Published in final edited form as:

Cancer Invest. 2009 November ; 27(9): 885–890. doi:10.3109/07357900903275142.

Gene Expression Profile Assays as Predictors of Distant Recurrence-Free Survival in Early-Stage Breast Cancer

Nicole M. Kuderer and Gary H. Lyman

Department of Medicine, Duke University Medicine and the Duke Comprehensive Cancer Center, Durham, NC, USA

INTRODUCTION

The main prognostic factors for early-stage breast cancer (ESBC) include tumor size, nodal status, hormone receptor status, HER2/neu status, histology, and tumor grade. While these and other prognostic factors have been incorporated into treatment guidelines, the remaining prognostic uncertainty means that some patients destined to recur do not receive adjuvant therapy and others are unnecessarily treated experiencing only unnecessary side effects. Clearly better prognostic tools are needed to guide treatment decisions in patients with ESBC (1). New, high-performance screening techniques using DNA microarrays have permitted the analysis of patterns of gene expression in thousands of genes simultaneously. Several investigators have reported efforts to define gene expression signatures based on their ability to predict the risk of disease recurrence and guide the use of adjuvant systemic therapy in ESBC. Before the role of such assays in actual clinical practice can be fully understood, a careful evaluation and validation of test performance in a variety of settings is essential. Such an evaluation ideally precedes the wide scale dissemination and utilization of new diagnostic and prognostic technologies.

The analysis reported here is part of an ongoing effort to systematically review the test performance of various classes of gene expression signatures in women with ESBC and to explore any source of heterogeneity observed.

The enormous cost, and duration of observation required for prospective controlled clinical trials of long-term outcomes often prompts interim observational studies to evaluate the performance characteristics of prognostic assays compared to an intermediate gold standard. While such studies are of limited scope, they permit the assessment of specific aspects of test performance in various populations and require much less time and fewer resources to conduct. The availability of archived clinical material and the development of assays that can be performed on such fixed specimens have further enhanced timely and efficient validation in groups of patients with information on long-term clinical outcomes.

METHODS

Data acquisition

A systematic review of the all-language published literature on gene expression profile studies in patients with ESBC was undertaken using the following data sources:

Copyright © Informa Healthcare USA, Inc.

Correspondence to: Nicole M. Kuderer, M.D., Duke University School of Medical Center, DUMC 3841, Durham, NC 27710, USA.

DECLARATION OF INTEREST

The authors report no conflict of interest. The authors alone are responsible for the content and writing of this paper.

MEDLINE®, EMBASE®, The Cochrane Library, DARE (Database of Abstract of Reviews of Effectiveness), and a review of references in identified articles. Search terms included gene expression profiles or profiling, gene signatures, gene expression chips, microarrays, and breast neoplasm or cancer. The study population had to be an original study group with duplicate articles based on the same group of patients excluded. Follow-up studies in which a subset of patients were reported, only the most recent article containing the most up-to-date results was reported. Eleven independent cohorts of patients with ESBC were identified in which the relationship between a gene expression signature and distant recurrence-free survival is available. The primary outcome of interest was distant recurrence-free survival based on gene expression risk category. Data extraction was undertaken by two independent investigators in a blinded fashion with any unresolved issues settled by a third independent investigator.

Statistical analysis

Primary outcomes—The primary outcome of interest was distant recurrence-free survival. Patients were stratified according to the gene expression profile risk grouping into high- or low-risk. Measures of test performance consisted of sensitivity, specificity, positive and negative predictive value, likelihood ratio, the diagnostic odds ratio, and the receiver-operating characteristic (ROC) curve. Sensitivity is defined as the proportion of patients with subsequent disease recurrence classified as high risk by the gene signature, while specificity is the proportion of patients without disease recurrence classified as low risk. Sensitivity and specificity will vary as the threshold or cutpoint defining risk is varied. A cutpoint associated with greater sensitivity will result in a loss of specificity and vice versa. The predictive value positive is defined as the probability of disease recurrence in those classified as high-risk, whereas the predictive value negative is the probability of no disease recurrence in those classified as low-risk. The predictive value depends both on model performance and the risk of distant recurrence or death. The likelihood ratio represents the probability of the assay result in someone with recurrence divided by the probability in someone without recurrence and is a useful summary measure for describing the discriminatory ability of high- or low-risk assay results. The diagnostic odds ratio is a useful measure of overall assay prognostic discrimination and can be thought of as the ratio of the likelihood ratio positive and likelihood ratio negative. The ROC curve plots the true positive rate (sensitivity) and the false positive rate ($1 - \text{specificity}$) from different studies. Overall test discrimination is associated with the area under the ROC curve corresponding to the c -statistic for dichotomous measures.

Descriptive statistics—The distributions of all outcomes were evaluated and summary measures of central tendency and variability were estimated. Measures of central tendency included the overall proportions, mean, median, and variance weighted summary measures based on the method of Mantel and Haenszel. Bivariate correlations between continuous measures were based either on the Pearson coefficient as a measure of linear correlation for data close to normal in distribution or the Spearman's rho as a nonparametric measure of correlation. Tests for linearity were derived from the sum of squares, degrees of freedom, and mean square associated with linear and nonlinear components based on ANOVA. Categorical measures were compared using a χ^2 test. The relationship of primary outcomes to study size, proportion with positive lymph nodes, technique used, and study quality was evaluated.

Meta-analysis—Study-by-study heterogeneity was assessed for each test performance measure based on the Q-statistic (2). The hypothesis that the studies are all drawn from a population of studies with the same effect size is rejected if the Q-statistic exceeds the upper 100 (1 - α) percentile of the χ^2 distribution. An inconsistency index (I^2) was estimated by

the method of Altman as $(H^2 - 1)/H^2$ where $H^2 = Q/(k - 1)$ and k the degrees of freedom (3). I^2 was estimated reflecting the proportion of variation in study estimates due to between study heterogeneity rather than random variation. Due to the significant heterogeneity between studies, weighted summary outcome estimates were based on a random effects model as proposed by DerSimonian and Laird (4). Under a random effects model, the studies are assumed to be from different populations with different true effect sizes. With this conservative approach, the true measure may differ between studies due to differences in patient populations, treatment variation, or because outcome measures differ from one study to the next. Therefore, two sources of variation are assumed consisting of random error and variation due to real differences between the studies. As a result, the standard error of the outcome measure estimates will approach zero as the sample size within studies increases, while the differences in the true effect sizes between studies will persist.

Hypothesis testing on summary effect estimates was based on a z-statistic with estimates of standard error and 95% confidence interval (CI) provided for all individual studies as well as the summary overall effect estimate. Results are presented as forest plots with effect estimates and 95% confidence limits (CLs) presented for each individual study and a summary measure and CLs across all studies. Evidence for a publication bias in studies included in the meta-analysis is evaluated using a funnel plot of the diagnostic odds ratio and its precision (1/standard error).

RESULTS

Summary of included studies

Of the 453 studies, in the published or presented literature, identified between 1990 and 2005, 11 were found to represent separate study populations of ESBC patients with validation based on distant recurrence or distant recurrence-free survival (5-14). Table 1 summarizes the 11 validation cohorts included in this systematic review. The 11 series included 1520 patients (range 20–668) of which 824 (54%) patients were classified as high-risk based on the gene assay and 354 (23%) experienced distant breast cancer recurrence during the period of observation. The reported recurrence rates were 35% (95% CI: 26–42) among gene expression profile high-risk patients and 9% (95% CI: 7–12) among low-risk patients. Median follow-up on patients ranged from 2 to 14 years. No adjuvant treatment was utilized in two studies, limited to chemotherapy in one and tamoxifen in one, whereas it was not restricted in three and not specified in the remaining four studies. Six cohorts were evaluated utilizing cross-validation techniques while five were studied in independent cohorts of patients. As shown in Table 1, reported series included six limited to lymph node negative patients, two restricted to node positive patients, and three with both node positive and negative patients. Only one series limited the study to hormone receptor positive women (11).

Test performance characteristics

Summary measures—Summary test performance characteristics for distant disease recurrence-free survival are shown in Table 2 including the overall rates pooling all studies as well as the average sensitivity, specificity, likelihood ratio positive and negative, the diagnostic odds ratio, and the positive and negative predictive values across studies. Among assay performance measures, significant heterogeneity was observed for sensitivity ($P = 0.0082$), specificity ($P < 0.0001$), positive predictive value ($P < 0.0001$), and the diagnostic odds ratio ($P = 0.0011$) but not for the negative predictive value ($P = 0.2112$), likelihood ratio positive ($P = 0.2632$), or the likelihood ratio negative ($P = 0.1795$).

Sensitivity and specificity—Weighted summary assay sensitivity and specificity were estimated at 82.4% (95% CI 76.1–88.7) and 53.3% (95% CI 43.9–62.7), respectively. The corresponding overall proportion of patients with recurrence classified as low-risk (false negative rate) was 17.8% (95% CI 14.2–22.1), while the proportion of patients with no recurrence but classified as high-risk (false positive) was 46.1% (95% CI 43.2–49.0).

Predictive value—Figures 1 and 2 present forest plots with individual study and weighted summary measures for assay positive and negative predictive values demonstrating weighted summary estimates of 42.5% (95% CI 32.2–52.7) and 89.8% (95% CI 86.7–93.0), respectively. The inconsistency index is 87% and 24% for the positive and negative predictive values, respectively. The overall estimate of distant recurrence risk in those with low-risk assays (1 — predictive value negative) is 9.2% (95% CI 7.2–11.6). Likewise, the overall estimate of no distant recurrence in those with high-risk assays is 64.7% (95% CI 61.4–67.9). As anticipated, the positive predictive value is directly proportional to the overall risk of recurrence in the study population.

Diagnostic odds ratio—Weighted summary estimates of 1.78 (95% CI 1.47–2.16) and 0.35 (95% CI 0.26–0.47) were calculated for assay likelihood ratio positive and negative, respectively. The weighted summary estimate of the diagnostic odds ratio as a measure of overall test discrimination was 6.44 (95% CI 3.42, 12.08). The estimated area under the ROC curve ($\pm$ standard error) was 0.8611 ± 0.0495 .

Sources of heterogeneity—No significant relationship was observed between the measures of assay performance considered and lymph node status, hormone receptor status, or type of adjuvant treatment. Assay performance measures were generally weaker in studies with independent validation but only reached statistical significance for predictive value positive with medians of 54% and 32% for crossover and independent validation studies, respectively ($P = 0.045$). Significant correlation was observed between the number of genes in the signature studied and assay sensitivity ($r_{sp} = 0.673$; $P = 0.047$) and diagnostic odds ratio ($r_{sp} = 0.648$; $P = 0.043$). The association between the number of genes in the signature of each study and the diagnostic odds ratio as a measure of overall model discrimination is shown in Figure 3.

DISCUSSION

Gene expression profile assays based on microarray analysis show early promise for predicting recurrence-free survival in patients with ESBC. However, the use of these assays in therapeutic decision-making must consider both the limitations of assay test performance and the specific patient population being evaluated (15-17). Systematic reviews of prognostic studies may contribute to a better understanding of these techniques by summarizing the results of test accuracy from several different investigations, identifying reasons for variation in the results of individual studies, and potentially improving the quality of future studies through a delineation of the methodological inadequacies of previous reports. Finally, such reviews may be capable of generating new hypotheses that may be tested in future definitive studies.

A systematic review of reported studies of gene expression assays in women with ESBC is presented here. The gene expression assays included in the analysis were limited to those providing validation in women with ESBC based on distant recurrence-free survival. A formal meta-analysis of test performance characteristics observed for these assays and outcomes was conducted.

The variability of assay prognostic performance characteristics may relate to both random error (chance) and to systematic differences or heterogeneity between studies. Most of the validation studies reported here were small with more than half including less than 100 patients. Since assay performance measures are more accurately estimated in larger studies, such studies deserve particular attention. The methodology employed in this systematic review, in fact, weighs the individual studies inversely in proportion to their contribution to the variance and, therefore, larger studies will generally contribute more to the summary estimates. Substantial heterogeneity was also observed across studies for most performance measures. Weighted pooled performance estimates were therefore based on a random effects model and emphasis was placed on the ROC analysis. While assay performance varied considerably across reports, it was limited to some degree in all studies. Causes of study-by-study variation include the use of different gene signatures and risk score cutpoints. In addition, the inclusion of studies based on cross-validation may overestimate the accuracy of an assay while independent validation offers the truest estimate of how an assay actually performs (15, 16). Summary estimates of prognostic performance in this review were found to be less in studies with independent validation, as anticipated. However, these differences were not statistically significant.

Systematic reviews of prognostic studies, like those of therapeutic trials, are dependent upon the quality of the individual studies included in its analysis (18-22). The patient selection process should be fully described and the reference standard outcome explicitly defined and generally accepted. The assay procedure and the definition of high-risk and low-risk results should also be fully described. Ideally, the investigator and interpreter of the test are blinded as to the disease status of the patients and the method of concealment should be detailed. The same uniform assay procedure and outcome evaluation should be employed for all patients. Inconsistent reporting of patient characteristics, assay characteristics, and statistical analysis, increases the inter-study variability, leading to increased variability in the results of the meta-analysis (23, 24). Variation in study population and clinical setting as well as the assay utilized can each result in variation in diagnostic accuracy. Additional sources of heterogeneity in assay performance may include the cutpoint or threshold defining high-risk and low-risk patients, measurement differences between observers, and differences in risk of recurrence.

Superior prognostic performance compared to conventional clinical prognostic markers or guidelines has been reported for some gene expression assays (6, 11, 14). Despite the prognostic information provided by these assays, however, clinicians need to be familiar with their limitations, in order to appropriately utilize them in clinical practice. As reported here, while the sensitivity of the gene assays for predicting recurrence was relatively high in some studies, the specificity for identifying those who remain disease-free was quite low. An assay is most useful for ruling out a future recurrence if it has high sensitivity, whereas it is most useful for confirming the risk of future recurrence if it has high specificity. Therefore, while the negative predictive value ranged from 75% to 100% across studies, the positive predictive value ranged from 25% to 88% or 43% overall. As expected, the predictive value of the gene signatures depends on not only the sensitivity and specificity of assay but also upon the risk of distant recurrence in the population under study. Assays utilized in higher risk women with ESBC will be associated with greater predictive value positive but lower predictive value negative despite a low-risk assay result.

The risk of distant recurrence, the potential effect of treatment on risk as well as the impact of disease and treatment on patient quality of life must all be considered in determining the appropriate role for such assays in clinical care. Although systematic reviews of prognostic accuracy can be quite valuable, particularly early in the evaluation of a new assay, such studies are not capable of addressing all clinically relevant questions. Clearly, large

controlled clinical trials represent the gold standard for evaluating any prognostic or therapeutic intervention that might have an impact on important clinical outcomes. However, such trials are handicapped by the great cost, resources, and time required. Clearly, the validity of meta-analyses of assay performance studies as well as controlled clinical trials are both dependent upon the quality of the original studies. In addition, the validity of systematic reviews is dependent upon capturing results from all available studies. This review was based on assay validation studies published in the peer-review literature or recently presented at major meetings.

Systematic evaluations such as that reported here should be useful not only to clinicians in evaluating and potentially incorporating the results of such assays into clinical practice but also to investigators involved in the development and validation of new gene expression signatures. Undoubtedly, the greatest challenge will be to define the potential role of these assays in clinical decision-making considering the clinical, economic, and quality of life impact of disease and treatments guided by the results of such prognostic tools. Current efforts to enhance the value of gene expression assays through incorporation of genes that are not only prognostic but also predictive of response and toxicity of specific therapies should further increase the clinical and quality of life impact as well as the cost-effectiveness of these tests.

Several specific recommendations can be put forth for current and future studies of gene expression signatures in ESBC as well as those related to other malignancies. Investigators should strive for independent validation in a large enough population of patients to provide unbiased and accurate estimates of assay performance. Transparency is needed in study design including the selection of genes for inclusion in the assay and the patient population utilized including clinical prognostic factors and methods of treatment. It is essential that any potential role for these assays be defined in comparison to or in combination with recognized clinical prognostic factors.

REFERENCES

1. Altman DG, Lyman GH. Methodological challenges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Res Treat.* 1998; 52:289–303. [PubMed: 10066088]
2. Laird MM, Mosteller F. Some statistical methods for combining experimental results. *Int J Technol Assess Health Care.* 1990; 6:5–30. [PubMed: 2361819]
3. Higgins JP, Thompson SG, Deeks JJ, Altman D. Measuring inconsistency in meta-analyses. *Brit Med J.* 2003; 327:557–560. [PubMed: 12958120]
4. DerSimonian R, Laird NM. Meta-analysis in clinical trials. *Contr Clin Trials.* 1986; 7:177–188.
5. van't Veer LJ, Dai H, van De Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature.* 2002; 415:530–536. [PubMed: 11823860]
6. van De Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med.* 2002; 347:1999–2009. [PubMed: 12490681]
7. Ahr A, Karn T, Solbach C, et al. Identification of high risk breast-cancer patients by gene expression profiling. *Lancet.* 2002; 359:131–132. [PubMed: 11809257]
8. Bertucci F, Nasser V, Granjeaud S, et al. Gene expression profiles of poor prognosis primary breast cancer correlate with survival. *Hum Mol Genet.* 2002; 11:863–872. [PubMed: 11971868]
9. Huang E, Cheng SH, Dressman H, et al. Gene expression predictors of breast cancer outcomes. *Lancet.* 2003; 361:1590–1596. [PubMed: 12747878]
10. Iwao K, Matoba R, Ueno N, et al. Molecular classification of primary breast tumors possessing distinct prognostic properties. *Hum Mol Genet.* 2002; 11:199–206. [PubMed: 11809729]
11. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node negative breast cancer. *N Engl J Med.* 2004; 351:2817–2826. [PubMed: 15591335]

12. Esteva FJ, Sahin AA, Coombes L, et al. Multi-gene RT-PCR assay for predicting recurrence in node-negative breast cancer patients that did not receive adjuvant tamoxifen nor chemotherapy. *Breast Cancer Res Treat.* 2004; 88(suppl 1):1–253. [PubMed: 15538040]
13. Esteva FJ, Sahin AA, Cristofanilli M, et al. Prognostic role of a multi-gene RT-PCR assay in patients with node-negative breast cancer not receiving adjuvant systemic therapy. *Clin Cancer Res.* 2005; 104:676–681.
14. Wang Y, Klijn JGM, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node negative primary breast cancer. *Lancet.* 2005; 365:671–679. [PubMed: 15721472]
15. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst.* 2003; 95:14–18. [PubMed: 12509396]
16. Ntzani EE, Ioannidis JPA. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet.* 2003; 362:1439–1444. [PubMed: 14602436]
17. Simon R. Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *Brit J Cancer.* 2003; 89:1599–1604. [PubMed: 14583755]
18. Lyman GH, Djulbegovic B. The challenge of systematic reviews in cancer screening, diagnostic and staging studies. *Cancer Treat Rev.* 2005; 31:628–639. [PubMed: 16225994]
19. Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. *BMJ.* 2001; 323:157–162. [PubMed: 11463691]
20. Vamvakas E. Meta-analysis of studies of the diagnostic accuracy of laboratory tests: a review of the concepts and methods. *Arch Pathol Lab Med.* 1998; 122:675–686. [PubMed: 9701328]
21. Irwig L, Tosteson ANA, Gatsonis C, et al. Guidelines for meta-analysis evaluating diagnostic tests. *Ann Intern Med.* 1994; 120:667–676. [PubMed: 8135452]
22. Shepes SB, Schechter MT. The assessment of diagnostic tests. *J Am Med Assoc.* 1984; 252(17): 2418–2422.
23. Lyman GH, Kuderer NM. The strengths and limitations of meta-analyses based on aggregate patient data. *BMC Med Res Methodol.* 2005; 5(14):1–7. [PubMed: 15636638]

Figure 1.

Forest plot of the positive predictive value of gene expression profile signatures $\pm$ 95% CLs in ESBC estimated by the Mantel-Haenszel method for each study. The weighted combined positive predictive value $\pm$ 95% CLs based on a random effects model is shown at the bottom.

Figure 2.

Forest plot of the negative predictive value of gene expression profile signatures $\pm 95\%$ CLs in ESBC estimated by the Mantel—Haenszel method for each study. The weighted combined negative predictive value $\pm 95\%$ CLs based on a random effects model is shown at the bottom.

Figure 3. Graph of the diagnostic odds ratio and the number of genes in the assay signature from each study of patients with early-stage breast cancer. The filled line is based on linear regression on the mean $\pm$ 95% CLs.

Table 1

Gene expression profile assays in early-stage breast cancer study characteristics

Study	Country	Patients	FU ^a	LN status	ER	Treatment
Ahr-1	Germany	27	2.0	Positive	Mixed	Unknown
Ahr-2	Germany	20	2.0	Negative	Mixed	Unknown
Bertucci	France	55	5.0	Both	Mixed	Chemotherapy
Huang	USA	52	3.0	Both	Mixed	Unknown
Iwao	Japan	119	4.5	Both	Mixed	Unknown
Paik	USA	668	14.0	Negative	Positive	Tamoxifen
Esteva	USA	149	10.0	Negative	Mixed	None
van De Vijver-1	The Netherlands	78	7.8	Negative	Mixed	Mixed
van De Vijver-2	The Netherlands	67	7.8	Negative	Mixed	Mixed
van De Vijver-3	The Netherlands	114	7.8	Positive	Mixed	Mixed
Wang	The Netherlands	171	5.0	Negative	Mixed	None

Abbreviations: ER, estrogen receptor; LN, lymph node.^aFollow-up in years.

Table 2

Gene expression profile assays in early-stage breast cancer summary performance measures

Measure	Mean	Overall (95% CLs)
Sensitivity	84.2%	82% (78, 86)
Specificity	57.4%	54% (51, 57)
Likelihood ratio positive	3.45	1.78 (1.65, 1.932)
Likelihood ratio negative	0.32	0.35 (0.26, 0.47)
Diagnostic odds ratio	14.98	6.02 (3.02, 11.97)
Predictive value positive	45.6%	43% (32, 53)
Predictive value negative	89.0%	90% (87, 93)