VOLUME 23 · NUMBER 29 · OCTOBER 10 2005

JOURNAL OF CLINICAL ONCOLOGY

REVIEW ARTICLE

Roadmap for Developing and Validating Therapeutically Relevant Genomic Classifiers

Richard Simon

From the Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, Bethesda, MD.

Submitted May 26, 2005; accepted June 18, 2005.

Terms in blue are defined in the glossary, found at the end of this issue and online at www.jco.org.

Author's disclosures of potential conflicts of interest are found at the end of this article.

Address reprint requests to Richard Simon, DSc, National Cancer Institute, 9000 Rockville Pike, MSC 7434, Bethesda, MD 20892; e-mail: rsimon@nih.gov.

0732-183X/05/2329-7332/$20.00

DOI: 10.1200/JCO.2005.02.8712

ABSTRACT

Oncologists need improved tools for selecting treatments for individual patients. The development of therapeutically relevant prognostic markers has traditionally been slowed by poor study design, inconsistent findings, and lack of proper validation studies. Microarray expression profiling provides an exciting new technology for relating tumor gene expression to patient outcome, but it also provides increased challenges for translating initial research findings into robust diagnostics that benefit patients and physicians in therapeutic decision making. This article attempts to clarify some of the misconceptions about the development and validation of multigene expression signature classifiers and highlights the steps needed to move genomic signatures into clinical application as therapeutically relevant and robust diagnostics.

J Clin Oncol 23:7332-7341.

INTRODUCTION

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 would not only save patients from unnecessary toxicity and inconvenience, but 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.

Microarray expression profiling has provided an exciting new technology for attempting to identify classifiers for tailoring treatments to patients. To date, however, no multigene expression signature has been widely adopted into oncology practice and very few are close to achieving such status. Development of biomarker classifiers useful for improving treatment decisions and sufficiently validated for broad clinical application is difficult, and more difficult for expression signature classifiers. The field of microarray expression profiling is also burdened with both unrealistic hype and excessive skepticism. In this article, I will attempt to clarify some of the misconceptions about the development and validation of multigene expression signature classifiers and highlight the steps needed to move genomic signatures into clinical application as therapeutically relevant and robust diagnostics.

WHY ARE SO FEW PROGNOSTIC FACTORS USED IN ONCOLOGY?

Although there is a large 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 relevance. Most prognostic factor studies are conducted using a convenience sample of patients for whom tissue is available, but the cohort is often far too heterogeneous with regard to stage and treatment to support therapeutically relevant conclusions. Additional problems in the prognostic marker literature derive from the fact that most studies develop prognostic

markers and prognostic models, but do not test prespecified models using independent data. Clinical drug trials are generally prospective, with patient selection criteria, primary end point, hypotheses, and analysis plan specified in advance in a written protocol. The consumers of clinical trial reports have been educated to be skeptical of data dredging to find something "statistically significant" to report in clinical trials. They are skeptical of analyses with multiple end points or multiple subsets, knowing that the chances of erroneous conclusions increase rapidly once one leaves the context of a focused, single-hypothesis clinical trial. Prognostic marker studies are generally performed with no written protocol, no eligibility criteria, no primary end point or hypotheses and no defined analysis plan. The analysis often includes numerous analyses of different end points and patient subsets. The problem is not just that the studies are for developing prognostic markers rather than validating previously specified markers, but that even as developmental studies the planning and analysis is relatively unfocused.[1]

Another feature that has hindered the use of prognostic markers in medical practice is the lack of studies demonstrating the reproducibility of results for assaying markers either between laboratories, between samples of the same tissue specimen, or between times and readers for the same laboratory.

Many of these problems apply to studies of prognostic classifiers on gene expression profiles. Some of the problems are even more formidable. Because of the number of genes available for analysis, microarray data can be a veritable fountain of false findings unless a structured approach to model development and validation is utilized.[2]

Some of the key steps in obtaining a classifier that is ready for "prime time" are listed in Table 1. These steps are discussed in the following sections. We have already discussed the importance of developing the classifier for a specific therapeutic decision problem and using cases relevant to that decision context. That is of key importance. There are, however, some well-defined therapeutic decision contexts where even accurate, reproducible, and well-validated classifiers are unlikely to be used widely. For example, consider the treatment of patients with advanced disease treated with a potentially curative treatment. A classifier for predicting the patients unlikely to respond to that therapy may not be widely used if there is no good alternative treatment. The classifier would have to have a very high negative predictive value in order to justify withholding a potentially curative therapy. It is important to evaluate carefully the context of therapeutic decision making if one wants to develop a classifier that has a sufficiently great chance of having clinical impact to warrant the large expense and time commitment required to achieve the other parts of Table 1.

Table 1. Key Steps in Development and Validation of Therapeutically Relevant Genomic Classifiers

Develop classifier for addressing a specific important therapeutic decision
Patients are sufficiently homogeneous and receiving uniform treatment so that results are therapeutically relevant
Treatment options and costs of mis-classification are such that a classifier is likely to be used
Perform internal validation of classifier to assess whether it appears sufficiently accurate relative to standard prognostic factors that it is worth further development
Translate classifier to platform that would be used for broad clinical application
Demonstrate that the classifier is reproducible
Independent validation of the completely specified classifier on a prospectively planned study

WHAT IS A MULTIGENE CLASSIFIER?

A multigene expression signature classifier is a function that provides a classification of a tumor based on the expression levels of the component genes. The classes are often good-risk or poor-risk, but classifiers can be defined to distinguish any set of classes for which a training set of cases exist for each class. The term "classifier" is somewhat over-restrictive because a multigene biomarker can be a function that provides a continuous risk score rather than a class identifier. Here we will use the term "classifier" however, because for validation purposes it is usually important that cutoff thresholds of a risk score be defined in advance.

Some people prefer the phrase "multigene biomarker" to "multigene classifier." This can lead to serious misunderstandings, however. A completely defined classifier can be used to select patients and stratify patients for therapy, and the clinical effectiveness of the classifier can potentially be validated. Specifying only the genes involved does not enable one to structure prospective clinical validation experiments in which patients are assigned or stratified in prospectively well-defined ways. Hence, one is forever correlating expression of individual genes against outcomes, but never evaluating the use of a defined diagnostic classifier that can be applied to patients. The gene sets identified as associated with outcome tend to be unstable because gene groups are correlated by co-regulation and the stringent criteria used for identifying differentially expressed genes results in reduced statistical power for gene selection. It is often much easier to develop a classifier that performs accurately than it is to identify exactly the optimal gene set.

The components of expression signature classifiers need not be valid biomarkers in the sense of the US Food and Drug Administration.[3] Those criteria require that the role of the biomarker be mechanistically understood and accepted as markers of disease activity. Such criteria are relevant for biomarkers used as surrogate end points but not for the components of expression

signatures used for tailoring treatments. It is, of course, desirable to understand the mechanistic relationship of the components of an expression signature, but the classifier can be validated without such understanding and clear biologic interpretation may be more difficult to achieve than accurate classification.[4]

The concept of "validation" has been problematic for the development of traditional disease biomarkers. Much of the confusion derives from attempting to define validation in an absolute sense. A much more pragmatic and productive approach is to focus on validation for a specified purpose. For example, an expression signature should be developed for the purpose of predicting outcome for a well-defined set of patients who receive a well-defined therapy. The signature classifier would be developed using data from such patients and would be validated for an independent set of such patients. The developmental study would identify the genes to be included in the classifier, usually by screening a much larger set of genes to find those whose expression is most correlated with outcome. The developmental study would also combine the genes into a completely specified classifier that can be used and potentially validated in a subsequent study. The validation does not consist of seeing whether the same genes are prognostic in the subsequent study. The validation should be focused on addressing whether the application of the previously defined classifier to a new set of patients results in clinical benefit. This is discussed further in a subsequent section.

DEVELOPING A GENOMIC CLASSIFIER

What Kinds of Classifiers Are Most Useful?

Many algorithms have been used effectively with DNA microarray data for class prediction. A linear discriminant is a function

$$l(\underline{x}) = \sum_{i \in F} w_i x_i$$

where x_i denotes the expression measurement for the ith gene, w_i is the weight given to that gene, and the summation is over the set F of features (genes) selected for inclusion in the classifier. For a two-class problem, there is a threshold value c that must be defined; a sample with expression profile defined by a vector $\underline{x}$ of values is predicted to be in class 1 or class 2 depending on whether $l(\underline{x})$ as computed from the equation is less than or greater than c.

Many kinds of classifiers used in the literature have the form shown in the preceding equation. They differ with regard to how the weights are determined. These classifiers include Fisher's linear discriminant analysis and diagonal discriminant analysis,[5] the compound covariate predictor of Radmacher et al,[6] the weighted voting method of Golub et al,[7] support vector machines with inner product kernel,[8] perceptrons,[9] and the naïve Bayes classifier for multivariate Gaussian distributions.[10]

When the number of genes (p) is greater than the number of cases (n), perfect separation of a training set is always possible with a linear classifier. In fact, there are an infinite number of linear classifiers that achieve perfect separation. That suggests that there may not be sufficient information in most datasets to effectively utilize nonlinear classifiers. Although complex nonlinear classifiers are popular, there is very little evidence that they perform any better than simpler methods.

In the study of Dudoit et al,[5] the simplest methods, diagonal linear discriminant analysis and nearest-neighbor classification, performed as well or better than the more complex methods. Nearest-neighbor classification is based on a distance function $d(\underline{x},\underline{y})$, which measures the distance between the expression profiles $\underline{x}$ and $\underline{y}$ of two samples. The distance function utilizes only the genes in the selected set of genes F. To classify a sample with expression profile $\underline{y}$, compute $d(\underline{x},\underline{y})$ for each sample $\underline{x}$ in the training set. The predicted class of $\underline{y}$ is the class of the sample in the training set that is closest to $\underline{y}$ with regard to the distance function.

Paik et al[11] used linear classifiers for predicting recurrence risk of patients with primary breast cancer. Paik et al identified 19 genes for inclusion in the classifier. These included five proliferation genes, four genes related to estrogen metabolism, two *Her2* genes, two genes related to tissue invasion, and three other genes. These genes were selected on the basis of their correlation with recurrence in a training set of data. The classifier was based on computing the average expression level for each gene group and then a weighted average of the gene group–specific averages. The genes not in the proliferation, estrogen, *Her2* or invasion groups were taken as members of singleton groups. The weights were determined to optimize prediction on the training set. The final component of the classifier determined based on the training set were two cutpoints for the weighted sum of gene expression in order to define groups with a low risk, intermediate risk, and high risk of recurrence.

How Many Genes Should Be Included in the Classifier?

Most classifiers do not use all of the genes whose expression is measured. Consequently, one step in developing a classifier is determining which genes to include; this is called feature selection. Using all of the genes means that all of the genes would have to be measured in the future for classification of new patients. That is particularly problematic if the classifier is going to be converted to a real-time reverse transcriptase polymerase chain reaction (RT-PCR) platform. Also, the number of genes that are actually differentially expressed between the classes (ie, "informative

genes") is usually small compared to the number of genes that are not differentially expressed ("noise genes"). Including too many noise genes can dilute the influence of the informative genes and reduce the accuracy of prediction. It also makes interpretation and future use of the predictor more difficult.

It is sometimes possible to distinguish very different cell types based on expression levels of a small number of genes. Even if such genes are not known a priori, they can be identified if they are very differentially expressed in the two cell types. This is often not the case for more difficult classification problems however. For these problems there may be a dozen or more differentially expressed genes, but the fold differences in expression may not be large and it may be difficult to identify these genes from among the thousands of noise genes. Omitting informative genes from a classifier has a greater deleterious effect on classification accuracy than does inclusion of noise genes, so long as the number of noise genes included is not too great. Consequently, in many cases accurate classifiers can be developed, but it is more difficult to develop such classifiers based on a very small number of genes.

INTERNAL VALIDATION OF A CLASSIFIER IN DEVELOPMENTAL STUDIES

It is useful to divide genomic classifier studies into developmental studies and validation studies. Developmental studies are the ones that first develop the classifiers and are analogous to phase II clinical trials. They should include an indication of whether the genomic classifier is promising and worthy of phase III evaluation. There are special problems in evaluating whether a genomic classifier is promising based on a developmental study, however. 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 classifiers that accurately classify the data on which they were developed even if there is no relationship between expression of any of the genes and outcome.[6] 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 for testing the model, or some form of cross validation based on repeated model development and testing on random data partitions. This internal validation should not, however, be confused with the kind of external validation of the classifier in a setting simulating broad clinical application.

Split-Sample Validation

The most straightforward method of estimating the accuracy of future prediction is the split-sample validation method of partitioning the set of samples into a training set and a test set. Rosenwald et al[12] used this approach successfully in their international study of prognostic prediction for large B cell lymphoma. They used two thirds of their samples as a training set. Multiple kinds of predictors were studied on the training set. When the collaborators of that study agreed on a single fully specified prediction model, they accessed the test set for the first time. On the test set there was no adjustment of the model or fitting of parameters. They merely used the samples in the test set to evaluate the predictions of the model that was completely specified using only the training data. In addition to estimating the overall error rate on the test set, one can also estimate other important operating characteristics of the test such as sensitivity, specificity, positive and negative predictive values.

The split-sample method is often used with so few samples in the test set, however, that the validation is almost meaningless. One can evaluate the adequacy of the size of the test set by computing the statistical significance of the classification error rate on the test set or by computing a confidence interval for the test-set error rate. Since the test set is separate from the training set, the number of errors on the test set has a binomial distribution.

Michiels et al[13] suggested that multiple training-test partitions be used, rather than just one. The split sample approach is mostly useful, however, when one does not have a well-defined algorithm for developing the classifier. When there is a single training set-test set partition, one can perform numerous unplanned analyses on the training set to develop a classifier and then test that classifier on the test set. With multiple training-test partitions however, that type of flexible approach to model development cannot be used. If one has an algorithm for classifier development, it is generally better to use one of the cross validation or bootstrap resampling approaches to estimating error rate because the split sample approach does not provide as efficient a use of the available data.[14] Some of the conclusions of Michiels et al about the inaccuracy of published expression profiles may be artifacts of their using inadequately small test sets.

Cross Validation

Cross validation is an alternative to the split sample method of estimating prediction accuracy.[6] Molinaro et al[14] describe and evaluate many variants of cross-validation and bootstrap resampling for classification problems where the number of candidate predictors vastly exceeds the number of cases. For illustration we will describe leave-one-out cross validation (LOOCV). LOOCV starts like split-sample cross validation in forming a training set of samples and a test set. With LOOCV, however, the test set consists of only a single sample; the rest of

the samples are placed in the training set. The sample in the test set is placed aside and not utilized at all in the development of the class prediction model. Using only the training set, the informative genes are selected and the parameters of the model are fit to the data. Let us call M_1 the model developed with sample 1 in the test set. When this model is fully developed, it is used to predict the class of sample 1. This prediction is made using the expression profile of sample 1, but obviously without using knowledge of the true class of sample 1. This predicted class is compared to the true class label of sample 1. If they disagree, then the prediction is in error. Then a new training set–test set partition is created. This time sample 2 is placed in the test set and all of the other samples, including sample 1, are placed in the training set. A new model is constructed from scratch using the samples in the new training set. Call this model M_2 . Although the same algorithm for gene selection and parameter estimation is used, since model M_2 is constructed from scratch on the new training set, it will in general not contain exactly the same gene set as M_1. After creating M_2, it is applied to the expression profile of sample 2, which was omitted. If this predicted class does not agree with the true class label of the second sample, then the prediction is in error. The process is repeated leaving each of the n biologically independent samples out of the training set, one at a time. During the steps, n different models are created and each one is used to predict the class of the omitted sample. The number of prediction errors is totaled and reported as the leave-one-out cross-validated estimate of the prediction error.

At the end of the LOOCV procedure, you have constructed n different models. They were constructed in order only to estimate the prediction error associated with the type of model constructed. The model that would be used for future predictions is one constructed using all n samples. That is the best model for future prediction and the one that should be reported in the publication. The cross-validated error rate is an estimate of the error rate to be expected in use of this model for future samples, assuming that the relationship between class and expression profile is the same for future samples as for the currently available samples. With two classes, one can use a similar approach to obtain cross-validated estimates of the sensitivity, specificity, and the negative and positive predictive values of the classification procedure. One could even estimate an entire receiver operating characteristics curve.

The cross-validated prediction error is an estimate of the prediction error associated with application of the algorithm for model building to the entire dataset. A commonly used invalid estimate is called the re-substitution estimate. You use all the samples to develop a model. Then you predict the class of each sample using that model. The predicted class labels are compared to the true class labels and the errors are totaled.

Simon et al[15] performed a simulation to examine the bias in estimated error rates for class prediction. Two types of LOOCV were studied: one with removal of the left-out specimen before selection of differentially expressed genes and one with removal of the left-out specimen before computation of gene weights and the prediction rule but after gene selection. They also computed the re-substitution estimate of the error rate. In a simulated dataset, 20 gene expression profiles of length 6,000 were randomly generated from the same distribution. Ten profiles were arbitrarily assigned to class 1 and the other 10 to class 2, creating an artificial separation of the profiles into two classes. Since no true underlying difference exists between the two classes class prediction will perform no better than a random guess for future biologically independent samples. Hence, the estimated error rates for simulated data sets should be centered around 0.5 (ie, 10 misclassifications of 20).

Figure 1 shows the observed number of misclassifications resulting from each level of cross validation for 2,000 simulated data sets. It is well known that the re-substitution estimate of error is biased for small data sets and the simulation confirms this, with an astounding 98.2% of the simulated data sets resulting in zero misclassifications even though no true underlying difference exists between the two groups. Moreover, the maximum number of misclassified profiles using the resubstitution method was only one.

Cross validating the prediction rule after selection of differentially expressed genes from the full data set does little to correct the bias of the re-substitution estimator: 90.2% of simulated data sets still result in zero misclassifications. It is not until gene selection is also subjected to cross validation that we observe results in line with our

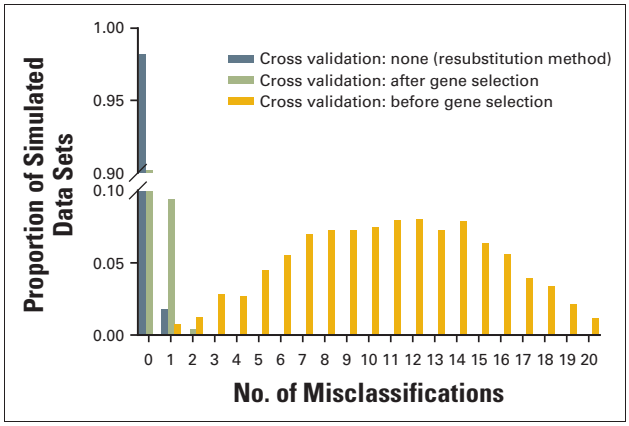

Fig 1. The effect of various levels of cross validation on the estimated error rate of a predictor. Two thousand datasets were simulated as described in the text. Class labels were arbitrarily assigned to the specimens within each dataset, and so poor classification accuracy is expected. Class prediction was performed on each dataset as described in the supplemental information, varying the level of leave-one-out cross validation used in prediction. Vertical bars indicate the proportion of simulated data sets (of 2,000) resulting in a given number of misclassifications for a specified cross-validation strategy. Reprinted from Simon et al.[15]

expectation: the median number of misclassified profiles jumps to 11, although the range is large (0 to 20).

The simulation results underscore the importance of cross validating all steps of predictor construction in estimating the error rate. A study of breast cancer also illustrates the point: van't Veer et al[16] predicted clinical outcome of patients with axillary node-negative breast cancer (metastatic disease within 5 years *v* disease free at 5 years) from gene expression profiles. The investigators controlled the number of misclassified recurrent cases (ie, the sensitivity of the test) in both situations, so here we focus attention on the difference in estimated error rates for the disease-free cases. Partial and complete cross validation resulted in estimated error rates of 27% (12 of 44) and 41% (18 of 44), respectively. The improperly cross-validated method results in a seriously biased underestimate of the error rate, probably largely due to overfitting the predictor to the specific dataset. Other examples of incorrect use of cross validation are described by Ambroise and McLachlan.[17] There are numerous articles in the most prominent journals, written by both biologists and methodologists, that make claims for gene expression classifiers and for new classification algorithms, which are invalid because they have cross validated improperly.

It is important to compute the statistical significance of the cross-validated estimate of classification error. This determines the probability of obtaining a cross-validated classification error as small as actually achieved if there were no relationship between the expression data and class identifiers. A flexible method for computing this statistical significance was described by Radmacher et al.[6] It involves randomly permuting the class identifiers among the patients and then recalculating the cross-validated classification error for the permuted data. This is done a large number of times to generate the null distribution of the cross-validated prediction error. If the value of the cross-validated error obtained for the real data lies far enough in the tail of this null distribution, then the results are statistically significant. This method of computing statistical significance of cross-validated error rate for a wide variety of classifier functions is implemented in the BRB-ArrayTools software (National Cancer Institute, Bethesda, MD).[18] Statistical significance, however, does not imply that the prediction accuracy is sufficient for the test to have clinical value, however.

DOES THE CLASSIFIER PERFORM BETTER THAN STANDARD PROGNOSTIC FACTORS?

Even if a classifier is developed for a set of patients sufficiently homogeneous and uniformly treated to be therapeutically relevant, it may be important to evaluate whether the classifier predicts more accurately than do standard prognostic factors or adds predictive accuracy to that provided by standard prognostic factors. For example, Rosenwald et al[12] developed a classifier of outcome for patients with advanced diffuse large B cell lymphoma receiving CHOP chemotherapy. The International Prognostic Index (IPI) is easily measured and prognostically important for such patients, however, and so it was important for Rosenwald et al to address whether their classifier provided added value.

The most effective way of addressing whether a classifier adds predictive accuracy to a standard classification system is to examine outcome for the new system within the levels of the standard system. This was the approach used by Rosenwald et al[12] for data in their separate test set. This is illustrated in Figure 2. The spread of the outcome survival curves for the classes defined by the new expression classifier within levels of the IPI indicate the extent to which the new system adds classification accuracy. When the classifier has been completely determined on a training set of data, then the statistical significance of the contribution of the new classifier to the standard IPI can be computed easily from a log-rank test using the test-set data.

Measuring whether a classifier adds predictive accuracy when there is not a separate test set is more difficult. Curves such as those shown in Figure 2 can be constructed using the predicted class of each case as determined by cross validation. The separation of the survival curves within levels of the standard prognostic factor is still a valid measure of the independent contribution of the expression classifier, but the statistical significance of the contribution can no longer be determined by computing a log-rank test of the separation in survival curves. The standard log-rank test is not valid because the classes were not determined independently of the data. The cross-validation process induces a dependence among cases that invalidates the standard statistical analysis. The statistical significance of the independent contribution of the new classifier can be determined using more complex permutation methods, however.[19]

Several important publications have attempted to determine the relative importance of an expression classifier and standard prognostic factors by using standard multivariate statistical models, such as the logistic model for binary response data and the proportional hazards model for survival data. The models often include standard prognostic factors and the predicted class of a case based on a cross-validation analysis.[16] Statistical significance and CIs for the regression coefficients corresponding to each factor are then computed using the usual formulas. This kind of analysis is problematic, however.[20] There is also a more fundamental problem with this kind of analysis. The value of an expression based classifier is determined by its prediction accuracy. Consequently, the analysis should emphasize estimating prediction accuracy, not the size of regression coefficients, in additive multivariate models.[21]

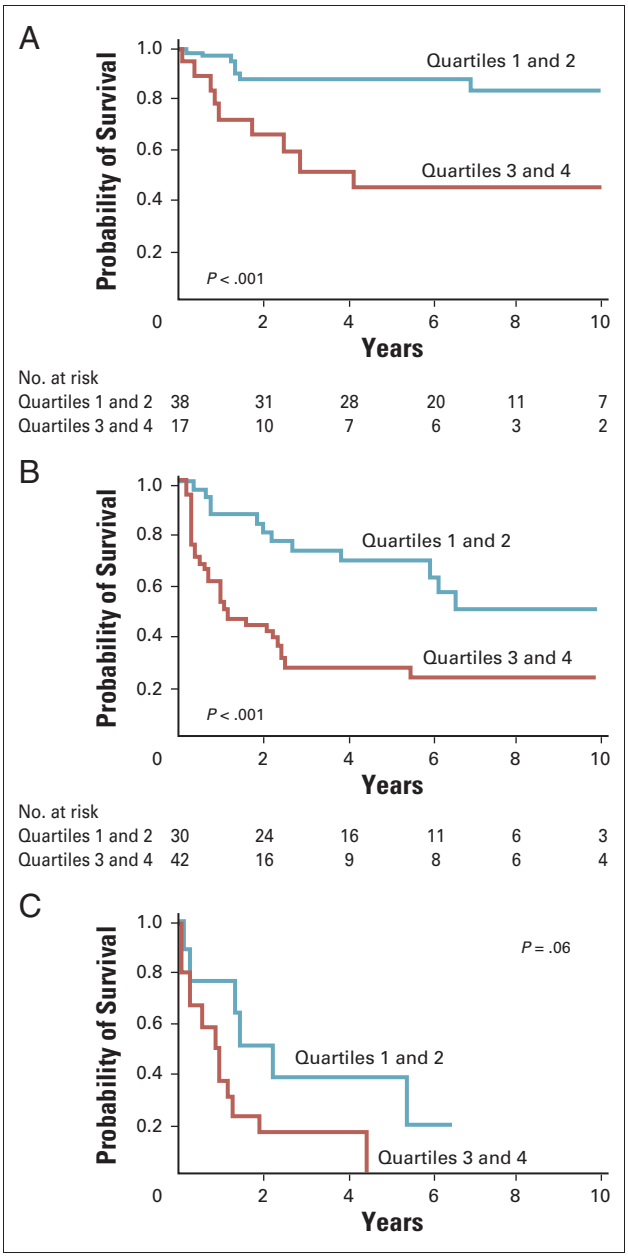

Fig 2. Survival curves for diffuse large-B-cell lymphoma patients by gene expression classifier stratified by three levels of International Prognostic Index (IPI) score: (A) IPI scores 0-1; (B) IPI scores 2-3; (C) IPI scores 4-5.Four prognostic classes were defined based on gene expression risk score. Graphs show survival curves for patients with risk score below the median (quartiles 1 and 2) versus patients with risk score above the median (quartiles 3 and 4). Reprinted from Rosenwald et al.[12]

TRANSLATION OF PLATFORMS AND DEMONSTRATING ASSAY REPRODUCIBILITY

The power of microarray expression profiling lies in the parallel measurement of expression levels for thousands of genes. This is useful for screening genes to find those that should be included in a classifier, but it is rarely necessary to measure expression for hundreds or thousands of genes in application of the classifier to subsequent cases. There are considerable challenges with microarray expression profiling of formalin-fixed paraffin-embedded (FFPE) tissue. With appropriately designed primers, however, RT-PCR can be performed on FFPE tissue.[22] Consequently, the developmental strategy of screening the genome using microarrays and then developing genomic classifiers based on a limited number of genes whose expression is measured using RT-PCR on FFPE tissue is potentially broadly applicable.

Whether the classifier is based on DNA microarray analysis or on RT-PCR analysis, it is important that the assay be standardized and that evaluations of reproducibility be conducted. The study by Dobbin et al[23] demonstrated that microarray protocols using Affymetrix arrays could be sufficiently standardized to achieve good inter- and intra-laboratory reproducibility. Achieving such reproducibility requires standardization of protocols and standardization of platform and reagents, however. One of the challenges in moving genomic classifiers to the clinic is the conduct of such studies. If a genomic classifier is used for identifying a patient population for which an experimental drug is shown to be effective, the drug sponsor has a financial incentive to adequately standardize and validate the classifier so that the classifier can be approved as a diagnostic test. In using genomic classifiers with commercially available therapy, however, it is not clear whether anyone has sufficient incentive to do the laborious but necessary studies needed to standardize and validate the reproducibility of the assay for measuring the classifier.

INDEPENDENT VALIDATION OF GENOMIC CLASSIFIERS

Although studies that develop classifiers often report a seemingly impressive accuracy for predicting outcome, there is abundant reason to demand external validation based on truly independent data. We refer to this as external validation because it is based on independent data external to the study used to develop the classifier. The analysis of high-dimensional gene expression data is complex and there are many examples of serious errors in internal estimates of accuracy included in publications in the best journals. There are also potential biases in internal estimates of accuracy based on tissue handling and assay reagent differences between cases and controls or responders and nonresponders. Developmental studies also often utilize patients selected in a manner that may not be representative of the diversity of patients to whom the classifier would be applied if it were adopted for broad clinical use. Developmental studies also often have the assay performed in one research laboratory based on archived specimens and this may not reflect the sources of assay variability likely to be encountered in broad practice.[24]

Often the initial study in which the classifier is developed will not be large enough to estimate the positive and negative predictive values of the test with sufficient

precision to determine whether the test has real clinical utility. It is important that the intended clinical use of the classifier be carefully considered in planning the external validation study so that these performance characteristics can be adequately estimated.

The objective of external validation is to determine whether use of a completely specified diagnostic classifier for therapeutic decision making in a defined clinical context results in patient benefit. The objective is not to repeat the developmental study and see if the same genes are prognostic or if the same classifier is obtained. An independent validation study could be a prospective clinical trial in which patients are randomly assigned to treatment assignment without use of the classifier versus treatment assignment with the aid of the classifier. Often, however, this design will be inefficient and require a huge sample size because many or most of the patients will receive the same treatment either way they are assigned. For example, consider women with lymph node-negative, estrogen receptor (ER) –positive breast cancers. Approximately one third of such patients might be expected to be classified as low risk for recurrence based on the Oncotype-DX expression signature–based risk score.[11] If one wants to test the strategy of withholding cytotoxic chemotherapy from the subset of patients classified as low risk, it would be inefficient to randomly assign all of the node-negative, ER-positive patients. If one randomly assigns all the patients and performs the assay on only the half assigned to have classifier based therapy, then the two randomization groups must be compared overall, although two thirds of the patients receive the same treatment in both arms. A more efficient alternative is to perform the assay up front for all patients, and then randomly assign only those classified as low risk. Those patients would be assigned to receive either tamoxifen alone or tamoxifen plus cytotoxic chemotherapy. If the low-risk patients do not benefit from cytotoxic chemotherapy, then the genomic classifier is clinically useful because it enables chemotherapy to be withheld from patients who otherwise would have received it.

Randomly assigning only the patients classified as low risk is more efficient than assigning all of the patients, but it still would require many patients. It is a therapeutic equivalence trial in the sense that finding no difference in outcome changes clinical practice; consequently it is important to be able to detect small differences. Since the expected recurrence rate is so low, it would take many patients to detect a difference between the treatment arms. But if the recurrence rate is as low as predicted by the classifier, then the benefit of chemotherapy is necessarily extremely small. Consequently, an alternative design for external validation is a single-arm study in which the patients classified as low risk are treated with tamoxifen alone. If, with long follow-up, these patients have a very low recurrence rate, then the classifier is considered validated for providing clinical benefit because it enabled the identification of patients whose prognosis was so good with tamoxifen monotherapy that they could be spared the toxicity, inconvenience and expense of chemotherapy. This was the approach used by Paik et al[11] for validation of the OncoType Dx classifier for patients with node-negative, ER-positive breast cancer. The genes that seemed prognostic were initially identified based on published microarray studies. Primers for measuring expression of those genes using RT-PCR of FFPE tissue were developed and a classifier was developed based on archived tissue from National Surgical Adjuvant Breast and Bowel (NSABP) studies. The completely prespecified classifier was then tested on 668 patients from NSABP B-14 who received tamoxifen alone as systemic therapy. Fifty-one percent of the assayed patients fell into the low-risk group. They had a distant recurrence rate at 10 years of 6.8% (95% CI, 4.0% to 9.6%). Much higher rates of distant recurrence were seen in the intermediate- and high-risk groups of the classifier (14.3% and 30.5%, respectively).

One might argue that treatment determination using a genomic classifier for women with stage I ER-positive breast cancer should not be compared with the strategy of administering to all such women tamoxifen plus chemotherapy, because there are practice guidelines available based on tumor size and age that withhold chemotherapy from some patients. Nevertheless, it would still be inefficient to randomly assign women to genomic classifier–determined therapy or nongenomic practice guidelines–determined therapy in which the genomic classifier is measured only on the women randomly assigned to its use. Most of the women will probably receive the same treatment in whichever arm they are assigned to. It is much more efficient to perform the assay for measuring the genomic classifier, and then randomly assign only the women for whom the two treatment strategies differ. The current plan for independently validating the classifier developed by van't Veer et al[16] for women with primary breast cancer utilizes this design strategy.

Phase III clinical trials generally attempt to utilize an intervention in a manner that it might be used if adopted in broad clinical practice. For evaluating a diagnostic classifier, a multicenter clinical trial provides the challenges of distributed tissue handling and real time assay performance that would be met in general use. The assays might be performed in multiple laboratories and cannot be batched in time with a single set of reagents as might be done in a retrospective study. Consequently, the prospective clinical trial is the gold standard for external validation of a genomic classifier.

External validation based on a new prospective clinical trial will require a long follow-up time for low-risk patients, however. In such circumstances it can be useful to conduct a prospectively planned validation using patients

treated in a previously conducted prospective multicenter clinical trial if archived tumor specimens are available for the vast majority of patients. The validation study should be prospectively planned with at least as much detail and rigor as for prospective accrual of new patients. Although assaying procedures probably cannot be distributed over time in the same way as for newly accrued patients, assay reproducibility studies should be conducted to demonstrate that the assay has been standardized and quality controlled sufficiently so that such sources of variation are negligible. A written protocol should be developed to ensure that the study is planned prospectively to evaluate the clinical benefit of a completely specified genomic classifier for a defined therapeutic decision in a defined population in a hypothesis testing manner as it would for a prospective clinical trial. The study of Paik et al[11] of the OncoType Dx classifier for women with node-negative, ER-positive breast cancer is an example of careful prospective planning of an independent validation study using archived specimens.

USE OF GENOMIC CLASSIFIERS IN NEW DRUG DEVELOPMENT

The objective of validation of a genomic classifier differs somewhat for existing therapy compared to an experimental therapy. With existing therapy, the emphasis should be on validation of the clinical benefit of using the classifier. With an experimental therapy, however, the emphasis should be on demonstrating effectiveness of the drug in a population identified by the classifier as being more likely to benefit. Simon and Maitournam[25] demonstrated that use of a genomic classifier for focusing a clinical trial in this manner can result in a dramatic reduction in required sample size, depending on the sensitivity and specificity of the classifier for identifying such patients. Not only can such targeting provide a huge improvement in efficiency in phase III development, it also provides an increased therapeutic ratio of benefit to toxicity and results in a greater proportion of treated patients who benefit.

Developing a genomic classifier of which patients are likely to benefit for targeting phase III trials may require larger phase II studies. This depends on the type of drug being developed. For example, if the drug is an inhibitor of a kinase mutated in cancer, then there is a natural diagnostic and no genome-wide screening is needed. Similarly, in the comparison of trastuzumab plus chemotherapy to chemotherapy alone in chemotherapy-naïve and -refractory metastatic breast cancer patients,[26,27] cases with less than a 2+ level of expression of the Her2/*neu* protein were excluded. In the development of gefitinib, had the phosphorylation domain of the *EGFR* gene been sequenced in responders and nonresponders on phase II trials of non–small-cell lung cancer patients, mutation status could have been used in focusing the phase III trials.[28,29] For many molecularly targeted drugs, however, the appropriate assay for selecting patients is not known, and development of a classifier based on comparing expression profiles for phase II responders versus phase II nonresponders may be the best approach. In such instances, one may not have sufficient confidence in the genomic classifier developed in phase II to use it for excluding patients in phase III trials. It may be better in this case to accept all conventionally eligible patients, and use the classifier to define a single subset analysis for the patients predicted to be most responsive to the new drug. The overall null hypothesis for all randomly assigned patients is tested at the .04 significance level. A portion 0.01 of the usual 5% false-positive rate is reserved for testing the new treatment in the subset predicted by the classifier to be responsive. This analysis strategy provides sponsors an incentive for developing genomic classifiers for targeting therapy in a manner that does not unduly deprive them of the possibility of broad labeling indications when justified by the data.

CONCLUSIONS

Oncologists need improved tools for selecting treatments for individual patients. The genomic technologies available today are sufficient to develop such tools. There is not broad understanding of the steps needed to translate research findings of correlations between gene expression and prognosis into robust diagnostics validated to be of clinical benefit. This article has attempted to identify some of the major steps needed for such translation. Many of these steps are not easy, nor cheap. For therapeutic decision settings of sufficient importance, attention should be devoted to establishing a means of funding and expeditiously carrying out these steps.

Author's Disclosures of Potential Conflicts of Interest

The authors indicated no potential conflicts of interest.

REFERENCES

1. Simon R, Altman DG: Statistical aspects of prognostic factor studies in oncology. Br J Cancer 69:979-985, 1994

2. Simon RM, Korn EL, McShane LM, et al: Design and analysis of DNA microarray investigations. New York, NY, Springer, 2003

3. FDA: Draft guidance for industry: Pharmacogenomics data submission. Rockville, MD, Food and Drug Administration, 2003

4. Pusztai L, Hess KR: Clinical trial design for microarray predictive marker discovery and assessment. Ann Oncol 15:1731-1737, 2004

5. Dudoit S, Fridlyand J, Speed TP: Comparison of discrimination methods for classification of tumors using gene expression data. J Am Stat Assoc 97:77-87, 2002

6. Radmacher MD, McShane LM, Simon R: A paradigm for class prediction using gene expression profiles. J Comput Biol 9:505-511, 2002

7. Golub TR, Slonim DK, Tamayo P, et al: Molecular classification of cancer: Class discovery and class prediction by gene expression modeling. Science 286:531-537, 1999

8. Ramaswamy S, Tamayo P, Rifkin R, et al: Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci USA 98:15149-15154, 2001

9. Khan J, Wei JS, Ringner M, et al: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine 7:673-679, 2001

10. Hand DJ, Yu K: Idiot's Bayes: Not so stupid after all? Int Stat Rev 69:385-398, 2001

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 351:2817-2826, 2004

12. Rosenwald A, Wright G, Chan WC, et al: The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med 346:1937-1947, 2002

13. Michiels S, Koscielny S, Hill C: Prediction of cancer outcome with microarrays: A multiple random validation strategy. The Lancet 365:488-492

14. Molinaro AM, Simon R, Pfeiffer RM: Prediction error estimation: A comparison of resampling methods. Bioinformatics 2005 (in press)

15. Simon R, Radmacher MD, Dobbin K, et al: Pitfalls in the analysis of DNA microarray data: Class prediction methods. J Natl Cancer Inst 95:14-18, 2003

16. van't Veer LJ, Dai H, Vijver MJVD, et al: Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530-536, 2002

17. Ambroise C, McLachlan GJ: Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci U S A 99:6562-6566, 2002

18. Simon R, Lam AP: BRB-ArrayTools (Version 3.3). Bethesda MD, Biometric Research Branch, National Cancer Institute, http://linus.nci.nih.gov/brb

19. Vasselli J, Shih JH, Iyengar SR, 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 100: 6958-6963, 2003

20. Lusa L, McShane LM, Radmacher MD, et al: Appropriateness of inference procedures based on within-sample validation for assessing gene expression microarray-based prognostic classifier performance. (Submitted for publication), 2005

21. Kattan MW: Judging new markers by their ability to improve predictive accuracy. J Natl Cancer Inst 95:634-635, 2003

22. Cronin M, Pho M, Dutta D, et al: Measurement of gene expression in archival paraffin-embedded tissues: development and performance of a 92-gene reverse transcriptase-polymerase chain reaction assay. Am J Pathol 164:35-42, 2004

23. Dobbin K, Beer DG, Meyerson M, et al: Inter-laboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. Clin Cancer Res 11: 565-572, 2005

24. Simon R: When is a genomic classifier ready for prime time? Nat Clin Pract Oncology 1:4-5, 2004

25. Simon R, Maitournam A: Evaluating the efficiency of targeted designs for randomized clinical trials. Clin Cancer Res 10:6759-6763, 2004

26. Baselga J: Herceptin alone or in combination with chemotherapy in the treatment of HER2-positive metastatic breast cancer: Pivotal trials. Oncology 61:14-21, 2001

27. Eiermann W: Trastuzumab combined with chemotherapy for the treatment of HER2-positive metastatic breast cancer: Pivotal trial data. Ann Oncol 12:S57-S62, 2001

28. Lynch TJ, Bell DW, Sordella R, et al: Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med 350:2129-2139, 2004

29. Paez JG, Janne PA, Lee JC, et al: EGFR mutations in lung cancer: Correlation with clinical response to gefitinib therapy. Science 304:1497-1500, 2004