to allow for the small amount of overfitting. For ease of interpretation, these scores have been multiplied by 100. Similar IHC4 scores were obtained if the end point was all recurrences, if only node-negative women were included, or if local grade was used in the clinical score and they were highly correlated (Pearson correlation > 0.93 for all pairwise comparisons). The χ^2 for the clinical variables using all patients was 147 (9 *df*). Similar models were obtained when analyses were restricted to node-negative patients or when any recurrence was used as an end point. A similar prognostic value was obtained for node-negative patients ($\chi^2_4 = 35.4$) but now the classic variables were less predictive ($\chi^2_7 = 40.7$) because of the omission of nodal status. For these patients, the IHC4 contained almost as much information as the remaining classical variables. A histogram of the IHC4 score for all patients is shown in Figure 2. The median is -4.2 and the interquartile range (IQR) is -29.9 to 29.9. The hazard ratio (HR) for a change from the 25th to 75th percentile of the IHC4 score for all patients was 5.7 (95% CI, 3.4 to 9.7) in a univariate analysis and 3.9 (95% CI, 2.4 to 6.7) when added to the clinical score. A slight skewness with a longer upper tail is noted. A similar histogram is obtained when restricted to node-negative patients, with median -6.0 and IQR of -31.0 to 27.0.

None of the patients in this study received trastuzumab if they were HER2-positive, as would be current practice. Thus, it is of interest to see how the model performed on the 1,066 patients who were HER2-negative (in whom there were 179 recurrences, 124 of which were distant recurrences). A model termed "IHC3" developed on such

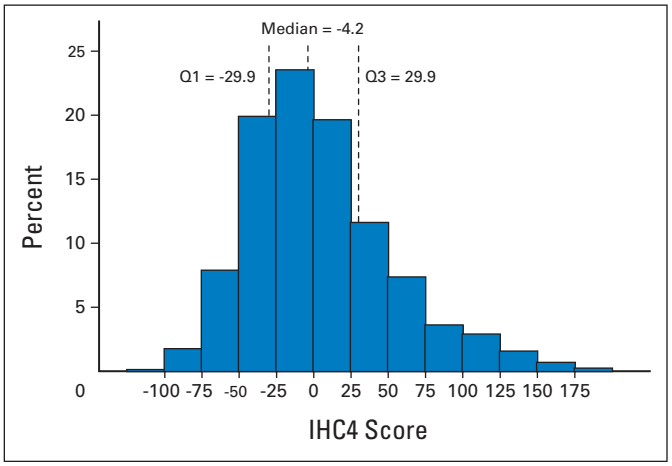

Fig 2. Histogram of IHC4 score (score for four immunohistochemical markers: estrogen receptor, progesterone receptor, human epidermal growth factor receptor 2, and Ki-67) with median and interquartile range for all patients (see Appendices 1 and 2, online only). Q, quartile.

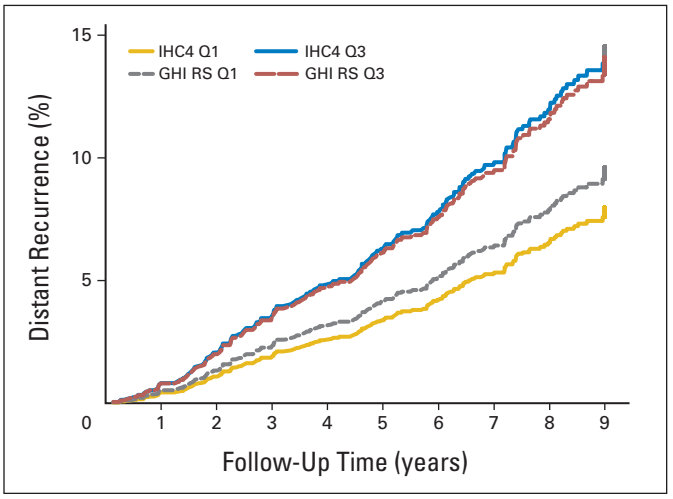

Fig 3. Example of predicted time to distant recurrence for a node-negative patient age ≥ 65 years with a poorly differentiated 1- to 2-cm tumor treated with anastrozole who is at either the 25th (quartile 1 [Q1]) or 75th (Q3) percentile of the IHC4 score [score for four immunohistochemical markers: estrogen receptor, progesterone receptor, human epidermal growth factor receptor 2, and Ki-67] or the Genomic Health recurrence score (GHI-RS).

patients was highly prognostic ($\chi^2_3 = 22.4$; $P < .0001$) and had the following form:

$$\text{IHC3} = 93.1 \times \{-0.086\ \text{ER}_{10} - 0.081\ \text{PgR}_{10} + 0.281 \ln(1 + 10 \times \text{Ki67})\},$$

which was virtually identical to IHC4 when HER2 was negative.

Comparison With GHI-RS

The mean changes in LR-χ^2 for the addition of either IHC4 or GHI-RS or both to the classical score in the validation halves of 100 random splits of the data are shown in Table 2. Higher values indicate more added prognostic information. Separate classical scores were computed when evaluating IHC4 score and GHI-RS, but the two classical scores were highly correlated ($r = 0.998$) and the choice of which one was used had a minimal impact on the outcome. The IHC4 score was modestly correlated with the GHI-RS ($r = 0.72$), indicating some overlap in information but also a considerable difference. Similar correlations were found when the GHI-RS was compared with the IHC4 score determined by recurrence or when restricted to node-negative patients ($r = 0.68$ for TTDR, N_0; $r = 0.71$ for time to recurrence [TTR], all; $r = 0.69$ TTR, N_0), indicating an R^2 of approximately 50% between these scores. These procedures indicate that the amount of prognostic information provided by the IHC4 score is similar to that provided by the GHI-RS. This was true regardless of whether TTDR or TTR was used as an end point, or whether all patients were included or just the node-negative subset. The IHC4 score was slightly more prognostic for distant recurrences whereas the GHI-RS was slightly more prognostic for all recurrences. Using both scores together provided slightly more information than using either of the scores individually. The average coefficient in the validation halves of the data for the IHC4 score derived from the training sets was 0.894, which agrees closely with the predicted value for the split sample of 0.889, supporting the shrinkage estimate used above for the final IHC4 score. However the shrinkage value predicted for the whole sample is nearer to unity because of the larger sample size.

Examples of Prognostic Power

To illustrate these results more fully, Kaplan-Meier curves for a node-negative, 1- to 2-cm poorly differentiated tumor receiving anastrozole are shown for either the 25th or 75th percentile of each score (Fig 3). At 9 years, the difference for distant recurrence was 7.6% versus 13.9% for the IHC4 score and 9.2% versus 13.4% for the GHI-RS. Figure 4 shows the relationship of the combined clinical/IHC4 score with the risk of distant

Table 2. Average Change in LR-χ^2 Value and 95% Bootstrap CIs Based on 100 Resamplings for Assessing the Amount of Information Added by the IHC4 Score or the GHI-RS to the Clinical Score and the GHI-RS to a Model Containing the Clinical Score and IHC4 Score

	TTDR				TTR			
	All Patients		Node-Negative Patients		All Patients		Node-Negative Patients	
Model	LR-χ^2	95% CI	LR-χ^2	95% CI	LR-χ^2	95% CI	LR-χ^2	95% CI
C + IHC4 *v* C (1 *df*)	29.3	27.7 to 30.3	29.9	28.5 to 31.2	21.1	19.5 to 21.6	23.0	22.3 to 24.8
C + GHI-RS *v* C (1 *df*)	25.3	25.2 to 25.9	20.9	20.7 to 21.6	25.6	25.2 to 25.9	25.7	25.4 to 26.4
C + IHC4 + GHI-RS *v* C (2 *df*)	33.3	32.6 to 34.2	32.6	31.8 to 34.1	29.4	28.6 to 29.8	30.6	30.3 to 31.8
C + IHC4 + GHI-RS *v* C + IHC4 (1 *df*)	3.1	2.8 to 3.9	1.6	1.3 to 2.1	7.3	7.1 to 8.6	6.7	5.7 to 7.3

NOTE. Results are given separately for TTDR, TTR, all patients, or only node-negative patients.
Abbreviations: C, clinical score; GHI-RS, Genomic Health recurrence score; IHC4, four immunohistochemical markers (estrogen receptor, progesterone receptor, human epidermal growth factor receptor 2, and Ki-67); LR-χ^2, χ^2 value based on the likelihood ratio statistic; TTDR, time to distant recurrence; TTR, time to recurrence.

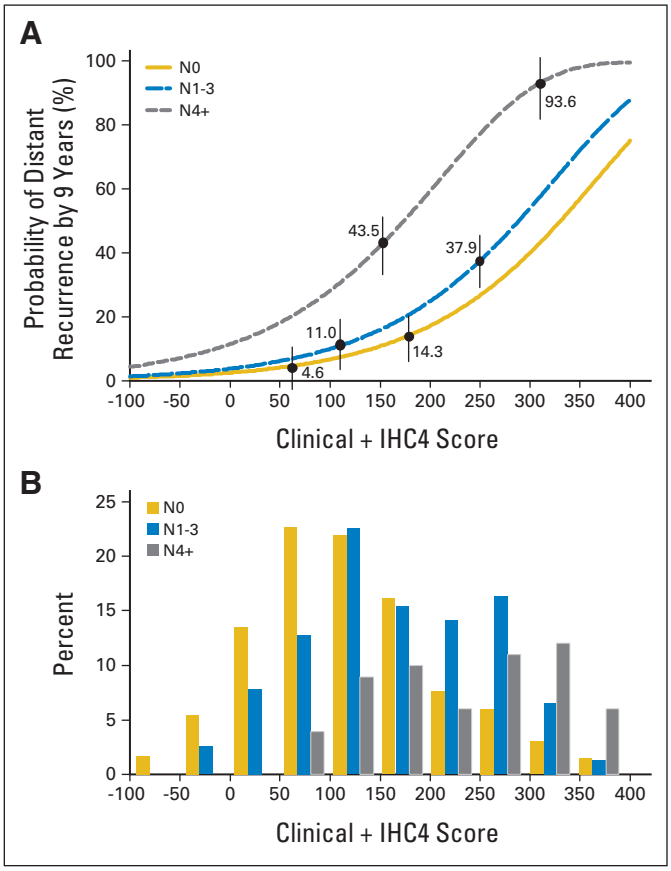

Fig 4. Predicted 9-year distant recurrence probabilities for different nodal status groups according to the sum of the IHC4 score (score for four immunohistochemical markers: estrogen receptor, progesterone receptor, human epidermal growth factor receptor 2, and Ki-67) and clinical score (after removing nodal status term) and histogram for clinical score + IHC4 score for each nodal status. The predicted 9-year distant recurrence rates for a patient at the 25th and 75th percentile of the score for each nodal category are indicated.

recurrence after 9 years separately for each nodal category. Differences between the two scores for predicted 9-year distant recurrence rates for a range of different clinical parameters are shown in Table 3.

Further Validation of IHC4 in the Nottingham Cohort

Further validation of the IHC4 score was performed by using a cohort of 786 ER-positive women treated in Nottingham. About half these women received no endocrine treatment and were somewhat younger because they did not have to be postmenopausal, but otherwise they were well matched to the TransATAC (the translational science substudy within the ATAC study) cohort (Appendix Table A2, online only). Because Ki-67 levels were on average about 2.5 times higher on account of the manual reading and use of the MiB1 antibody (Appendix Table A2), they were rescaled in the IHC4 score by this amount (ie, a multiplier of 4 was used instead of 10 before taking logarithms). The adjusted IHC4 score was highly significantly prognostic for outcome (HR, 4.1; 95% CI, 2.5 to 6.8) for a change from the 25th to 75th percentile in a univariate analysis and gave similar results when added to the clinical score (HR, 3.9; 95% CI, 2.3 to 6.5; $\Delta\chi^2 = 25.89$; $P < .0001$). Patients were divided into three prognostic tertiles using the combined score derived from the TransATAC data, and observed and expected distant recurrence Kaplan-Meier curves (based on the average of the predicted survival curves within tertile using the TransATAC model) are shown in Figure 5. Similar results are seen if the data are restricted to the patients treated with tamoxifen.

Table 3. Examples of Predicted 9-Year Distant Recurrence Probabilities for 25th and 75th Percentiles of the IHC4 and GHI-RS Scores for Different Grades and Nodal Status for a Women Age ≥ 65 Years With a 1- to 2-cm Tumor Treated With Anastrozole

Nodal Status	Grade / Percentile Score	Poor or Undifferentiated IHC4	Poor or Undifferentiated GHI-RS	Moderate IHC4	Moderate GHI-RS	Well Differentiated IHC4	Well Differentiated GHI-RS
Node Negative	25	7.1	8.3	4.8	5.8	2.2	2.5
	75	13.1	12.1	8.9	8.4	4.2	3.6
Node Positive	25	10.4	12.1	7.1	8.4	3.3	3.6
	75	18.8	17.3	13.0	12.2	6.2	5.3

Abbreviations: GHI-RS, Genomic Health recurrence score; IHC4, four immunohistochemical markers (estrogen receptor, progesterone receptor, human epidermal growth factor receptor 2, and Ki-67).

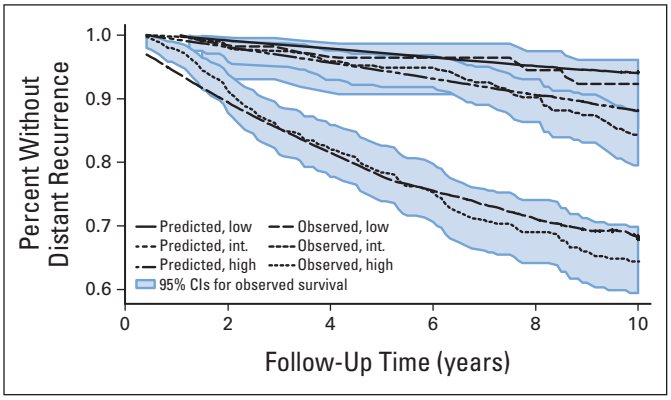

Fig 5. Predicted versus observed time to distant recurrence (Kaplan-Meier curves plus 95% CIs) for the tertiles of the combined score in the Nottingham data. int., intermediate.

DISCUSSION

We have shown that accurate quantitative IHC measures of ER, PgR, HER2, and Ki-67 provide additional prognostic information in this population that is at least as informative as the RNA-based GHI-RS. We have created a prognostic model that integrates this information with classical clinical and pathologic variables and may prove helpful in managing early ER-positive breast cancer in postmenopausal patients.

The GHI-RS was created by analyses performed within a single laboratory and has been shown to be highly reproducible. Although it has become widely used in the United States, its cost is an impediment to its use in many centers in other countries. An attraction of IHC4 is that three of the tests (ER, PgR, and HER2) are conducted in the routine workup of almost all breast cancers and that Ki-67 is also measured in many centers. The Ki-67 method in this study used the SP6 antibody and image analysis. This provides highly correlated but lower values than manual scoring with the widely used MiB1 antibody,[7] which would require an adjusted coefficient within the IHC4. On the basis of the Nottingham data, this adjustment should be by a factor of about 2.5. Thus IHC4 constitutes an alternative, inexpensive test battery that can provide prognostic utility similar to that of the GHI-RS and can identify women at such low risk of

recurrence that no meaningful benefit from chemotherapy can be envisaged for them. Lack of reproducibility of IHC assays is, however, an issue of concern in considering extension of the use of IHC4 to other laboratories. Differences in IHC values can occur as a result of variability in several factors including fixation, antigen retrieval, reagents, and interpretation. Several quality assurance programs have been created, such as the United Kingdom National External Quality Assessment Service (NEQAS), and these have been shown to lead to marked improvements in between-laboratory agreement.[10] Further improvements in the standardization of assays may also be aided by the American Society of Clinical Oncologists/College of Pathologists (ASCO/CAP) guidelines for HER2[11] and more recently for ER and PgR.[12] We have validated our results using another data set in which the assays were done in a separate laboratory, but because IHC4 offers the possibility to do this test in local laboratories, full validation would require evaluation of the IHC4 score when carried out in a range of local laboratories.

None of the clinical or IHC variables provided predictive information regarding the choice between tamoxifen or anastrozole. In addition, the analysis was conducted in women who did not receive chemotherapy because the aim was to identify women at such low risk of distant recurrence after endocrine treatment that they might be safely spared chemotherapy. Development of a model that would predict the magnitude of benefit from chemotherapy is also desirable. The low- and high-risk categories identified by GHI-RS have been reported to separate women into those who would gain little or substantial benefit from adjuvant chemotherapy, respectively.[13,14] A similar relationship is may exist for the IHC4 score, but direct verification is needed.

Overall, these data suggest that four standard IHC assays performed in a high-quality laboratory can provide amounts of prognostic information similar to that provided by the GHI-RS for endocrine-treated ER-positive breast cancer patients. This information is clinically relevant and the difference between being at the 25th percentile and the 75th percentile translates into an almost two-fold difference in 9-year distant recurrence rates (Table 3). This approach has a wide applicability and could extend the circumstances in which improved prognostic information is routinely available.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Although all authors completed the disclosure declaration, the following author(s) indicated a financial or other interest that is relevant to the subject matter under consideration in this article. Certain relationships marked with a "U" are those for which no compensation was received; those relationships marked with a "C" were compensated. 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.

Employment or Leadership Position: None **Consultant or Advisory Role:** Jack Cuzick, AstraZeneca (C); Mitch Dowsett, AstraZeneca (C); Anthony Howell, AstraZeneca (C) **Stock Ownership:** None **Honoraria:** Jack Cuzick, AstraZeneca; Mitch Dowsett, AstraZeneca; Elizabeth Mallon, Roche; Anthony Howell, AstraZeneca; John F. Forbes, AstraZeneca **Research Funding:** Jack Cuzick, AstraZeneca; Mitch Dowsett, AstraZeneca; Anthony Howell, AstraZeneca **Expert Testimony:** Mitch Dowsett, AstraZeneca (C) **Other Remuneration:** None

AUTHOR CONTRIBUTIONS

Conception and design: Jack Cuzick, Mitch Dowsett, John F. Forbes
Provision of study materials or patients: Andrew R. Green, Ian O. Ellis, Anthony Howell, Aman U. Buzdar, John F. Forbes
Collection and assembly of data: Jack Cuzick, Mitch Dowsett, Christopher Wale, Janine Salter, Emma Quinn, Lila Zabaglo, Elizabeth Mallon, Andrew R. Green, Ian O. Ellis, Aman U. Buzdar
Data analysis and interpretation: Jack Cuzick, Mitch Dowsett, Christopher Wale, Silvia Pineda, Anthony Howell, Aman U. Buzdar, John F. Forbes
Manuscript writing: Jack Cuzick, Mitch Dowsett, Christopher Wale, Silvia Pineda, Janine Salter, Emma Quinn, Lila Zabaglo, Elizabeth Mallon, Andrew R. Green, Ian O. Ellis, Anthony Howell, Aman U. Buzdar, John F. Forbes
Final approval of manuscript: Jack Cuzick, Mitch Dowsett, Christopher Wale, Silvia Pineda, Janine Salter, Emma Quinn, Lila Zabaglo, Elizabeth Mallon, Andrew R. Green, Ian O. Ellis, Anthony Howell, Aman U. Buzdar, John F. Forbes

REFERENCES

1. Dowsett M, Goldhirsch A, Hayes DF, et al: International Web-based consultation on priorities for translational breast cancer research. Breast Cancer Res 9:R81, 2007
2. Dowsett M, Cuzick J, Wale C, et al: Prediction of risk of distant recurrence using the 21-gene recurrence score in node-negative and node-positive postmenopausal patients with breast cancer treated with anastrozole or tamoxifen: A TransATAC study. J Clin Oncol 28:1829-1834, 2010
3. Mauriac L, Keshaviah A, Debled M, et al: Predictors of early relapse in postmenopausal women with hormone receptor-positive breast cancer in the BIG 1-98 trial. Ann Oncol 18:859-867, 2007
4. Dowsett M, Allred C, Knox J, et al: Relationship between quantitative estrogen and progesterone receptor expression and human epidermal growth factor receptor 2 (HER-2) status with recurrence in the Arimidex, Tamoxifen, Alone or in Combination trial. J Clin Oncol 26:1059-1065, 2008
5. Aleskandarany MA, Rakha EA, Macmillan RD, et al: MIB1/Ki-67 labelling index can classify grade 2 breast cancer into two clinically distinct subgroups. Breast Cancer Res Treat 127:591-599, 2011
6. Abd El-Rehim DM, Ball G, Pinder SE, et al: High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses. Int J Cancer 116:340-350, 2005
7. Zabaglo L, Salter J, Anderson H, et al: Comparative validation of the SP6 antibody to Ki67 in breast cancer. J Clin Pathol 63:800-804, 2010
8. Arimidex, Tamoxifen, Alone or in Combination (ATAC) Trialists' Group, Forbes JF, Cuzick J, et al: Effect of anastrozole and tamoxifen as adjuvant treatment for early-stage breast cancer: 100-month analysis of the ATAC trial. Lancet Oncol 9:45-53, 2008
9. Gruber MHJ: Improving Efficiency by Shrinkage. New York, NY, Marcel Dekker, 1998
10. Rhodes A, Jasani B, Balaton AJ, et al: Immunohistochemical demonstration of oestrogen and progesterone receptors: Correlation of standards achieved on in house tumours with that achieved on external quality assessment material in over 150 laboratories from 26 countries. J Clin Pathol 53:292-301, 2000
11. Wolff AC, Hammond ME, Schwartz JN, et al: American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. J Clin Oncol 25:118-145, 2007
12. 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
13. 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
14. Albain KS, Barlow WE, Shak S, et al: Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: A retrospective analysis of a randomised trial. Lancet Oncol 11:55-65, 2010

Acknowledgment

Supported by the Royal Marsden National Institute for Health Research Biomedical Research Centre, Cancer Research United Kingdom Program Grant No. C569-10404 and grants from Breakthrough Breast Cancer and from AstraZeneca.

Presented at the 32nd Annual San Antonio Breast Cancer Symposium, San Antonio, TX, December 9-13, 2009.

Appendix

Appendix 1

Our initial pilot studies demonstrated the suitability of the TransATAC (the translational science substudy within the Arimidex, Tamoxifen, Alone or in Combination [ATAC] study) tissue microarrays (TMAs) for estrogen receptor (ER) and Ki-67 analysis but not for progesterone (PgR) and human epidermal growth factor receptor 2 (HER2) analysis. As a result, ER and Ki-67 analyses were performed on 4-μm sections from the triplicate TMA blocks, and PgR and HER2 analyses were performed on single 4-μm whole sections from the donor blocks used in the TMA construction. Sections were picked up on charged slides, dewaxed in xylene, and rehydrated in decreasing grades of industrial methylated spirits. Antigen retrieval was performed for all markers: ER, PgR, and Ki-67 were microwaved for 10 minutes in citrate buffer pH 6.0, and HER2 was heated for 40 minutes in HercepTest antigen retrieval buffer (Dako, Copenhagen, Denmark) at 97°C in a waterbath. All slides were stained on the Dako autostainer by using either the REAL detection kit protocol or HercepTest. ER, PgR, and Ki-67 were demonstrated by using the 6F11 antibody (Vector Laboratories, Burlingame, CA) diluted 1:40, clone 16 (Vector Laboratories) diluted 1:100, or SP6 antibody (Abcam, Cambridge, MA) diluted 1:100, respectively. All dilutions and washes were performed with Dako antibody diluent and Dako wash buffer, respectively. Sections were then counterstained with Mayer's hematoxylin. HER2 was demonstrated by using the HercepTest kit per manufacturer's instructions followed by Vysis PathVysion (fluorescent in situ hybridization [FISH]) in those samples scored at 2+ by immunohistochemisty (IHC).

ER was quantified by using the H-score, which is defined as the percentage of cells staining weakly plus two times the percentage of cells staining moderately plus three times the percentage of cells staining strongly. ER was considered positive if the H-score was greater than 1. The variable ER_{10} was obtained by dividing the H-score by 30 to obtain a variable with a range of 0 to 10. PgR was scored as a percentage of cells staining positive with a positive cutoff of 10%. PgR_{10} was obtained by dividing this percentage by 10 to obtain a variable with a range of 0 to 10. HER2 was scored according to the manufacturer's recommendation: 3+ was positive, and equivocal 2+ cases underwent FISH analysis to determine the level of HER2 amplification. Tumors that were 3+ positive or 2+ positive with a FISH ratio of more than 2.0 were regarded as HER2 positive.

Ki-67–stained slides were scanned with the Applied Imaging Ariol image analysis system (Genetix, San Jose, CA) by using the TMAsight assay with a ×20 objective. Images acquired through three filters (red, green, and blue) were converted by Ariol software into color reconstructions. MultiStainHighRes script was used to analyze images by using classifiers established during training. The analysis was performed only on invasive tumor areas in the individual cores. Ki-67 scores were recorded as the percentage of positively staining malignant cells.

Further validation of the immunohistochemistry score for four markers (IHC4; ER, PgR, HER2, and Ki-67) was performed by using a cohort of 786 women treated for primary operable invasive breast cancer in Nottingham from 1990 to 1998. All of these patients were ER positive (H-score $>$ 10) and received either adjuvant tamoxifen or no endocrine treatment. Information on local, regional, and distant recurrence and survival is maintained on a prospective basis.

Patients were followed up at 3-month intervals initially, then every 6 months, then annually. Similar IHC methods that used the standard streptavidin-biotin complex method and scoring algorithms were used for the Nottingham cohort, except that the MiB1 antibody (MIB1 clone; Dako) was used on whole sections for Ki-67[5] and TMAs were used for ER (1D5 clone; Dako), PgR (PgR636 clone; Dako), and HER2 (Herceptest, Dako), as previously described.[6] Ki-67 scoring was performed in areas with the highest number of positive nuclei (hot spot) within the invasive component of the tumor and expressed as the percentage of 1,000 malignant cells.[5] At least two TMA cores, one core obtained from the center and the other from the periphery of the tumor, were evaluated from each tumor for ER and PgR by using the H-score[6]. HER2 status was evaluated by using HercepTest guidelines (Dako) and confirmed by using CISH (DuoCISH, Dako).

Appendix 2

The variables used to derive the IHC4 score were determined prospectively before examining the data. These were evaluated by both univariate and multivariate analyses in a proportional hazards model by using age ($<$ 65, $\geq$ 65 years), nodal status (0, 1-3, 4+), tumor size ($\leq$ 1 cm, $>$ 1 to $\leq$ 2 cm, $>$ 2 to $\leq$ 3 cm, $>$ 3 cm) centrally read grade (poor, intermediate, well differentiated), and randomized treatment (anastrozole *v* tamoxifen). Analyses with locally read grade were also undertaken. The primary end point was time to distant recurrence. Thus locoregional recurrence or contralateral disease was ignored, and patients were censored at the time of death, if this occurred before a distant recurrence. Additional analyses were also undertaken by using any recurrence (local, distant, or contralateral), censoring only at death without recurrence. Follow-up was based on the 100-month median follow-up database.[6] The contribution of each of the four variables was evaluated by the change in likelihood ratio χ^2 (LR-χ^2; 1 *df*) in three ways: by univariate analysis, as an addition to a model containing only the clinical variables, and as a decrease in LR-χ^2 when the variable was removed from the full model containing the clinical variables and all four IHC variables. Finally, a model was created that used all the data to give the best IHC4 score when combined with the classical variables leading to an overall prognostic score, which was split into the sum of a clinical score and the IHC4 score. The

information contained in the IHC4 score was assessed by change in the LR-χ^2 (with 4 *df*) when the four variables were added to a model containing the classical clinical variables. A shrinkage correction[9] was then applied to adjust for the small amount of overfitting associated with this approach by rescaling the IHC4 score to have an LR-χ^2 that equaled the observed LR-χ^2 minus the number of degrees of freedom (df = 4). The validity of this was confirmed in the split-sample analyses described below.

Illustrative examples of the difference in 9-year distant recurrence percentages at the 25th versus the 75th percentile of the IHC4 score were given for selected clinical parameter combinations, and a graph was created to estimate these proportions for any value of the combined clinical and IHC4 score.

Comparison of the IHC4 with the Genomic Health, Inc., recurrence score (GHI-RS) presented additional difficulties, since the GHI-RS was predefined, and no such predefined score existed for IHC4. To overcome this, we split the entire sample randomly into two equal halves. The first half (training set) was used to compute an IHC4 score for the added value of these four variables to the classical features. A clinical score was also created from the coefficients obtained from the full model to reflect the contribution of the classical variables as a single score. These two scores were then used as the only independent variables in the remaining half of the data (validation), where the prognostic value of the IHC4 score was assessed by the increase in the LR-χ^2 when this score was added to the classical score. The validation set was then used in the same manner to determine the increase in LR-χ^2 when the GHI-RS was added to a newly computed clinical score on the basis of using the same clinical variables in the training set but in the presence of the GHI-RS. Use of the clinical score obtained from the IHC4 model instead of this score had little effect on the results because they were highly correlated. Since the IHC4 was created and tested on independent data sets, any overfitting due to the estimation of the coefficients for the model was eliminated. The process was then repeated after interchanging the role of the training and validation sets, and the two χ^2 values were summed on the square root scale (ie, each LR-χ^2 was squared-rooted, summed, and then squared). To minimize any fluctuation due to the random splitting of the data, this was repeated 100 times, and the resulting LR-χ^2 values were averaged on the square root scale, as above. For comparison the same process was done for the GHI-RS and a model including both the ICH4 score and the GHI-RS. To obtain CIs for these averaged LR-χ^2 values, a bootstrap analysis was performed with 100 replications.

To assess the validity of our shrinkage correction on the whole sample, the coefficients obtained for the IHC4 score and clinical score (omitting nodal status) in the validation half of the samples were averaged over 1,000 sample splits. Values near to unity indicate minimal overfitting, and smaller values quantify the extent of overfitting. The shrinkage adjustment estimated by this approach was then compared against the estimates based on LR-χ^2 described above, when applied to the split-sample data sets.

Further validation was performed in a separate data set from Nottingham. Exactly the same coefficients were used for the components of the IHC4 score as well as the clinical score, but the Ki-67 values were rescaled by a factor of 2.5 to account for the higher values obtained from manual reading. The only other modification was to add an additional term for no endocrine treatment, because such a group did not exist in the TransATAC data set. Patients were divided into three prognostic tertiles that used the combined score derived from the TransATAC data (with adjustment for no endocrine treatment), and observed and predicted distant recurrence Kaplan-Meier curves (based on the average of the predicted survival curves within tertile) were computed. The model was found to discriminate quite well despite use of a different antibody for Ki-67.

Our conclusions are based on a split-sample evaluation within a single large multicenter trial, whereas the GHI-RS was verified in a separate study. However, the verification half of our study (n = 562) was similar in size to the entire GHI-RS validation study (National Surgical Adjuvant Breast and Bowel Project (NSABP) B-14; N = 675[12]), and the training half was larger than the trials used to formally develop the recurrence score from a panel of 250 candidate genes (NSABP-20 and other studies; N = 447 (Paik S, et al: N Engl J Med 351:2817-2826, 2004; Mamounas EP, et al: J Clin Oncol 28:1677-1683, 2010). We have also taken great care to adjust for any overfitting of the final IHC4 score and have validated it in an entirely independent patient cohort.

Table A1. Association Between Individual Terms Comparing IHC4 Score and Clinical Variables

	PgR (Spearman)			Ki-67 (Spearman)			HER2 Positive/Negative (Wilcoxon)		Nodes (N_0, N_{1-3}, N_{4+}) (Cuzick trend)		Tumor size (mm) (Spearman)			Grade (well-differentiated, moderate, poor or undifferentiated) (Cuzick trend)		Age (years) (Spearman)		
Term	ρ	95% CI	*P*	ρ	95% CI	*P*	Median	*P*	Median	*P*	ρ	95% CI	*P*	Median	*P*	ρ	95% CI	*P*
ER (H-score)	0.14	0.08 to 0.20	< .0001	0.1	0.04 to 0.16	.0008	156 *v* 150	< .0001	156 *v* 155 *v* 155	.6	−0.10	−0.04 to −0.16	.001	155 *v* 156 *v* 154	.8	0.21	0.15 to 0.27	< .0001
PgR (%)				−0.07	−0.01 to −0.13	< .0001	70 *v* 17	< .0001	68 *v* 62 *v* 59	.3	−0.05	0.008 to −0.11	.07	72 *v* 67 *v* 55	.001	−0.04	0.02 to −0.08	.2
Ki-67 (%)							37 *v* 45	< .0001	68 *v* 62 *v* 59	.3	0.13	0.07 to 0.19	< .0001	30 *v* 37 *v* 47	< .0001	0.13	0.07 to 0.19	< .0001
HER2																		
Positive (%)									10.3 *v* 10.4 *v* 10.3	.9*				4.4 *v* 9.0 *v* 21.4	< .0001*			
Median											17 *v* 19.5		.016†			63.6 *v* 62.1		.5†

NOTE. Tests used were Spearman's rank correlation, Pearson's χ^2, linear Logit model, Wilcoxon, and Cuzick trend test (Cuzick J: Stat Med 4:87-90, 1985).
Abbreviations: ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; H-score, the percentage of cells staining weakly plus two times the percentage of cells staining moderately plus three times the percentage of cells staining strongly divided by 30 to obtain a variable with a range of 0 to 10; IHC4, four immunohistochemical markers (ER, PgR, HER2, and Ki-67); LR, likelihood ratio; PgR, progesterone receptor.
*Linear logit (LR trend test).
†Wilcoxon.

Table A2. Comparison of Clinical and Immunohistochemical Characteristics of Patients in the Nottingham and TransATAC Data Sets

Characteristic	Nottingham (N = 786)		TransATAC (N = 1,125)	
	No.	%	No.	%
Treatment				
No endocrine treatment	410	52	0	
Tamoxifen	376	48	565	50
Anastrozole	0		560	50
Age, years				
Median	55		64	
IQR	48-63		57-70	
Tumor size, cm				
≤ 1	105	13	177	16
1-2	415	53	574	51
2-3	190	24	272	24
> 3	76	10	102	9
Tumor grade				
Poor	178	23	206	18
Moderate	336	43	690	61
Well differentiated	272	34	229	21
Unknown	0		49	4
Nodal status				
Negative	487	62	793	70
Positive	299	38	288	26
Not known	0		44	4
HER2				
Positive	41	5	116	10
ER H-score				
Median %	141		165	
IQR	99-201		135-186	
PgR				
Median %	50		62	
IQR	12-77		22-92	
Ki-67*				
Median %	11.0		4.2	
IQR	5.0-35		1.8-8.9	
Median time to distant recurrence, months	9.5		7.7	
Distant recurrence (within 10 years)	174	22	145	13

Abbreviations: ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; H-score, defined as the percentage of cells staining weakly plus two times the percentage of cells staining moderately plus three times the percentage of cells staining strongly divided by 30 to obtain a variable with a range of 0 to 10; IQR, interquartile range; PgR, progesterone receptor; TransATAC, the translational science substudy within the Arimidex, Tamoxifen, Alone or in Combination study.
*Different antibodies used (SP6 for TransATAC *v* MiB1 for Nottingham).