Development and Evaluation of Therapeutically Relevant Predictive Classifiers Using Gene Expression Profiling

Richard Simon

Gene expression profiles offer the possibility of improving risk predication and optimizing treatment selection for individual patients. Two articles in this issue of the Journal describe clinical studies of gene expression profiling—one a developmental study and the other a validation study. Asgharzadeh et al. *(1)* address the development of a prognostic classifier for patients with metastatic neuroblastoma lacking amplification of the MYCN gene. Buyse et al. *(2)* report the validation of a gene expression–based prognostic classifier for patients with early breast cancer.

Asgharzadeh et al. *(1)* developed their classifier based on the expression of 55 genes that appears to predict risk of disease progression more accurately than does patient age, histologic type, or other currently used risk features. The claims of improved risk prediction are based on an internal estimate of prediction accuracy computed by Asgharzadeh et al. The approach taken by the authors allowed them to avoid one of the major pitfalls of developmental studies, which is that they often provide highly biased estimates of accuracy. The fundamental principle is that the same data should not be used for developing a predictive classifier and for evaluating the accuracy of that classifier. This principle is especially important for microarray-based studies because the number of candidate predictors (genes) is generally orders of magnitude greater than the number of cases. In this setting, the bias of using the same data for developing and evaluating a classifier is overwhelming *(3)*.

Some developmental studies avoid this bias by separating the data into a training set used for model development and a test set used for evaluating the predictive accuracy of the model. Although this split-sample approach is useful, it represents an inefficient use of the data in that the training set may be too small to develop an accurate classification model and the test set may be too small to provide an accurate estimate of prediction accuracy. Molinaro et al. *(4)* showed that various cross-validation approaches can provide better estimates of predictive accuracy. Such methods are based on repeatedly partitioning the sample into a relatively large portion that is used for classifier development and a small portion that is used for classifier evaluation and then averaging the results over the multiple partitions. In this case, the estimate of prediction accuracy pertains to the model that was developed using the full dataset, which is the model that will be used in future studies. The cross-validation procedure develops multiple classifiers based on reduced training sets only as steps in computing an estimate of prediction accuracy for the classifier developed using the full dataset.

In developing a classifier on a reduced data training set, the model development algorithm must be applied from scratch, without using any information based on data not part of that reduced data training set. This point is frequently overlooked by biomedical scientists, statisticians, and computer scientists. Asgharzadeh et al. were careful to use cross-validation methods properly to avoid the large biases that can result from incomplete cross-validation *(3,5)*. They used nested cross-validation to optimize the genes selected for the model and to estimate prediction accuracy. They then developed cross-validated Kaplan–Meier curves of progression-free survival and evaluated the statistical significance of the log-rank statistic used as a measure of the separation of the Kaplan–Meier curves by permutational methods, as initially described by Vaselli et al. *(6)*. Other authors have assessed the statistical significance of cross-validated Kaplan–Meier curves using the usual chi-square distribution, but that distribution is not valid for cross-validated curves because the independence assumption is violated.

Asgharzadeh et al. *(1)* also properly focused on evaluating their gene expression classifier by assessing its predictive accuracy relative to that of standard classifiers that are based on age and histology. Many studies, by contrast, make the mistake of focusing on the statistical significance of the estimated correlation between clinical outcome and predicted risk group. Often, the predicted risk group is used in a multivariable regression model with standard prognostic variables, and the conclusions are based on the statistical significance of the regression coefficients. Such measures of statistical significance are generally not only invalid *(7)* but also potentially misleading *(8)*. For studies in which the intent is to develop classifiers that can be used for predicting patient outcome, it is predictive accuracy and related characteristics such as sensitivity, specificity, and positive and negative predictive values of the classifier relative to predictions based on standard prognostic factors that matter, not the size of regression coefficients, their statistical significance, or whether one variable is an independent predictor of outcome *(9)*.

The article by Asgharzadeh et al. exhibits many good methodologic features that are missing from many other published reports. Most important, it reflects close collaboration between biomedical scientists and biostatisticians. The fundamental challenge in using genomic technology for enhancing therapeutics development and the development of predictive medicine is not just the management of the large volumes of data involved but also the complex issues involved in the design and analysis of appropriate studies. In a field that is filled with misinformation, hype, and inappropriate cynicism, it is essential for biomedical scientists to obtain the guidance and collaboration of biostatisticians. Most of the methods utilized by Asgharzadeh et al. are included in the BRB-ArrayTools software package, freely

Correspondence to: Richard Simon, DSc, Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, 9000 Rockville Pike, Bethesda, MD 20892-7434 (e-mail: rsimon@nih.gov).

DOI: 10.1093/jnci/djj364

available for noncommercial purposes from the National Cancer Institute at http://linus.nci.nih.gov/brb. BRB-ArrayTools is designed to be used by biomedical scientists to guide them in use of state-of-the-art methods for analysis of microarray gene expression data.

Even with all the good methodologic features of the study of Asgharzadeh et al., is their 55-gene classifier of progression risk ready for prime time use? I believe that the answer is no. Developmental studies should provide unbiased estimates of predictive accuracy, but such estimates are not a substitute for external validation. Validation studies should ideally establish the predictive accuracy and clinical utility of the classifier under conditions that simulate the prospective broad clinical application of the classifier. Asgharzadeh et al. used archived frozen specimens from patients previously treated on a variety of clinical trials and performed microarray assays at a single laboratory under controlled conditions. Consequently, their data do not reflect the many potential sources of variation in real-world conditions, with prospective tissue handling, assay drift, and reagent batch effects within an assay laboratory, as well as interlaboratory assay variation.

Buyse et al. *(2)* carried out a validation study of a 70-gene classifier for patients with early breast cancer. This classifier was previously developed by investigators at the Netherlands Cancer Institute *(10)* for a mixed population of node-negative and node-positive patients younger than 55 years of age who had received no systemic therapy. Buyse et al. properly recognized that the purpose of a validation study is to see whether a previously developed classifier accurately predicts and has clinical utility for an independent set of patients. The purpose of a validation study is not to see whether redeveloping a classifier with new data results in selection of the same genes, although this is one of the most common misunderstandings of validation. Indeed, gene expression classification studies have been widely and inappropriately criticized in the medical literature because different studies of the same disease result in classifiers with very divergent gene sets. What many critics fail to understand is that a set of genes is not a classifier. The classifier is a function that transforms the expression levels for the selected genes to a risk score or a predicted risk group. The expression of many genes is correlated, and hence it is to be expected that the genes selected for inclusion in a classifier will not be stable among studies. This instability is exacerbated by the very stringent significance levels that are generally used for identifying genes whose expression is correlated with patient outcome and are therefore included in the classifier. At these stringent significance levels, the statistical power of selecting a specific gene for inclusion is quite low. Consequently, one should not expect reproducibility in the gene sets selected in developing classifiers for different data sets. What matters, however, is whether a classifier provides accurate prediction for independent data *(11)*. Buyse et al. *(2)* applied the classifier to the new patients properly, without modifying it. Many investigators cannot resist the temptation to “improve” on previously reported classifiers. Unfortunately, this means that their studies are no longer proper validation studies and, themselves, require independent validation.

To be useful, a prognostic classifier must be therapeutically relevant. Many prognostic factor developmental studies use highly heterogeneous populations of patients for whom biological specimens are available. A classifier that is prognostic for such a mixed group of patients generally has very limited therapeutic relevance. The Netherlands Cancer Institute classifier was developed using patients who were heterogeneous with regard to nodal status and hormonal receptor status, but none of the patients had received any systemic chemotherapy. Buyse et al. limited their validation study to patients who were node negative, were of heterogeneous hormone receptor status, were less than 61 years of age at diagnosis, and had received no systemic therapy of any type.

Buyse et al. report that the 70-gene classifier was more predictive of risk of recurrence, risk of distant metastatic recurrence, and death from any cause than several of the currently used clinicopathologic prognostic systems, including that used by Adjuvant! software. Patients classified as having high risk of recurrence using the 70-gene classifier had a 10-year survival of 0.69, regardless of the risk group predicted by Adjuvant! software. Patients classified as having a low risk of recurrence using the 70-gene classifier had a 10-year survival of approximately 0.89, regardless of the risk group predicted by the Adjuvant! software.

The study by Buyse et al. has some limitations as a validation study, however. For example, the authors used archived frozen tumor specimens, had them assayed at a single reference laboratory, and limited eligibility to patients less than 61 years of age. These features may limit the real-world relevance of the results. The more important limitation, however, is that the study, despite the use of the phrase “clinical utility” in the title, did not actually evaluate the clinical utility of the 70-gene classifier. Establishing clinical utility means establishing that patient benefit is improved as a result of using the classifier.

Buyse et al. allude to the prospective MINDACT clinical trial that is under way to evaluate whether use of the 70-gene classifier is associated with clinical benefit. The most straightforward design for such a study would be to randomly assign a defined group of patients to have their treatment determined on the basis of either 1) the gene classifier or 2) standard practice guidelines. If the patients in the first group have better outcomes than those in the second—i.e., better tumor control or similar tumor control with fewer adverse events—then clinical utility is established. Such clinical trials generally require huge numbers of patients, however, because many of the patients in the two groups being compared will receive the same therapy. The MINDACT study attempts to improve the efficiency of the study by measuring the 70-gene classifier on all potential candidates but randomly assigning only those patients for whom practice guidelines and the gene classifier imply different treatments. Nevertheless, the study plans to prospectively accrue 6000 patients. Long follow-up will then be required before the results are evaluable.

Another approach to establish clinical utility is that used by Paik et al. *(12)*, who were able to provide evidence for the clinical utility of the Oncotype Dx classifier based on a retrospective study that developed a classifier of prognosis for node-negative, estrogen receptor (ER)–positive patients who received tamoxifen following local therapy for primary breast cancer. The Oncotype Dx classifier identified patients who had very low risk of recurrence on tamoxifen alone as systemic therapy. Such patients presumably do not require adjuvant treatment with cytotoxic chemotherapy, which represents clinical utility. In contrast, the study of Buyse et al. included both ER-negative and ER-positive patients and no patients received any systemic treatment. It is not clear whether the 70-gene profile is prognostic for ER-positive patients receiving tamoxifen. The 70-gene classifier does classify as high risk nearly all the ER-negative patients. However, the Adjuvant! classifier does so as well. Consequently, most of the patients randomly assigned in the prospective MINDACT trial to evaluate clinical utility of the 70-gene classifier will presumably have ER-positive tumors because few discrepancies in treatment selection would be expected for the ER-negative patients. It might have been advantageous, therefore, to have focused the development and validation of the gene classifier on ER-positive patients who were receiving tamoxifen.

The studies in this issue, taken together, illustrate many desirable features of gene expression profiling studies for optimizing treatment selection for individual patients. Clinical trial designs for using predictive classifiers in conjunction with the development of new drugs are very different than those described here, however *(13–16)*. Although the expression profiles utilized by Asgharzadeh et al. *(1)* and Buyse et al. *(2)* provide some biologic information about the nature of these cancers under study and potential molecular targets, it is important to recognize that the objective of predicting outcome for patients receiving a treatment is distinct from the objective of understanding the pathogenesis of the disease. Also, it is a mistake to criticize or reject a predictive classifier because its components are not seen as valid "disease biomarkers." Establishing a biomarker as a valid surrogate for measuring disease progression and clinical benefit is much more difficult than establishing that a classifier has clinical benefit for treatment selection. Because of the very distinct uses of the term "biomarker" and the very distinct kinds of "validation" that are appropriate, these terms should be used very carefully.

References

(1) Asgharzadeh S, Pique-Regi R, Sposto R, Wang H, Yang Y, Shimada H, et al. Prognostic significance of gene expression profiles of metastatic neuroblastomas lacking MYCN gene amplification. J Natl Cancer Inst 2006;98:1193–203.

(2) Buyse M, Loi S, van't Veer L, Viale G, Delorenzi M, Glas AM, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst 2006;98:1183–92.

(3) 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.

(4) Molinaro A, Simon R, Pfeiffer R. Prediction error estimation: a comparison of resampling methods. Bioinformatics 2005;21:3301–7.

(5) Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 2006;7:91.

(6) Vasselli J, Shih JH, Iyengar SR, Maranchie J, Riss J, Worrell R, et al. Predicting survival in patients with metastatic kidney cancer by gene expression profiling in the primary tumor. Proc Natl Acad Sci U S A 2003; 100:6958–63.

(7) Lusa L, McShane LM, Radmacher MD, Shih JH, Wright GW, Simon R. Appropriateness of inference procedures based on within-sample validation for assessing gene expression microarray-based prognostic classifier performance. Stat Med. In press 2006.

(8) Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic or screening marker. Am J Epidemiol 2004;159:882–90.

(9) Kattan MW. Judging new markers by their ability to improve predictive accuracy. J Natl Cancer Inst 2003;9:634–5.

(10) van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–6.

(11) Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DSA, Nobel AB, et al. Concordance among gene-expression-based predictors for breast cancer. New Engl J Med. In press 2006.

(12) 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.

(13) Simon R. Roadmap for developing and validating therapeutically relevant genomic classifiers. J Clin Oncol 2005;23:7332–41.

(14) Simon R, Maitnouram A. Evaluating the efficiency of targeted designs for randomized clinical trials. Clin Cancer Res 2005;10:6759–63.

(15) Freidlin B, Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clin Cancer Res 2005;11:7872–8.

(16) Simon R, Wang SJ. Use of genomic signatures in therapeutics development. Pharmacogenomics J 2006;6:166–73.