Background: Three gene expression–based prognostic breast cancer tests have been licensed for use.

Purpose: To summarize evidence on the validity and utility of 3 gene expression–based prognostic breast cancer tests: Oncotype DX (Genomic Health, Redwood City, California), MammaPrint (Agendia BV, Amsterdam, the Netherlands), and H/I (AvariaDX, Carlsbad, California).

Data Sources: MEDLINE, EMBASE, and Cochrane databases (from 1990 through January 2007), Web sites of test manufacturers, and discussion with the manufacturers.

Study Selection: Original data studies published in English that addressed prognostic accuracy and discrimination or treatment benefit prediction of any of the 3 tests in women with breast cancer.

Data Extraction: Information was extracted about the clinical characteristics of the study population (particularly clinical and therapeutic homogeneity), tumor characteristics, and whether the marketed test or underlying signature was evaluated.

Data Synthesis: The tests are based on distinct gene lists, using 2 different technologies. Overall, the body of evidence showed that this new generation of tests may improve prognostic and therapeutic prediction, but the tests are at different stages of development. Evidence shows that the tests offer clinically relevant, improved risk stratification over standard predictors. Oncotype DX has the strongest evidence, closely followed by MammaPrint and H/I (which is still maturing).

Limitations: For all tests, the relationship of predicted to observed risk in different populations and their incremental contribution over conventional predictors, optimal implementation, and relevance to patients receiving current therapies need further study.

Conclusion: Gene expression technologies show great promise to improve predictions of prognosis and treatment benefit for women with early-stage breast cancer. More information is needed on the extent of improvement in prediction, characteristics of women in whom the tests should be used, and how best to incorporate test results into decision making about breast cancer treatment.

Currently, 3 commercially available prognostic breast cancer tests based on gene expression (see Glossary) technology are available: Oncotype DX (Genomic Health, Redwood City, California), MammaPrint (Agendia BV, Amsterdam, the Netherlands), and H/I (AvariaDX, Carlsbad, California). Although measurement of gene expression is now a core research method, these commercial assays represent the first introduction of these technologies into clinical application.

Gene expression is the technical term to describe how active a particular gene is—that is, how many times it is expressed, or transcribed, to produce the protein it encodes (Figure 1). The transcription (see Glossary) of the gene’s DNA into messenger RNA (mRNA) is the first step in this process; modern molecular biological tools measure this activity by counting the number of mRNA molecules in a given cell type or tissue. Because the mRNA molecule is translated within the ribosome to produce a complete protein, counting mRNA transcripts provides an estimate of the number of corresponding proteins. High-throughput technologies, such as DNA microarray (see Glossary) and real-time reverse transcriptase polymerase chain reaction (RT-PCR) (see Glossary), allow simultaneous counting of many gene transcriptions (up to tens of thousands). This creates a snapshot of a tissue’s global gene activity, called the transcriptome.

Gene expression measurements have been used to develop new biological concepts, refine disease classification, improve diagnostic and prognostic accuracy, and identify new molecular targets for drugs, especially in cancer research (1–9). Results are commonly reported in the form of a list of genes that are differentially expressed between normal and diseased patients or that correlate with different prognoses or phenotypes. These lists are called gene expression profiles or signatures (see Glossary).

“Breast cancer” is increasingly understood as an umbrella designation for various tumor subtypes that differ in their prognoses and responses to therapy. An important decision for many patients with early-stage breast cancer, especially patients who have tumors that express hormone receptors and will be given antiestrogen therapy, is whether they should also be treated with systemic chemotherapy. Although adjuvant chemotherapy is frequently recommended in this setting, many women will remain recurrence-
A molecular biology technique that allows amplification and quantification in real time of defined RNA molecules from specific specimens. This technology has been used for several years in research and clinical settings. In brief, in the first step, DNA copies of the investigated RNA molecules are obtained by a process called reverse transcription, and DNA amplification is then obtained by using polymerase chain reaction. The quantification of the accumulating DNA product is accomplished by the use of specific fluorescent reagents. In this technique, the quantification of the target RNA molecule is based on the analysis of the accumulation curve of the complementary DNA, as measured by the fluorescence detected at each cycle of the reaction. In biochemistry, the enzymatic reaction carried on by the RNA-dependent DNA polymerase. This enzyme, known as reverse transcriptase, is a DNA polymerase enzyme that copies single-stranded RNA into DNA. This process is the reverse of normal transcription, which involves the synthesis of RNA from DNA.

The process by which DNA sequences are copied into complementary RNA molecules by the enzyme RNA polymerase. This reaction represents the transfer of genetic information from DNA into RNA. The RNA sequence that is transcribed from a DNA molecule is called a transcript.

free at 10 years without it, especially those with small, estrogen receptor–positive tumors without axillary nodal involvement. Patients and their physicians must weigh the possible benefit of chemotherapy in reducing recurrence against its toxicity and other attendant costs.

Practicing oncologists frequently base their decisions about therapy on prognostic clinical algorithms that include demographic data; tumor stage; and other tumor characteristics, such as grade and estrogen receptor expression. These conventional combination predictors include the National Institutes of Health (NIH) Consensus Development criteria (10, 11); the St. Gallen expert opinion criteria (12, 13); the National Comprehensive Cancer Network guideline (14–16); and a Web-based algorithm, Adjuvant! Online (17, 18). Gene expression profiling has been proposed to potentially augment or replace these prognostic tools.

Oncotype DX is based on a 21-gene profile developed by Paik and colleagues (19). MammaPrint is based on a 70-gene prognostic signature developed by van’t Veer and colleagues (8), and H/I is based on a 2-gene signature (HOXB13–IL17BR) developed by Ma and colleagues (20). The gene sets on which these tests are based have minimal overlap. The 21-gene and the 70-gene expression signatures that form the basis of Oncotype DX and MammaPrints, respectively, share only 1 gene in common. Two technologies are used to determine gene expression: real-time RT-PCR (Oncotype DX and H/I) and DNA microarray (MammaPrint). All 3 tests use pathologic review of specimens to check tumor content and evaluate RNA preparation and quality. The 2 RT-PCR based assays (Oncotype DX and H/I) are done in formalin-fixed, paraffin-embedded tumor tissues, whereas fresh unfixed tumor tissue is required for MammaPrint. We review evidence on the prognostic accuracy of these 3 tests and their ability to predict treatment benefit.

**METHODS**

The Agency for Healthcare Research and Quality commissioned the review for the Centers of Disease Control and Prevention’s Evaluation of Genomic Applications in Practice and Prevention program. Additional details about the methods and results are found in a comprehensive evidence report that is available through the Agency for Healthcare Research and Quality (www.ahrq.gov/clinic/epicindex.htm).
Data Sources

On 9 January 2007, we searched the MEDLINE and EMBASE databases by using Medical Subject Headings and other terms relevant to breast cancer, gene expression profiling, and Oncotype DX or MammaPrint. On 7 February 2007, we searched the Cochrane database, including Cochrane Reviews, CENTRAL, and CINAHL. We supplemented this search by updating searches in MEDLINE and by hand-searching added publications that appeared after the initial search (January 2007 to July 2007).
2007) and studies related to H/I. The test manufacturers were also asked to provide any published or unpublished data relating to our study questions. Searches were limited to publications in English.

Study Selection
Two investigators independently reviewed titles and abstracts to identify original data studies that involved the use of any of the 3 assays in women with breast cancer.

Data Extraction
We extracted and double-checked information on the clinical characteristics of the study population, tumor characteristics, and whether the marketed test or underlying signature was evaluated. To assess the quality of studies, we applied (where appropriate) the general principles of the REMARK (REporting recommendations for tumor MARKer prognostic studies) (21, 22) and Standards for Reporting of Diagnostic Accuracy (23, 24) guidelines.

We synthesized data on the ability of a test to accurately predict recurrence risk (clinical validity) and treatment benefit (clinical utility). We distinguished between gene expression signatures and the gene expression–based marketed tests. The gene signature is the collection of genes whose expression levels are measured in a given test. A gene signature can be measured by using various technologies (RT-PCR or complementary DNA [cDNA] array) and procedures (for example, different reagents, controls, sample acquisition, preparation, and transport procedures), which may not be identical to those used in the marketed test.

We report limited information on the technical performance characteristics of the tests, sometimes called analytic validity. The analytic validity of a test usually is assessed by determining how observed measurements differ from standard reference values. However, no reference standard exists for gene expression measurements outside of the technologies used for these tests. Because analytic validity affects predictive ability, our assessment of predictive ability incorporates the effect of less-than-perfect analytic validity. In the full evidence report (available at www.ahrq.gov/clinic/epcindex.htm), we summarize data on the reproducibility of a test when applied repeatedly to the same patient or when repeated over time, as well as variability as a function of tumor sampling and handling, specimen preparation, and biological variation within tumor samples. Appendix Tables 1, 2, and 3 (available at www.annals.org) show the evidence summary.

Role of the Funding Source
The Agency for Healthcare Research and Quality and the Centers of Disease Control and Prevention’s Evaluation of Genomic Applications in Practice and Prevention program helped formulate the initial study questions but did not participate in the literature search, determination of study eligibility criteria, data analysis, or interpretation.

RESULTS
Figure 2 shows the number of studies considered at each phase of title, abstract, and article review. The final set of 26 studies was heterogeneous in focus and quality. Few reports addressed technical aspects of the tests. Ten reports focused on prognostic prediction. Only 1 study, involving Oncotype DX, examined the prediction of treatment benefit. Most of the published evidence available for Oncotype DX was conducted with the marketed assay. The evidence relevant to MammaPrint was a mix of studies of the underlying signature and of the marketed test. Only 1 study used the marketed version of H/I (41), and it was not clear whether the laboratory doing the assay was the same as the one with current rights to do the test. All other studies relevant to H/I examined the underlying 2-gene signature, using somewhat different measurement tech-

![Figure 2. Systematic search strategy and results.](https://annals.org)
niques and algorithms than those implemented in the marketed test.

Analytic Validity and Variability

We found limited evidence about the laboratory procedures used for Oncotype DX (26, 27) and MammaPrint (34, 35), including information about their reproducibility. Such evidence was reported in methodological studies (26–35) and in clinical studies that focused on predictive validity (19, 30). In the case of MammaPrint, results from different laboratories depended on RNA labeling protocols (34), suggesting that MammaPrint results may not be identical if done in different laboratories (only 1 laboratory currently offers the test). The overall proportion of samples that were successfully tested with the various methods ranged from 67.7% to 98.9% for Oncotype DX and 67.7% to 80.9% for MammaPrint (36) (Appendix Tables 1 and 2, available at www.annals.org). No reports investigated the reproducibility of H/I; Ma and colleagues (41) reported a success rate of 98%.

Predicting Disease Outcomes

Oncotype DX

Oncotype DX was developed on the basis of a prospectively chosen 250-candidate gene set, which was measured on 447 patients with breast cancer who were treated in 1 of 3 completed randomized trials with long-term follow-up. From these 250 genes, 21 genes (16 cancer-related and 5 references) were chosen to predict 10-year breast cancer recurrence. The expression levels of these genes are measured by using RT-PCR combined with a published quantitative algorithm to produce a number between 0 and 100, which is the recurrence score. In this review, “recurrence score” indicates the numeric value generated from Oncotype DX (19). The recurrence score is categorized into 3 risk strata: low (score <18), intermediate (score >18 but <30), or high (score ≥30).

Four studies assessed the clinical validity of Oncotype DX (19, 30, 31, 25). Paik and colleagues (19) studied 668 women in a randomized, controlled trial conducted by the National Surgical Adjuvant Breast and Bowel Project (NSABP) (Appendix Table 1, available at www.annals.org). The parent study (the NSABP B-14 trial), which enrolled patients from 1982 to 1988, examined the effect of tamoxifen therapy versus placebo in women with lymph node–negative, estrogen receptor–positive, early-stage breast cancer (42–45). The recurrence score algorithm and risk categories for the group treated with tamoxifen were pre-specified on the basis of the development studies. Stratification into the 3 risk categories yielded univariate actuarial 10-year recurrence risks of 7% (low), 14% (intermediate), and 31% (high) (P < 0.001) (Appendix Table 1, available at www.annals.org). The recurrence score was the strongest predictor among all traditional risk factors, with an adjusted hazard ratio (HR) of 2.8 (CI, 1.7 to 4.6) for a 50-point change in the recurrence score.

Glas and colleagues (35) examined the clinical perfor-

mance of the Oncotype DX assay to predict breast cancer death at 10 years in a community-based population of lymph node–negative, estrogen receptor–positive patients treated with tamoxifen. Two hundred twenty case patients (dead) and 570 matched control patients (alive) were selected, and 165 estrogen receptor–positive case patients and 55 case patients who received tamoxifen treatment (among estrogen receptor–positive patients) formed the final study sample. For patients treated with tamoxifen, results paralleled those of the NSABP B-14 trial (which examined recurrence): The probability of death at 10 years was 2.8%, 11%, and 16% in the low-, medium-, and high-risk groups, respectively. These rates were about 3 percentage points lower than those among patients not treated with tamoxifen. Prognostic value persisted after stratification by tumor grade and disease stage. The continuous recurrence score also showed a relationship with mortality risk in 52 estrogen receptor–negative patients after adjustment for tumor grade and disease stage (relative risk [RR], 1.4 per 10-unit increase in recurrence score [CI, 1.04 to 2.0]). Another study (36) showed no predictive value of the recurrence score in a small population of patients who received neither tamoxifen nor chemotherapy. However, worse tumor grade predicted better prognosis in that study, suggesting that the results were not reliable. A study on the signature alone, in which the main purpose was to contrast the different tests (33), is described in the “Comparison of Signatures” section.

No published study showed how the recurrence score reclassified patients into different risk strata after initial classification by conventional predictors. However, in 2004, Paik and colleagues (46) presented such information in poster form. They reported that the recurrence score had predictive power beyond that of the St. Gallen or National Comprehensive Cancer Network risk stratification guidelines, sufficient to change some patient decisions about chemotherapy. (The St. Gallen test did not include human epidermal growth factor receptor 2 at that time, which is included in the recurrence score). On the basis of the 2004 National Comprehensive Cancer Network guidelines, the study indicated that about half of the 92% of patients who were in the high-risk National Comprehensive Cancer Network category were reclassified as low-risk by the recurrence score, with a 10-year relapse risk of 7% (CI, 4% to 11%); this is similar to the risk seen in the low-risk recurrence score group without the National Comprehensive Cancer Network information. The same information for Adjuvant! Online was part of an oral presentation in 2005 (Table 1) (47). Compared with the Adjuvant! Online criteria, roughly 40% of women assessed to be at high risk (22% relapse) were reclassified into a group with an observed 8% risk if they had a low recurrence score. With Adjuvant! Online, 39% of women classified as high risk (31% recurrence) had a 9% recurrence risk after a low recurrence score. These findings showed that the greatest contribution of the test is probably the reclassification of
patients from high to low risk (that is, reducing the number of patients who might unnecessarily undergo chemotherapy) and that combining this test with conventional predictors yields the most information.

**MammaPrint**

MammaPrint is based on the 70-gene signature derived from an initially unselected set of more than 25,000 candidate genes on a cDNA array. The test was developed in 2002 at the Netherlands Cancer Institute by using 78 lymph node–negative patients younger than age 55 years who did not carry a breast cancer gene mutation and who had tumors that were less than 5 cm in diameter (8). The end point for this training set was 5-year distant recurrence. Patients are classified by calculating the correlation coefficient between a patient’s expression levels of the 70 genes and an average good-prognosis expression profile. If the correlation coefficient exceeds 0.4, the patient is classified as having a good prognosis; if less, they are classified as having a poor prognosis.

van de Vijver and coworkers (9) validated this signature in a series of 295 consecutive patients with stage I or II breast cancer and small tumors (<5 cm) who were younger than age 53 years. The population was mixed in terms of lymph node positivity, estrogen receptor status, and receipt of chemotherapy and tamoxifen. Sixty-one of the 295 patients in the validation study were also used to develop the signature. Patients with a good prognosis had dramatically better 5-year (95% vs. 61%) and 10-year (85% vs. 51%) recurrence-free survival and overall survival (95% vs. 55% at 10 years) than patients with a poor prognosis. Multivariable analysis showed that prognosis group, tumor size, and adjuvant chemotherapy were the strongest predictors of distant metastases, and patients with the poor-prognosis signature had the largest HR (4.6 [CI, 2.3 to 9.2]). Results of analyses excluding the 61 patients from the training cohort were similar. Fifteen percent of patients with the good-prognosis signature had recurrence by 10 years, demonstrating that when the 70-gene signature is used alone in this mixed population, the long-term risk in the good-prognosis group may not be low enough to justify forgoing chemotherapy.

Kaplan–Meier analyses showed the absolute risks associated with various predictors and combinations of predictors. Overall, the 70-gene signature placed 40% of the cohort (60 of 151) into the good-prognosis group, with a 10-year recurrence rate of about 15% (imputed from figures). The St. Gallen index placed only 15% (22 of 151) of the cohort into a low-risk group, with an estimated 10-year recurrence rate slightly greater than 20%, and the NIH criteria placed only 7% (11 of 151 patients) at low risk, with a long-term risk (based on small numbers) of slightly less than 20%. The 70-gene signature reclassified 33% (43 of 129) of St. Gallen high-risk patients and 38% (53 of

---

**Table 1. Patient Reclassification by Gene Expression Testing with Oncotype DX**

<table>
<thead>
<tr>
<th>Comparator Risk Group</th>
<th>Patients in Risk Group, n (%)</th>
<th>10-Year Risk for Distant Relapse, %</th>
<th>Oncotype DX</th>
<th>Risk Group</th>
<th>Patients in Risk Group, n (%)</th>
<th>10-Year Risk for Distant Relapse, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low</td>
<td>38 (72)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Medium</td>
<td>12 (22)</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>High</td>
<td>3 (6)</td>
<td>43</td>
</tr>
<tr>
<td>St. Gallen expert criteria</td>
<td></td>
<td></td>
<td></td>
<td>Low</td>
<td>134 (60)</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Medium</td>
<td>51 (23)</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>High</td>
<td>37 (17)</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>53 (8)</td>
<td>5</td>
<td></td>
<td>Low</td>
<td>166 (42)</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>222 (33)</td>
<td>9</td>
<td></td>
<td>Medium</td>
<td>86 (22)</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>393 (59)</td>
<td>18</td>
<td></td>
<td>High</td>
<td>141 (36)</td>
<td>33</td>
</tr>
<tr>
<td>2004 National Comprehensive Cancer Network guidelines</td>
<td></td>
<td></td>
<td></td>
<td>Low</td>
<td>38 (72)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Medium</td>
<td>12 (22)</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>High</td>
<td>3 (6)</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>53 (8)</td>
<td>5</td>
<td></td>
<td>Low</td>
<td>300 (49)</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>615 (92)</td>
<td>15</td>
<td></td>
<td>Medium</td>
<td>137 (22)</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>High</td>
<td>178 (29)</td>
<td>30</td>
</tr>
<tr>
<td>Adjuvant! Online criteria</td>
<td></td>
<td></td>
<td></td>
<td>Low</td>
<td>216 (61)</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Medium to high</td>
<td>138 (39)</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>354 (53)</td>
<td>8</td>
<td></td>
<td>Low</td>
<td>122 (39)</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>314 (47)</td>
<td>22</td>
<td></td>
<td>Medium to high</td>
<td>192 (61)</td>
<td>31</td>
</tr>
</tbody>
</table>

---
of NIH criteria high-risk patients into a lower-risk group, with a 10-year recurrence risk just a few percentage points greater than that of the 70-gene good-signature group without stratification.

Glas and colleagues (35) reanalyzed 145 patients from van de Vijver and coworkers’ (9) cohort and all 78 patients from the training set by using the marketed MammaPrint assay instead of the signature (35). A different reference RNA and different quantification method were used. Although odds ratios (ORs) and HRs were similar to those found in the earlier studies, approximately 9% (7 of 78) of patients were placed into different risk categories, most of which had borderline correlations.

The MammaPrint test was validated in a multicenter European study of 302 patients not treated with chemotherapy or tamoxifen, and it provided prognostic information beyond that of Adjuvant! Online (36). Frozen tumor specimens from node-negative patients younger than age 60 years who did not receive systemic adjuvant chemotherapy were tested. Ninety patients were estrogen receptor–negative, and none of the estrogen receptor–positive patients received tamoxifen. The median follow-up was 13.6 years, and the overall rate of distant metastasis was 25%. The area under the receiver-operating characteristic curves indicated that both methods had similarly modest discriminatory power in absolute terms (0.68 for MammaPrint and 0.66 for Adjuvant! Online), but MammaPrint provided better reclassification of patients in risk groups (Table 2). Hazard ratio estimates between high- and low-risk categories for distant recurrence in van de Vijver and colleagues’ study were substantially higher than those in this validation study (unadjusted HR >15 vs. 2.3, respectively; adjusted HR, 4.6 vs. 2.1). Compared with van de Vijver and colleagues’ (9) study, this validation cohort was observed for a longer period (median, 13.6 vs. 6.7 years), included older women, and excluded patients who received adjuvant therapy. This study also found that HRs for all end points decreased steadily with an artificial increase in censoring time from 5 to 10 years.

The H/I Test

Ma and colleagues (20) identified the 2 genes that are the basis for H/I by screening 22,000 genes in 60 patients with estrogen receptor–positive, lymph node–positive or negative breast cancer treated with tamoxifen. High expression of HOXB13 predicted recurrence, and high expression of IL17BR predicted nonrecurrence; therefore, a higher ratio of the 2 genes strongly predicted recurrence in this training set (interquartile OR, 10.2; adjusted OR, 10.2).

Reid and colleagues (37) examined 58 tamoxifen-treated patients with estrogen receptor–positive breast cancer whose disease was more advanced than in Ma and colleagues’ sample (48). No relationship between the expression of these genes and distant relapse was observed in these patients or in an additional 99 patients derived from a previously studied cohort (5) that the authors investigated after the initial negative result. In 2006, Goetz and colleagues (38) analyzed 206 estrogen receptor–positive patients treated in the tamoxifen-only arm of a phase III randomized trial. Expression values were normalized by using a different approach than that used by Ma and colleagues (20), and different cutoff points were calculated for the ratio that best predicted relapse-free survival, disease-free survival, and overall survival. The ratio had modest predictive strength in the entire cohort, with cross-validated HRs near 1.5 and P values around 0.05, and the predictive ability was restricted to node-negative patients.

In a large validation study, Ma and colleagues (41) examined a consecutive series of 852 patients with stage I or II breast cancer with a median follow-up of 6.8 years. The investigators used a slightly different method from the one they previously used (20) to combine and normalize the expression of the 2 genes into an index that is now the basis of the H/I assay. In a stratified analysis, the HOXB13–IL17BR ratio was predictive only in patients with node–negative, estrogen receptor–positive disease. The investigators optimized the threshold (maximizing the HR) differently in patients treated with tamoxifen and those who were not. The adjusted HR incorporating other risk factors was 3.9, regardless of tamoxifen treatment. Classification probabilities were not presented, and the incremental value of the HOXB13–IL17BR ratio compared with conventional combined predictors was not reported, although some components of those predictors were included in the multivariable analyses.

Jansen and associates (39) evaluated the ability of the HOXB13–IL17BR ratio to predict disease-free survival in 1252 patients with breast cancer who had undergone various surgical treatments. In this group, 73% of tumors were estrogen receptor–positive; 52% of patients were lymph node–positive, 14% were treated with tamoxifen, 17% received chemotherapy, and 55% received tamoxifen or chemotherapy after relapse. Jansen and associates (39) used different populations, protocols, normalization strategy, and ratio thresholds than Ma and colleagues (41). The overall relapse rate was high at 51% after a median of 6

### Table 2. Kaplan–Meier Analysis of Survival Stratified by MammaPrint and Adjuvant! Online

<table>
<thead>
<tr>
<th>MammaPrint Prognosis Group</th>
<th>Adjuvant! Online Risk Group</th>
<th>Low Risk</th>
<th>High Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Good</strong></td>
<td></td>
<td>52</td>
<td>59</td>
</tr>
<tr>
<td>Patients, n</td>
<td></td>
<td>0.88 (0.74–0.95)</td>
<td>0.89 (0.77–0.95)</td>
</tr>
<tr>
<td>10-year overall survival</td>
<td></td>
<td>(95% CI)</td>
<td>(95% CI)</td>
</tr>
<tr>
<td><strong>Poor</strong></td>
<td></td>
<td>28</td>
<td>163</td>
</tr>
<tr>
<td>Patients, n</td>
<td></td>
<td>0.69 (0.45–0.84)</td>
<td>0.69 (0.61–0.76)</td>
</tr>
<tr>
<td>10-year overall survival</td>
<td></td>
<td>(95% CI)</td>
<td>(95% CI)</td>
</tr>
</tbody>
</table>
years follow-up. The HOXB13–IL17BR ratio was examined in 468 patients with lymph node–negative, estrogen receptor–positive disease who did not receive adjuvant systemic chemotherapy. The ratio was associated with poor disease-free survival in a multivariable model (HR, 1.6; P = 0.02) and poor overall survival (HR not reported; P < 0.001). Prognostic value was also shown for untreated patients with estrogen receptor–positive, lymph node–positive tumors and for progression-free survival in patients with relapse who received first-line tamoxifen monotherapy (Appendix Table 3, available at www.annals.org). The ratio was not compared with conventional combination risk indices, and classification probabilities for the models with and without the ratio were not provided.

Jerevall and coworkers (40) investigated whether the HOXB13–IL17BR ratio predicted a differential benefit between 264 patients with postmenopausal breast cancer who received tamoxifen for 2 and 5 years and 93 premenopausal patients who did not receive systemic therapy. Seventy-two percent of patients had lymph node–positive disease and 74% had estrogen receptor–positive disease (Appendix Table 3, available at www.annals.org). The authors dichotomized the HOXB13–IL17BR ratio at the median, which differed from the approach used by Ma and colleagues (41). Jerevall and coworkers (40) concluded that IL17BR might be an independent prognostic factor in breast cancer and suggested that HOXB13 may be correlated with tamoxifen resistance. However, the HOXB13–IL17BR ratio had no prognostic value in postmenopausal patients with estrogen receptor–negative disease. Neither the patient profile nor the methods of calculation of the ratio were identical to those used in previous studies, and the results differed from previous reports because the HOXB13–IL17BR ratio predicted worse outcomes in patients with lymph node–positive disease.

Comparison of Signatures

Fan and colleagues (33) used the same data set to evaluate both the agreement between gene expression tests and other predictors and the individual performance of the tests. The Oncotype DX recurrence score and the HOXB13–IL17BR ratio were estimated from microarray gene expression data (that is, not RT-PCR) and thus were not obtained according to the protocols and methods used in the marketed assays. These data are therefore described as derived scores. Fan and colleagues (33) used the same 295 samples from patients with stage I or II breast cancer that had been used to develop the 70-gene signature (9). Therefore, Fan and colleagues’ (33) comparison would be expected to favor the 70-gene profile over the derived recurrence score or the HOXB13–IL17BR ratio.

The 70-gene signature and the derived recurrence score predicted overall survival and disease-free survival, but the derived HOXB13–IL17BR ratio did not predict either (HR, about 1); however, measurement of the derived 2-gene ratio may have been flawed (49). The inter-

Predicting Treatment Response

The ability of Oncotype DX to predict chemotherapy benefit was investigated in patients from the NSABP B-20 trial (50). In this study, Paik and colleagues (50) examined 10-year, distant recurrence-free survival in 651 patients with estrogen receptor–positive, lymph node–negative disease who were randomly assigned to receive tamoxifen alone or tamoxifen with chemotherapy. An overall benefit was seen from chemotherapy, but when the data were stratified by risk group, the benefit was restricted to patients with a high recurrence score (RR, 0.26 [CI, 0.13 to 0.53]), a finding that persisted in multivariate analyses. However, even though no benefit was seen in the low recurrence-score group, the point estimate had very wide CIs; a clinically relevant benefit could therefore not be excluded.

Two studies examined whether the recurrence score predicted pathologic response in patients receiving preoperative systemic therapy (28, 29). Neither study was done at the laboratory offering Oncotype DX (Appendix Table 1, available at www.annals.org.) One study found that the recurrence score predicted complete response (28), whereas the other study (29) found no such relationship. Finally, Chang and colleagues (32) assessed chemotherapy response prediction in 12 patients with complete clinical response among 72 women enrolled in phase II studies of docetaxel and found that a high recurrence score was associated with complete response (P = 0.008). When the recurrence score was used as a continuous variable, a 14-unit increase in the score (the difference between the low- and high-risk groups, as defined by the standard thresholds) were modestly predictive of a clinical complete response (OR, 1.7 [CI, 1.15 to 2.60]).

No study investigated the ability of MammaPrint to predict treatment response. One study reported that the HOXB13–IL17BR ratio could predict whether 5 years of tamoxifen therapy would provide survival benefit over 2 years of tamoxifen treatment in estrogen receptor–positive patients (40).

Discussion

This body of evidence on the 3 marketed gene expression tests for breast cancer prognosis shows that these tests have considerable potential for improving prognostic and
therapeutic prediction. It also provides valuable lessons about the complexity of evaluating such tests.

Because the role of the genes included in these tests in determining prognosis is not completely understood, it is often unclear which clinical or tumor characteristics are being measured. Intrinsic tumor aggressiveness, ability to metastasize, and responsiveness to treatment (hormonal, radiation, or chemotherapy)—each of which might involve different genes—can determine prognosis. However, the characteristic being assessed in a particular study must often be inferred from the treatment, tumor, and clinical characteristics of the study population. Results from populations that are clinically and therapeutically heterogeneous may not be optimal in determining the prognosis or risk for a particular woman.

End points also varied. Survival was defined in the studies as disease-free, distant recurrence–free, or overall, measured at 5 or 10 (or more) years. Prediction strength varied considerably depending on what end point the test was optimized for. Finally, performance of the underlying gene signature is not necessarily identical to the marketed test, because many test procedures, including pretest sample preparation and transport, can differ. It is therefore critical to pay close attention to the test description, population, and end points for each study to understand which studies are mutually supportive, which are adding qualitatively different information, and to whom they apply.

Despite the clinical novelty of these tests, their development must follow the same principles and procedures as those for any multivariate clinical prediction rule. These principles are outlined in detail in the clinical literature (51, 52) and have been agreed on in reporting guidelines (22) and articulated with respect to expression-based predictors in various review articles (53–55).

Each of the 3 marketed tests is at a different point in the developmental pathway. Almost all studies of Oncotype DX have used the marketed test as opposed to the signature, whereas the evidence on MammaPrint comes from studies examining the signature or the assay (only the large multicenter validation by Buyse and colleagues [36] used the marketed assay). The study that compared the results of the marketed MammaPrint test versus its signature on the same samples showed that about 9% of the patients were placed into different risk groups when the marketed test was used. Almost all of the studies based on H/I calculated or implemented the test in subtly different ways, and only 1 seemed to use the marketed H/I test (48).

In terms of the clinical and therapeutic homogeneity of the underlying populations, Oncotype DX focused on a narrower, more clinically and therapeutically homogeneous group than the other tests, which is reflected in its claimed indications: patients with estrogen receptor–positive, lymph node–negative, stage I or II disease who receive tamoxifen. MammaPrint has been tested in heterogeneous populations including a mix of treated and untreated patients, patients with lymph node–negative and —positive disease, and those with estrogen receptor–positive and —negative disease. The claimed indications are therefore broader than those for Oncotype DX. Although the indications for MammaPrint match the populations in whom the test has been evaluated, whether to consider these populations prognostically homogenous is a critical question.

The gene signature underlying H/I has been investigated in large, heterogeneous populations, and differences were found in its prediction ability for specific subgroups. The signature has been variably formulated as a simple ratio or as an index, normalized to different sets of genes or standardized with calibration RNA, and stratified by using thresholds optimized within each study. Whereas plausible mechanisms support the test rationale, the marketed test requires further validation and comparison with conventional combination predictors. The eligibility criteria for H/I are similar to those for Oncotype DX.

The utility of these tests for clinical decision making is a critical question. No study has addressed whether MammaPrint or H/I can predict the clinical benefit of chemotherapy. Oncotype DX is the only gene expression test that can predict such a benefit. Although this claim is based on past data, the study design was strong and was based on a randomized, clinical trial of tamoxifen plus chemotherapy versus tamoxifen alone (50). This study suggested that chemotherapy had no benefit in women in the low-risk group, but wide CIs around the observed zero effect did not rule out a meaningful effect. Even with few data on prediction of treatment benefit, the risk for long-term recurrence or death serves as an effective ceiling on the degree of chemotherapy benefit. If that risk is sufficiently low, some patients may forgo chemotherapy. Both the magnitude of the low-risk estimates and the proportion of patients who fall into those categories are therefore of considerable interest. Because various standard risk prediction tools are freely available, the question is how much the new tests add.

Only a few sources provide evidence on long-term absolute risks after conventional combination risk predictors are taken into account. For Oncotype DX, results of these analyses are published only in abstract form, although the findings are derived from the same NSABP B-14 cohort that provided main original validation evidence and are reported by the same authors (46, 47). This showed that Oncotype DX can reclassify patients in the highest-risk categories by conventional indices (18% for 2003 St. Gallen, 15% for 2004 National Comprehensive Cancer Network, and 22% for Adjuvant! Online) into clinically relevant lower risks (8%, 8%, and 9%, respectively), although the upper confidence limits on those new lower predictions all exceed 10%. For women at the lowest risk by conventional metrics, being placed in a low-risk stratum seems to lower the risk even further, information that patients might find useful; however, the number of patients in this group and on which this finding is based is small (approximately 30 to 60). For MammaPrint, the only reclassification data reported are those obtained in combina-
tion with Adjuvant! Online for a 10-year outcome (36). Adjuvant! Online had no predictive power for survival after the data were stratified by MammaPrint risk group, and it had only a very modest effect for 10-year distant recurrence. Similarly, findings were reported for the 70-gene signature in combination with the NIH and St. Gallen criteria (9). The risks in the good-signature MammaPrint groups were higher than in the Oncotype DX low-risk stratum, in part because the more heterogeneous validation population was at higher risk.

The exact values of the test results provide information that is lost when patients are assigned to risk categories, and the cutoffs for these categories may not correspond to optimal decision thresholds, particularly in combination with other predictors. How the results of such tests are conveyed to and understood by patients and physicians—for example, as absolute probabilities or as qualitative descriptors (“low risk”)—is critical as these tests become more widely used.

The ideal assessment of value of these tests would be to randomly assign patients to use them or not, as part their therapeutic decision making. The 2 ongoing prospective randomized trials, TAILORx (Trial Assigning Individualized Options for Treatment) (56) and MINDACT (Microarray In Node-negative Disease May Avoid Chemotherapy) (57), do not use such a design. The TAILORx compares disease-free survival among women with previously resected axillary node–negative breast cancer who had an Oncotype DX recurrence score between 11 and 25 and received adjuvant chemotherapy and hormonal therapy versus women who received hormonal therapy alone. All patients receive the test (56); those with a score of 10 or less do not receive chemotherapy, and all those with a score greater than 25 receive chemotherapy. These thresholds are lower than those conventionally used to designate high risk (score ≥30) and low risk (score <18).

The MINDACT study is a multicenter, prospective, phase III, randomized study directly comparing MammaPrint with Adjuvant! Online in selecting patients for adjuvant chemotherapy in node–negative breast cancer. Patients at low risk by both MammaPrint and Adjuvant! Online criteria do not receive chemotherapy; patients at high risk by both criteria receive chemotherapy; and patients with discordant criteria are randomly assigned to use either the MammaPrint or Adjuvant! Online results to determine treatment. This trial comes much closer to testing directly the clinical value of MammaPrint (versus Adjuvant! Online) than does TAILORx for Oncotype DX, although both studies will provide valuable evidence bearing on that question.

Our review identified important issues that may arise as these tests are applied in clinical practice, are modified, and as similar tests proliferate. These issues are described in Table 3. As these tests are modified and new ones are marketed, questions will arise about how the tests compare with one another and whether combining the tests has value. Answering these questions will require comparative effectiveness research. The U.S. government and industry often do not fund comparative effectiveness studies because such studies may not offer as much therapeutic promise as new discoveries and because industry is not eager to fund direct comparisons with competitive products. This same dynamic could take hold in the risk-prediction arena. Early in development, oversight of test development and research funding should encourage contrasts with existing expression-based predictors. Otherwise, new tests that all claim to offer similar guidance, or perhaps new guidance in previously neglected clinical subsets, will flood the market, and physicians and patients will have no way to evaluate the claims.

In conclusion, the introduction of gene expression tests has ushered in a new era in which many conventional clinical markers may be seen merely as surrogates for more fundamental genetic and physiologic processes that can be measured with these tests. The multidimensional nature of

### Table 3. Future Issues*

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evolving treatments and predictors</td>
<td>The current evidence is based on patients treated with past therapies (e.g., tamoxifen); thus, it is not known how these results would apply to patients treated with current therapeutic regimens (aromatase inhibitors, anti-HER2 drugs) and with different chemotherapy agents.</td>
</tr>
<tr>
<td>Technical performance</td>
<td>Specimen handling and preparation can substantially affect test quality because of the instability of RNA. Centralization of specimen processing is a strength, but additional scrutiny may be needed if other laboratories offer such testing. Wider utilization may tax reliability of central pathologic review and stretch other laboratory resources, requiring ongoing monitoring of predictive accuracy.</td>
</tr>
<tr>
<td>Genetic variability and gene expression</td>
<td>It is unknown whether gene expression profiles are more or less likely than traditional biomarkers to be generalizable across populations with varying genetic background. Gene expression patterns have also been associated with specific genetic mutations (i.e., BRCA1), indicating that specific DNA mutations or polymorphisms (8, 58) may affect the performance of a signature.</td>
</tr>
<tr>
<td>The need for databases, reproducibility, and standards</td>
<td>The MIAME standards (59) represent the basis for the proper collection and storage of microarray data and should be used going forward for the archiving of the tests done in real patients. Databases with complete data on each patient are needed (without identifiers), including all analyses and procedures used to produce a risk estimate from a tumor sample. Such databases could renew and expand the currently limited pool of validation databases.</td>
</tr>
<tr>
<td>Real-world implementation</td>
<td>It is important to understand what women are being told about test interpretation, whether absolute risks are conveyed, and what women's risk or benefit thresholds are for acceptance of chemotherapy.</td>
</tr>
<tr>
<td>Comparative effectiveness studies</td>
<td>Early in development, oversight of test development and research funding should encourage contrasts with existing expression-based predictors. Otherwise, new tests that all claim to offer similar guidance, or perhaps new guidance in previously neglected clinical subsets, will flood the market, and physicians and patients will have no way to evaluate the claims.</td>
</tr>
</tbody>
</table>

* HER2 = human epidermal growth factor receptor 2; MIAME = minimal information about a microarray experiment.
these predictors demands that large numbers of clinically homogeneous patients be used in the validation process and that exceptional rigor and discipline be applied in evaluation. Every study provides an opportunity to modify a genetic signature, but we must find the right balance between speed of innovation and development of reliable tools. It will be important to preserve genetic and clinical information from tested patients to facilitate further evaluation and innovation in current populations. Although these tests show great promise to improve predictions of prognosis and treatment benefit for women with early-stage breast cancer, more must be learned about the extent of that improvement, in whom it is most improved, and how the tests are best incorporated into decision making about current breast cancer treatment.

From Johns Hopkins University, School of Medicine, Baltimore, Maryland.

Disclaimer: The authors of this report are responsible for its content. Statements in the report should not be construed as endorsement by the Agency for Healthcare Research and Quality or the U.S. Department of Health and Human Services.

Grant Support: This project was funded under contract no. 290-02-0018 from the Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services.

Potential Financial Conflicts of Interest: None disclosed.

Requests for Single Reprints: Steven N. Goodman, MD, MHS, PhD, Johns Hopkins University School of Medicine, 550 Building, Room 11-03, Baltimore, MD 21205; e-mail, sgoodman@jhmi.edu.

Current author addresses are available at www.annals.org.

References


Current Author Addresses: Drs. Marchionni and Wolff: Johns Hopkins University School of Medicine, Oncology Cancer Biology, Baltimore, MD 21287. Ms. Wilson, Dr. Marinopoulos, and Dr. Bass: Johns Hopkins University School of Medicine, General Internal Medicine, Baltimore, MD 21287. Dr. Parmigiani: Johns Hopkins University, School of Medicine Oncology Informatics, Baltimore, MD 21287. Dr. Goodman: Johns Hopkins University School of Medicine, Oncology Biostatistics, Baltimore, MD 21287.
### Studies on the Oncotype DX Gene Expression Test

<table>
<thead>
<tr>
<th>Study, Year (Reference)</th>
<th>Assay development</th>
<th>Patients Analyzed/Eligible, n (%)</th>
<th>Population Characteristics</th>
<th>End Points</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colsher et al., 2003 (25)</td>
<td>Rush University Medical Center (Chicago, Illinois), 21-gene signature selection</td>
<td>76/80 (97.7)</td>
<td>Mean age, 57 ± 8; not reported; LN-negative, 100%; T1, 0–2 cm (31%); T2, &gt;2 cm (36%); T3, 36%; tamoxifen, 64%; adjacent CMF, 80%</td>
<td>Distinct disease-free survival at 10 y</td>
<td>21-gene signature selection</td>
</tr>
<tr>
<td>Pali et al., 2004 (19)</td>
<td>Providence St. Joseph Hospital (Burbank, California), 21-gene signature selection</td>
<td>36</td>
<td>ER-positive, proportion not specified; LN-negative, proportion not specified; tamoxifen, 31%; chemotherapy, 39%</td>
<td>Distinct disease-free survival at 10 y</td>
<td>21-gene signature selection</td>
</tr>
</tbody>
</table>

### Assay validation

<table>
<thead>
<tr>
<th>Study, Year (Reference)</th>
<th>Assay development</th>
<th>Patients Analyzed/Eligible, n (%)</th>
<th>Population Characteristics</th>
<th>End Points</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pali et al., 2004 (19)</td>
<td>NSABP B-20, randomized, controlled trial</td>
<td>233</td>
<td>ER-positive, 100%; LN-negative, 100%; tamoxifen, 100%; chemotherapy, 0%</td>
<td>Distinct disease-free survival at 10 y</td>
<td>21-gene signature selection; recurrence score algorithm and risk groups</td>
</tr>
</tbody>
</table>

### Other studies

<table>
<thead>
<tr>
<th>Study, Year (Reference)</th>
<th>Assay development</th>
<th>Patients Analyzed/Eligible, n (%)</th>
<th>Population Characteristics</th>
<th>End Points</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gianni et al., 2005 (28)</td>
<td>Neoadjuvant doxorubicin–paclitaxel, followed by chemotherapy, randomized control trial</td>
<td>45/57 (80.9)</td>
<td>Mean age, 57 ± 8; not reported; LN-negative, 100%; T1, 0–2 cm (31%); T2, &gt;2 cm (36%); T3, 36%; tamoxifen, 64%; adjacent CMF, 80%</td>
<td>Distinct disease-free survival at 10 y</td>
<td>21-gene signature selection; recurrence score algorithm and risk groups</td>
</tr>
</tbody>
</table>

### Appendix Table 1

#### Studies on the Oncotype DX Gene Expression Test

<table>
<thead>
<tr>
<th>Assay development</th>
<th>Patients Analyzed/Eligible, n (%)</th>
<th>Population Characteristics</th>
<th>End Points</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assay validation</td>
<td>Pali et al., 2004 (19)</td>
<td>NSABP B-14, randomized, controlled trial</td>
<td>668/675 (98.9)</td>
<td>Age: &lt;50 y (29%); 50–60 y (24%); &gt;60 y (47%); LN-negative, 100%; T2, &gt;2 cm (36%); tamoxifen, 64%; chemotherapy, 0%</td>
</tr>
<tr>
<td>Other studies</td>
<td>Chang et al., 2007 (12)</td>
<td>To evaluate whether RS can predict clinical response to adjuvant therapy</td>
<td>80/97 (82.4)</td>
<td>Mean age, 48.5 ± 8; ER-positive, 60.1%; LN-negative, 90.2%; HBR2-positive, 8.23%; T1, 0–2 cm (34%); T2, &gt;2 cm (22%); chemotherapy, 65%; AUC, 0.73</td>
</tr>
<tr>
<td>Other studies</td>
<td>Gambi et al., 2005 (28)</td>
<td>Neoadjuvant docetaxel-plus–trastuzumab, followed by chemotherapy, randomized control trial</td>
<td>89/95 (93.7)</td>
<td>Mean age, 49.9 ± 8; ER-positive, 56%; LN-negative, 16%; stage T1, 1%; T2, 5%, 79%; T3, 18%; T4, 24%; chemotherapy, 80%</td>
</tr>
<tr>
<td>Other studies</td>
<td>Minas et al., 2007 (29)</td>
<td>Not performed at Genomic Health; adjacent docetaxel-plus–trastuzumab</td>
<td>45/57 (80.9)</td>
<td>Mean age, 49 ± 8; ER-positive, 57%; LN-negative, 16%; chemotherapy, 100%</td>
</tr>
</tbody>
</table>

* AUC = area under the receiver-operator characteristic curve; CMF = cyclophosphamide, methotrexate, and fluorouracil; ER = estrogen receptor; HER2 = human epidermal growth factor receptor 2; HR = hazard ratio; KM = Kaplan–Meier; LN = lymph node; NSABP = National Surgical Adjuvant Breast and Bowel Project; OR = odds ratio; PR = progesterone receptor; RECEPT = response evaluation criteria in solid tumors; RS = recurrence score; TG = tumor grade; TS = tumor size.
### Appendix Table 2. Studies on the MammaPrint Gene Expression Test*

<table>
<thead>
<tr>
<th>Study, Year (Reference)</th>
<th>Design, Description, or Aim</th>
<th>Patients Analyzed/Eligible, n (%)</th>
<th>Population Characteristics</th>
<th>End Points</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>van’t Veer et al., 2002 (8)</td>
<td>Signature development; all age &lt;50 y; all ≤5 cm</td>
<td>Training set, 78; test set, 19</td>
<td>Mean age, 44.9 y; ER-positive, 70.2%; LN-negative, 100%; PR-positive, 57.7%; T5 &lt;2 cm (41.2%); TG, good (12%); TG, poor (49%); tamoxifen, 4%; chemotherapy, 4%</td>
<td>Distant metastasis as first relapse event at 5 y</td>
<td>Identification of the 70-gene signature; high- and low-risk groups algorithm: Training, OR (95% CI); Multivariate: 18 (3–94); P = 0.001 Validation of the prognostic value of the 70-gene signature; the 70-gene signature was associated with age, TG, ER (P &lt; 0.001), and T5 (P = 0.012) OR (95% CI): 67 LN-negative (not in van’t Veer et al., 2002 [8]); OR, 15.3 (95% CI, 1.6–127); P = 0.003; 180 LN-negative and negative (not in van’t Veer et al., 2002 [8]); CR, 14.6 (CI, 4.3–50); P &lt; 0.001 Univariate HR (95% CI): All patients: 5.1 (2.9–9.0); P &lt; 0.001 151 LN-negative patients: 5.5 (2.5–12.2); P &lt; 0.001 Multivariate HR (95% CI): All patients: 4.8 (2.9–9.2); P &lt; 0.001 Similar ORs and HRs compared with those obtained in the original cohorts were found. van’t Veer et al., 2002 (8) series, OR (95% CI); MammaPrint, 13.95 (95% CI); 778 differently classified by MammaPrint, 151 LN-negative patients from van’t Veer et al., 2002 (8) (HR; 95% CI); MammaPrint, 5.6 (2.4–9.3); P &lt; 0.001, similar results were obtained for overall survival.</td>
</tr>
<tr>
<td>van de Vijver et al., 2002 (9)</td>
<td>Signature validation; all age &lt;53 y; all ≤5 cm; 61 in common with van’t Veer et al., 2002 (8)</td>
<td>Total, 295; poor prognosis, 190; good prognosis, 115</td>
<td>Poor prognosis: ER-positive, 63%; LN-negative, 51%; T5 &lt;2 cm, 47%; TG-1, 11%; TG-3, 18%; tamoxifen, 13%; chemotherapy, 37% Good prognosis: ER-positive, 97%; LN-negative, 52%; T5 &lt;2 cm (82%); TG-1, 49%; TG-3, 12%; tamoxifen, 15%; chemotherapy, 28%</td>
<td>Distant metastasis as first relapse event at 5 y; overall survival</td>
<td>Similar ORs and HRs compared with those obtained in the original cohorts were found. van’t Veer et al., 2002 (8) series, OR (95% CI); MammaPrint, 13.95 (95% CI); 778 differently classified by MammaPrint, 151 LN-negative patients from van’t Veer et al., 2002 (8) (HR; 95% CI); MammaPrint, 5.6 (2.4–9.3); P &lt; 0.001, similar results were obtained for overall survival.</td>
</tr>
<tr>
<td>Glas et al., 2006 (15)</td>
<td>Assay development; reanalysis of previous cohorts</td>
<td>van’t Veer et al., 2002 (8), 78; van de Vijver et al., 2002 (9), 145</td>
<td>All 78 patients from van’t Veer et al., 2002 (8); 145 LN-negative patients from van de Vijver et al., 2002 (9)</td>
<td>Distant metastasis as first relapse event at 5 y; overall survival</td>
<td>Validation of the prognostic value of the 70-gene signature: the 70-gene signature was associated with age, TG, ER (P &lt; 0.001), and T5 (P = 0.012) OR (95% CI): 67 LN-negative (not in van’t Veer et al., 2002 [8]); OR, 15.3 (95% CI, 1.6–127); P = 0.003; 180 LN-negative and negative (not in van’t Veer et al., 2002 [8]); CR, 14.6 (CI, 4.3–50); P &lt; 0.001 Univariate HR (95% CI): All patients: 5.1 (2.9–9.0); P &lt; 0.001 151 LN-negative patients: 5.5 (2.5–12.2); P &lt; 0.001 Multivariate HR (95% CI): All patients: 4.8 (2.9–9.2); P &lt; 0.001 Similar ORs and HRs compared with those obtained in the original cohorts were found. van’t Veer et al., 2002 (8) series, OR (95% CI); MammaPrint, 13.95 (95% CI); 778 differently classified by MammaPrint, 151 LN-negative patients from van’t Veer et al., 2002 (8) (HR; 95% CI); MammaPrint, 5.6 (2.4–9.3); P &lt; 0.001, similar results were obtained for overall survival.</td>
</tr>
<tr>
<td>Buyse et al., 2006 (38)</td>
<td>Assay validation; multicenter study; all age &lt;60 y; all ≤5 cm</td>
<td>Total, 324/401 (80.9%); clinical information available for 302 patients; Adjuvant! Online/MammaPrint: Low/low risk: n = 92 Low/high risk: n = 9 High/low risk: n = 28 High/high risk: n = 163</td>
<td>All patients: tamoxifen, 0%; chemotherapy, 0% Adjuvant! Online/MammaPrint: Low/low risk: ER-positive, 100%; T5 &lt;2 cm, 67%; TG, good (41%); TG, poor (59%) Low/high risk: ER-positive, 100%; T5 &lt;2 cm (93%); TG, good (43%); TG, poor (57%) High/low risk: ER-positive, 91%; T5 &lt;2 cm (29%); TG, good (12%); TG, poor (88%) High/high risk: ER-positive, 48%; T5 &lt;2 cm (25%); TG, good (3%); TG, poor (90%)</td>
<td>Distant metastasis as first relapse event at 5 and 10 y and beyond: disease-free survival and overall survival</td>
<td>HR (95% CI); MammaPrint adjusted by Adjuvant! Online: time to distant metastasis, 2.13 (1.19–3.82); disease-free survival, 1.36 (0.93–2.32); overall survival, 2.43 (1.49–4.04) Development of metastases within 5 years: Sensitivity for MammaPrint: 0.90 (CI, 0.78–0.95); sensitivity for Adjuvant! Online: 0.87 (CI, 0.75–0.94); specificity for MammaPrint: 0.42 (CI, 0.36–0.48); specificity for Adjuvant! Online: 0.39 (CI, 0.24–0.35) AUC: time to distant metastasis: MammaPrint, 0.681; Adjuvant! Online, 0.648 AUC, overall survival: MammaPrint, 0.659; Adjuvant! Online, 0.616</td>
</tr>
</tbody>
</table>

* AUC = area under the receiver-operating characteristic curve; ER = estrogen receptor; HR = hazard ratio; LN = lymph node; OR = odds ratio; PR = progesterone receptor; TS = tumor size; TG = tumor grade.

---

**Note:** This table provides data for studies involving the MammaPrint Gene Expression Test, detailing various aspects such as assay validation, signature development, and endpoint results. It includes information on patient characteristics, end points, and statistical results, such as the area under the curve (AUC) for different survival metrics.
### Appendix Table 3. Studies on the H/I Gene Expression Test*

<table>
<thead>
<tr>
<th>Study, Year (Reference)</th>
<th>Design and validation of the 2-gene ratio signature; gene identification by microarray analysis; implementation by real-time RT-PCR analysis of frozen (training set) and FFPE (test set) specimens</th>
<th>Patients Analysed or Eligible, n (%)</th>
<th>Population Characteristics</th>
<th>End Points</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ma et al., 2004 (20)</td>
<td>Development and validation of the 2-gene ratio signature; gene identification by microarray analysis; implementation by real-time RT-PCR analysis of frozen (training set) and FFPE (test set) specimens</td>
<td>Training set, 59/60; recurrence, 28; nonrecurrence, 32</td>
<td>All patients: tamoxifen, 100%; chemotherapy, 0%</td>
<td>Disease-free survival</td>
<td>Training set, frozen specimens: Recurrent disease-free survival: 54.8 mo (range, 5–137 mo); nonrecurrence disease-free survival: 115.6 mo (range, 61–569 mo); 48% analysis, log-rank test; P = 0.001 Test set, FFPE specimens: Recurrent disease-free survival: 51.4 mo (range, 15–117 mo); nonrecurrence disease-free survival: 95.8 mo (range, 25–123 mo); 48% analysis, log-rank test; P = 0.002 Test set classification results: 16/20 correctly classified</td>
</tr>
<tr>
<td>Ma et al., 2006 (41)</td>
<td>Development of the 2-gene index used by the H/I assay; optimized risk groups stratification</td>
<td>All samples, 52/670 (HR); tamoxifen treated, 286; not treated with tamoxifen, 566</td>
<td>All samples: age &lt; 50 y, 82%; ER-positive, 73%; TN-negative: 72%</td>
<td>Disease-free survival at 5 years</td>
<td>H/I index (continuous variable), 5 y recurrence risk in untreated patients (95% CI): HR = 2.0–15.9, P = 0.001–0.005; H/I 2 = 0–5 (8.8%–20.5%), HR = 2.0–36.0 (8.8%–49.2%) Multivariate Cox regression analysis, HR for optimally dichotomized H/I with 95% CI in ER-positive, ER-negative, tamoxifen treated and untreated (n = 225): HR, 3.9 (1.5–10.3), P = 0.007</td>
</tr>
<tr>
<td>Signature validation</td>
<td>2-gene signature validation</td>
<td>All samples: age &gt; 50 y, 93.1%; ER-positive, 100%; TN-negative, 22.5%; PR-positive, 22.5%; PR-negative, 77.5%; HER2-positive, 22.5%; HER2-negative, 77.5%;</td>
<td>Age &gt; 50 y, 93.1%; ER-positive, 100%; TN-negative, 22.5%; PR-positive, 22.5%; PR-negative, 77.5%; HER2-positive, 22.5%; HER2-negative, 77.5%;</td>
<td>Disease-free survival</td>
<td>The prognostic value of the 2-gene ratio was not confirmed in this population; univariate logistic regression OR (95% CI), 1.30 (0.88–1.93), P = 0.18; similar results by the other methods</td>
</tr>
<tr>
<td>Goiti et al., 2006 (38)</td>
<td>Randomized NCCTG 89-30-52 trial about tamoxifen treatment</td>
<td>206/211 (97.6)</td>
<td>ER-positive, 100%; TN-negative, 63%; HER2-positive, 18%; TN = 3 cm (76%); TG-1, 26%; TG-3, 18%; tamoxifen, 100%; chemotherapy, 0%</td>
<td>Disease-free survival, disease free survival; overall survival</td>
<td>All patients, multivariate HR (95% CI); RFS, 1.45 (0.93–2.27); disease-free survival, 1.57 (1.04–2.48); overall survival, 1.29 (0.81–2.06) LN-negative (n = 130) HR (95% CI): disease-free survival, 1.72 (0.53–5.23); disease-free survival, 1.77 (0.99–3.14); overall survival, 2.21 (1.03–4.99)</td>
</tr>
<tr>
<td>Jamien et al., 2007 (39)</td>
<td>Evaluate the prognostic value of the 2-gene ratio; optimized stratification in risk groups different from Ma et al., 2006 (25)</td>
<td>Total, 1252/1699</td>
<td>Disease-free survival in 468 patients: ER-positive, 100%; TN-negative, 63%; chemotherapy, 0%</td>
<td>Disease-free survival</td>
<td>Disease-free survival, overall survival; PFS, overall survival</td>
</tr>
<tr>
<td>Immervoll et al., 2007 (40)</td>
<td>Randomized, controlled trial about the benefit of tamoxifen treatment duration; optimized stratification in risk groups different from Ma et al., 2006 (25)</td>
<td>Total, 357/373 (95.7); postmenopausal, 264</td>
<td>Postmenopausal: ER-positive, 74%, TN-negative, 28%; 2 y tamoxifen, 62%; 5 y tamoxifen, 38%; TS &gt;2 cm (29%)</td>
<td>Recurrence rate (95% CI):</td>
<td>The ratio was significantly associated with TS, P = 0.006; OR (95% CI), 3.9 (0.17–91); HR (95% CI), 2.0 (0.18–22); HR (95% CI), 2.0 (0.18–22); HR (95% CI), 2.0 (0.18–22); HR (95% CI), 2.0 (0.18–22)</td>
</tr>
</tbody>
</table>