

## Development and Validation of Therapeutically Relevant Multi-Gene Biomarker Classifiers

*Richard Simon*

In June 2004 Ma et al. (1) described a two-gene expression ratio that they claimed accurately predicted clinical outcome of early-stage breast cancer patients treated with adjuvant tamoxifen monotherapy. In the current issue of this journal, Reid et al. (2) reported their failure to confirm the usefulness of the two-gene expression ratio on independent data. I will attempt to try to provide possible explanations for the inconsistency of results of the two studies and to draw some general conclusions.

Oncologists need improved tools for selecting treatments for individual patients. Most cancer treatments benefit only a minority of the patients to whom they are administered. Being able to predict which patients are most likely to benefit not only would save patients from unnecessary toxicity and inconvenience but also might facilitate their receiving drugs that are more likely to help them. In addition, the current overtreatment of patients results in major expense for individuals and society, an expense that may not be indefinitely sustainable.

Although there is a large body of literature on prognostic factors for cancer patients, very few such factors are used in clinical practice. Prognostic factors are unlikely to be used unless they are therapeutically relevant, and most publications do not establish such therapeutic relevance. Most prognostic factor studies are conducted by use of a convenience sample of patients for whom tissue is available, and the cohort is often far too heterogeneous with regard to stage and treatment to support therapeutically relevant conclusions.

Microarray expression profiling has provided an exciting new technology for attempting to identify classifiers for tailoring treatments to patients. One serious limitation of microarray expression profiling, however, is that it is an RNA assay that requires fresh or well-preserved frozen tissue for extraction of viable RNA. This limits the availability of material for the development of therapeutically relevant predictive models and complicates the clinical application of such predictive models. Consequently, some investigators use microarrays to screen the genes for those factors associated with outcome and then develop classifiers for clinical applications that are based on subsequent studies using reverse transcription–polymerase chain reaction (RT–PCR) of the selected genes. This was the approach used by Paik et al. (3) in developing the Oncotype Dx risk score for patients with lymph node–negative estrogen receptor (ER)–positive breast cancer receiving tamoxifen monotherapy. RT–PCR can be conducted on RNA extracted from formalin-fixed paraffin-embedded tissue.

It is useful to divide genomic classifier studies into developmental studies and validation studies. Developmental studies define the multi-gene classifiers and are analogous to phase II clinical trials. They should indicate whether the genomic classifier is promising and worthy of evaluation in a phase III trial. There are special problems in evaluating whether a genomic classifier is promising on the basis of a developmental study.

The difficulty derives from the fact that the number of candidate genes available for use in the classifier is much larger than the number of cases available for analysis. In such situations, it is always possible to find linear classifiers that perfectly classify the data on which they were developed. This apparently perfect classification can be achieved even if there is no relationship between expression of any of the genes and outcome (5). Consequently, even in developmental studies, some kind of validation on data not used for developing the model is necessary. This “internal validation” is usually accomplished either by splitting the data into two portions, one used for training the model and the other used for testing the model or some form of cross-validation that is based on repeated model development and testing on random data partitions (5). This internal validation should not, however, be confused with external validation of the classifier in a setting simulating broad clinical application.

Ma et al. developed their classification model by use of the expression of 22000 genes measured on 60 patients. They restricted attention to the 5475 genes with greatest variation in expression across the samples. They did *t* tests on those genes to identify those whose mean expression was statistically significantly different among the patients who relapsed and patients who remained disease free. For their laser-capture microdissected samples, they found only nine genes that were statistically significant at the appropriately stringent .001 level. The fact that only nine genes were statistically significant would usually suggest that an accurate predictor of outcome would not be possible on the basis of those data. With 5475 genes tested, one would expect at least 5.4 false positives by chance alone ( $5475 \times 0.001$ ). This 60% false discovery rate is much greater than desired. Reid et al. re-analyzed Ma’s data with standard classifier and cross-validation procedures and obtained a cross-validated error rate of 39%.

Ma et al., however, did not use this standard approach to predictive model development. They focused instead on developing a model that used two of the three genes that were statistically significant in both their data from laser-capture microdissected material and their data from whole tissue breast tumor specimens from the same patients. They measured gene expression in the same set of patients with RT–PCR and developed a cut-off for distinguishing relapse patients from continuous disease-free patients that was based on the HOXB13/IL17BR ratio. Because they were using all of their original data to select the genes and the

---

*Affiliation of author:* Biometric Research Branch, National Cancer Institute, Bethesda, MD.

*Correspondence to:* Richard Simon, DSc, National Cancer Institute, 9000 Rockville Pike, MSC 7434, Bethesda, MD 20892 (e-mail: rsimon@nih.gov).

DOI: 10.1093/jnci/dji168

*Journal of the National Cancer Institute*, Vol. 97, No. 12, © Oxford University Press 2005, all rights reserved.

functional form for combining the gene expression measurements, Ma et al. selected an independent set of 20 patients for evaluating their two-gene ratio model. They applied the model developed on their initial set of 60 patients to predict outcome for the 20 additional patients and showed that the model correctly predicted recurrence or no recurrence for 16 of the 20 patients.

Ma et al. applied their model to the test set of 20 patients and evaluated prediction accuracy. Some investigators re-analyze a test set to see whether they would come up with the same set of genes. The latter is not appropriate, however. The set of genes selected for inclusion into a multi-gene classifier may be unstable because of correlations among gene expression. It is the prediction accuracy of a completely specified model that should validate, not the statistical significance of the individual gene components.

The “validation set” of Ma et al. consisted of 20 ER-positive patients with primary breast cancer treated with adjuvant tamoxifen monotherapy between 1991 and 2000 for whom both medical records and formalin-fixed paraffin-embedded blocks were available. The promise of the Ma et al. two-gene ratio rested heavily on the fact that the two-gene expression ratio predicted correctly for 16 of the 20 patients in the test set. It is unfortunate that the test set consisted of only 20 ER-positive patients receiving tamoxifen monotherapy. It seems surprising that although 19 of the 20 patients were lymph node negative, 10 of the 20 patients relapsed. This result may reflect some aspect of the selection process. A much larger independent test set of patients would have been preferable. The test set of Ma et al. should be viewed, however, as the set providing evidence that their two-gene ratio was promising, not for providing an adequate external validation of the classifier, and this distinction was acknowledged in the article.

The validation set of Reid et al. consisted of 58 patients with ER-positive primary breast cancer treated at the Istituto Nazionale Tumori between March 1991 and December 1997. The expressions of the two genes identified by Ma et al. and their ratio were not predictive of outcome in the Milan data. In the Milan cohort, 77.5% were lymph node positive compared with 47.2% for the original cohort of Ma et al. and 5% (1 of 20 specimens) for the test set. The Milan patients also had larger tumors, and 20.7% were HER2 positive compared with 5.4% for the initial cohort of Ma et al. In general, an external validation study should consist of patients who are the target population for use of the proposed classifier. The target population was not clearly defined by Ma et al., but patients with large, lymph node-positive tumors do not seem like good candidates today for tamoxifen monotherapy, particularly not those who are HER2 positive. Consequently, it would have been useful for Reid et al. to report on an analysis of the two-gene ratio for the subset of their patients who were lymph node negative. They had only about 13 such patients, however, and this sample size would not be sufficient to either validate or refute the findings of Ma et al. for the lymph node-negative population.

Reid et al. also re-analyzed microarray expression data of Ma et al. and the expression data for a subset of the cohort previously published by Sotiriou et al. (6). The cohort of Sotiriou et al. consisted of 99 patients, only 33 of whom were ER positive and lymph node negative. The initial cohort of Ma et al. included only 28 patients who were ER positive and lymph node negative. Neither of these cohorts is really ideal for developing therapeutically relevant gene expression-based classifiers or for concluding that microarray data are insufficient for developing such models. The results of Paik et al. (3) were based on larger studies and provide strong evidence that prediction of outcome for ER-positive

patients with lymph node-negative breast cancer receiving tamoxifen is possible by analyzing expression data. Although the results of the study of Reid et al. do not support the usefulness of the two-gene ratio reported by Ma et al., evaluation of the two-gene ratio in a larger study of ER-positive, lymph node-negative patients receiving tamoxifen monotherapy is desirable. Sgroi et al. (7) reported that their confirmatory study in a cohort of ER-positive patients from a randomized clinical trial indicated that the two-gene ratio was “a more robust predictor in lymph node negative patients, as compared with lymph node positive patients.”

In general, I make the following recommendations:

- Both developmental and validation studies should be based on cohorts of patients that are sufficiently homogeneous for therapeutically relevant classifiers to be developed. Generally, this goal is best achieved by studying patients who were included in a single large randomized clinical trial.
- Developmental studies should be sufficiently large so that they can incorporate either cross-validation or split sample validation and demonstrate that the internally validated prediction error is statistically significantly less than would be expected by chance (8).
- Independent validation studies are essential before results are accepted into medical practice (9).
- Validation studies should apply the classifier completely specified (including cut offs) by the developmental study and measure prediction accuracy. The size of the validation study should be sufficient so that meaningful confidence intervals on predictive accuracy and positive and negative predictive values can be reported. The size of the validation study should also be sufficient so that the extent to which the classifier adds predictive accuracy to established prognostic factors can be meaningfully evaluated (10).
- Raw expression data from developmental studies should be made publicly available to enable others to re-analyze the data and perform subset and meta-analyses.

## REFERENCES

- (1) Ma X, Wang Z, Ryan PD, Isakoff SJ, Barmettler A, Fuller A, et al. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* 2004;5:607–16.
- (2) Reid JF, Lusa L, De Cecco L, Coradini D, Veneroni S, Grazia Daidone M, et al. Limits of predictive models using microarray data for breast cancer clinical treatment outcome. *J Natl Cancer Inst* 2005;97:927–30.
- (3) Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated node-negative breast cancer. *N Engl J Med* 2004;351:2817–26.
- (4) Simon R. Diagnostic and prognostic prediction using gene expression profiles in high dimensional microarray data. *Br J Cancer* 2003;89:1599–604.
- (5) Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the analysis of DNA microarray data: class prediction methods. *J Natl Cancer Inst* 2003;95:14–8.
- (6) Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A* 2003;100:10393–8.
- (7) Sgroi DC, Haber DA, Ryan PD, Ma Xj, Erlander MG. RE: A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* 2004;6:445.
- (8) Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. *J Computational Biol* 2002;9:505–11.
- (9) Simon R. When is a genomic classifier ready for prime time? *Natl Clin Practice Oncol* 2004;1:4–5.
- (10) Kattan MW. Evaluating a new marker’s predictive contribution. *Clin Cancer Res* 2004;10:822–4.