www.nature.com/clinicalpractice/onc

When is a genomic classifier ready for prime time?

Richard Simon

R Simon is Chief of the Biometric Research Branch at the National Cancer Institute, Bethesda, MD, USA.

DNA microarray technology is widely used to examine gene expression for the elucidation of biological mechanisms. Gene-expression profiles are now being used as classifiers of patients' prognosis and response to therapy.[1]

The development of traditional prognostic and diagnostic biomarkers has been largely disappointing. The extensive literature on this topic is often contradictory and relatively few oncologic biomarkers have been adopted in clinical practice.[2,3] One reason is that biomarker studies are generally performed retrospectively using heterogeneous specimens with no prespecified hypotheses. Generally, this approach does not provide an appropriate context for reliable development of a therapeutically relevant classifier. Many studies of primary breast cancer, for example, include node-negative and node-positive patients, some of whom are receiving chemotherapy. Such studies often contain numerous data-driven subset analyses, and are not conducted according to accepted standards for reliable prospective clinical trials of therapeutics.[4]

Gene-expression-profile classifiers are subject to additional potential problems that result from the number of candidate predictors being orders of magnitude greater than the number of cases available. In this situation, a model cannot be judged by its fit to the data used to develop it, which raises the issue of what kind of validation should be required.[5]

Internal validation is appropriate for the initial studies in which a genomic-profile classifier is developed. Such studies are often based on specimens from one institution, and the microarray assay is conducted in one laboratory. Consequently, the data might not reflect the full range of clinical variability in prognostic influences and tissue-handling procedures. Nevertheless, it is important to provide an unbiased estimate of the prediction accuracy of the classifier within this restricted context.

There are two types of unbiased internal validation: split-sample validation and cross validation. In both cases, a model-development algorithm is applied to a training dataset, and the resulting genomic-profile classifier is evaluated using a test dataset. For split-sample validation, t he division of available samples into a training set and a test set is often done at random, with a third to a half of the samples reserved for the test set. When the model is applied to the samples in the test set, no final adjustment of model parameters or selection among candidate models is permitted; the samples in the test set should be classified and the number of errors counted.

Cross validation involves repeatedly splitting the data into a training set containing most of the samples and a test set containing the remaining samples. The number of errors is determined for the predictions in the test set. This process is repeated for the many training-set and test-set partitions, and the resulting error rates observed for the test sets are averaged. The predictions for each test set are unbiased estimates of the true prediction error because the test-set samples are not used in the development of the model used for their prediction. The test sets are very small, but precision is improved by averaging error estimates over the repeated training-test partitions.

The many variants of cross-validation[6] can be used only when a well-defined model development algorithm is used for the classifier. Model-development algorithms are, however, complex; an algorithm is used to select genes to be incorporated in the model, to specify the type of predictive model, and to fit model parameters.[7,8] When used properly, cross validation can be more efficient than split-sample validation. Unfortunately, cross validation is often used incorrectly.[5] Correct cross-validation requires that the genomic-profile-classifier model be built from scratch for each training set. Many reports use the full set of data to preselect the genes to be used in building the model and then only partially cross validate the model, which strongly biases the results.[5]

The initial study that develops a genomic-profile classifier is analogous to a phase II

Correspondence
Biometric Research Branch
Division of Cancer
Treatment and Diagnosis
National Cancer Institute
9000 Rockville Pike
Bethesda
MD 20892
USA
rsimon@nih.gov

Received 15 July 2004
Accepted 10 August 2004
www.nature.com/clinicalpractice
doi:10.1038/ncponc0006

therapeutic study. It should provide a fully specified classifier and a preliminary estimate of the error rate of the classifier. However, that estimate will be limited in precision and accuracy because of the limited sample size of the original study, and because the variability seen in broad clinical application is not incorporated.[9] The classifier should not be adopted in clinical practice on the basis of such initial studies because previous experience indicates the possibility of overly optimistic results, even without the complexities of properly analyzing thousands of genes. Consequently, a prospectively planned phase III evaluation is required.

The external validation study should be a prospectively planned evaluation of a fully specified classifier. It should have prespecified hypotheses and endpoints, and a sample size sufficient to give precise estimates of sensitivity, specificity and positive and negative predictive values of the genomic-profile classifier.[10] The patients should be selected from a prospective multicenter clinical trial to ensure the prespecified therapeutic question is addressed, that results can be generalized to multiple locations and that tissue handling is not confounded with outcome. Genomic studies should generally address assay reproducibility for tumor heterogeneity, tissue handling and laboratory variation. The assays should be blinded to outcome to ensure lack of bias and lack of confounding of assay reagents or machine variability with outcome.

Ideally, external validation would encompass a prospective clinical trial designed to evaluate whether real-time collection of tissue, performance of the assay and use of the classifier results in improved therapeutic decision making. For example, a genomic-profile classifier for newly diagnosed patients with estrogen-receptor-positive breast tumors 1–4 cm in diameter, but without evidence of axillary node or distant metastases, might be validated in the following way. Patients classified as low risk based on a prespecified threshold for the classifier are randomized to receive tamoxifen alone after local treatment or systemic chemotherapy in addition to tamoxifen. Since chemotherapy is widely used for estrogen-receptor-positive patients, if chemotherapy provides no benefit to the low-risk subset, this classifier is beneficial for therapeutic decision making. Such a study would require thousands of patients.

A more limited form of external validation is possible by conducting a prospectively planned evaluation of the genomic-profile classifier based on archived tumor specimens from a randomized clinical trial. Many of the objectives of an external validation can be achieved in this way. Generally, however, archived specimens will not be available for all patients in a clinical trial and there may be concern about whether available specimens are representative.

Studies that develop genomic-profile classifiers should be viewed as phase II studies and should generally include a minimum of 50 patients. There are many pitfalls in the development of such classifiers; internal methods of validation should be used properly and the raw data made publicly available so that the analyses can be independently verified. The classifiers resulting from such studies should generally not be considered for clinical application until the results are externally validated. External validation should be based on evaluation of a completely prespecified classifier on patients from a multicenter clinical trial and should incorporate evaluation of assay reproducibility. Such studies generally require hundreds to thousands of patients. Although, many objectives of external validation can be achieved in prospective analysis of archived specimens from patients in large multicenter clinical trials, the gold standard for external validation involves real-time assaying of specimens from prospectively accrued patients.

Competing interests
The author declared he has no competing interests.

References

1 Rosenwald A *et al*. (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma. *N Engl J Med* **346:** 1937–1947
2 Hilsenbeck SG *et al*. (1992) Why do so many prognostic factors fail to pan out? *Breast Cancer Res Treat* **22:** 197–206
3 Hayes DF *et al*. (1998) Assessing the clinical impact of prognostic factors: when is "statistically significant" clinically useful? *Breast Cancer Res Treat* **52:** 304–319
4 Simon R and Altman DG (1994) Statistical aspects of prognostic factor studies in oncology. *Br J Cancer* **69:** 979–985
5 Simon R *et al*. (2003) Pitfalls in the analysis of DNA microarray data: class prediction methods. *J Natl Cancer Inst* **95:** 14–18
6 Hastie T *et al*. (2001) *The Elements of Statistical Learning*. New York: Springer
7 Simon R (2003) Diagnostic and prognostic prediction using gene expression profiles in high dimensional microarray data. *Br J Cancer* **89:** 1599–1604
8 Simon RM *et al*. (2003) *Design and Analysis of DNA Microarray Investigations*. New York: Springer
9 Covinsky JA and Berlin JE (1999) Assessing the generalizability of prognostic information. *Ann Intern Med* **130:** 515–524
10 Pepe MS (2003) The statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press