A Checklist for Evaluating Reports of Expression Profiling for Treatment Selection

Richard Simon, DSc

Dr. Simon is Chief, Biometric Research Branch in the Division of Cancer Treatment & Diagnosis at the National Cancer Institute in Bethesda, Md.

Address correspondence to:
Richard Simon, DSc, Biometric Research Branch, Division of Cancer Treatment & Diagnosis, National Cancer Institute, 9000 Rockville Pike, MSC 7434, Bethesda, MD 20892-7434; Tel: 301-496-0975; E-mail: rsimon@nih.gov.

Abstract: Oncologists need improved tools for selecting the right treatment for a given patient because many patients do not benefit from administered therapies. The use of expression profiling of tumors has increased dramatically and many claims are made for the value of expression signatures in treatment selection. It is difficult, however, for oncologists to critically evaluate published results in this technology- and statistics-intensive field. A checklist is presented to help oncologists evaluate publications on expression profiling of human tumors to determine whether the results are ready for use with their patients.

Keywords
Biomarkers, expression profiling, pharmacogenomics, predictive factors

The use of microarrays for profiling gene expression in tumors has increased dramatically. Numerous studies are conducted to study basic biological mechanisms using experimental tumor models, to identify new therapeutic targets, and to identify prognostic factors or factors predictive of response to a given treatment. This article is presented in an effort to help hematologists and oncologists evaluate clinical reports of expression profiling to determine whether the results can help in treatment selection. Selected key issues for evaluating publications of the prognostic and predictive methods used in studies will be addressed. These issues are particularly relevant for patient classifiers based on high-dimensional data such as gene expression profiling. A checklist of key issues is presented in Table 1. The list of items has been purposely kept short so as to be useful for physicians who are not greatly interested in technical details but are looking for keys to help them evaluate whether a given report appearing in an important journal is applicable to their clinical practice. Additional recommendations and more detailed descriptions are available in previously published material.[1-7]

Biomarkers, Genomic Classifiers, and Predictive Indices

Biological measurements used to inform treatment selection are sometimes called biomarkers, but the term invites misinterpretation. Many people think of biomarkers as measures of disease activity, increasing as the disease progresses and decreasing as the disease responds. Such disease biomarkers would have considerable utility as surrogate endpoints for clinical trials. Consequently, a great deal

Table 1. Checklist of Some Key Issues

- Does the study provide a completely specified classifier or predictive index or does it just identify biological measurements correlated with outcome?
- Is the study a developmental or validation study?
- Does it develop a classifier or use a previously developed classifier?
- Are patients sufficiently homogeneous to be therapeutically relevant?
- Were patients enrolled in one clinical trial?
- Does the study address prognosis or response to therapy?
- Does the study address predictive accuracy or clinical utility?
- Is the patient outcome measure clinically relevant?
- Are alternative treatments considered?
- Are standard prognostic/predictive factors considered?
- Does the study provide information about assay reproducibility?
- Were there procedures to avoid bias from confounding tissue handling or assay drift with patient outcome?
- Are there obvious statistical flaws?
  - Use of cluster analysis?
  - Use of multivariate analysis?
  - Lack of statistician as coauthor?
- For developmental studies that use a cross-validation strategy that repeatedly partitions the data into training and test sets:
  - Is the model built from scratch for each training set, including gene selection, or were the genes preselected using all the data?
- Does the study provide at least 20 patients per class (eg, 20 responders and 20 nonresponders) for training set development of the classifier?
- Does the study demonstrate that the prediction accuracy is statistically significantly better than chance?
- Does the study provide confidence limits for the misclassification rate, sensitivity, specificity, and positive and negative predictive values?

of money and time is spent in attempting to identify such measures. The US Food and Drug Administration (FDA) is, of course, very concerned about the basis for accepting a surrogate endpoint. Consequently, stringent criteria have been established for "validating" such biomarkers.[8,9] Generally speaking, it is very difficult to establish that a biological measurement is a valid disease biomarker, and the time required to do so may be much longer than the time required to develop a new drug without such a surrogate. The focus here is not on surrogate endpoints, and thus it would be best to omit the confusion engendered by using the term biomarker. Biological features, such as gene expression levels, protein expression levels, and the presence or absence of gene mutations, polymorphisms, or amplifications, do not need to be "valid disease biomarkers" to be tremendously valuable for treatment selection. Similarly, the criteria that have been developed by the FDA and by academic researchers for establishing the validity of disease biomarkers are not relevant for pretreatment biological measurements to be used in treatment selection.

Sometimes multiple biological measurements are combined to give one composite measurement for use in treatment selection. This is particularly true for gene expression studies. Combined gene expression measurements generally are more effective than single gene expression measurements for distinguishing the patients who respond to a treatment from those who do not respond. The component measurements are combined in a defined way, such as a weighted average of log expression levels, with the weights determined based on expression profiles for a sample of responders and a sample of nonresponders. A wide variety of methods for combining the component measurements have been studied, but the details are beyond the scope of this article.[4,10] The combination can either be represented as a continuous "predictive index," which indicates the likelihood of response, or as a two-level or multilevel classifier, which indicates whether the patient is likely to respond, unlikely to respond, or intermediate. Binary classifiers are particularly useful because they incorporate a treatment strategy that can be validated: treat the patients who are predicted to be responsive to the therapy under consideration with that therapy, and treat the others with some other therapy. Classifiers and predictive indices of risk of recurrence can also be derived with disease-free survival data; it is not necessary to convert such data into binary outcomes in order to develop predictors of recurrence risk.[11,12]

Developmental or Validation Study?

In trying to determine whether the results of a published study are ready for application to patients, it is important to distinguish developmental studies from validation studies. A developmental study is one that develops a classifier whereas a validation study uses a classifier developed previously. One reason the distinction is important is because it is problematic to use the same data for developing a classifier and for evaluating it, particularly for high-dimensional assays. In developing a classifier, one selects the biological measurements that best fit the patient outcomes. If one measures expression of 20,000 genes, there are many variables to select from. Because the authors have optimized the classifier to the data, and therefore the degree of fit of the classifier to that data is not an adequate measure of predictive accuracy

for independent data. Simon and colleagues[13] showed that even with completely random data in which there is no true difference in expression profile between responders and nonresponders, it is always possible to find a classifier that perfectly distinguishes the responders from the nonresponders. Consequently, it is best to have a completely independent set of data on which to validate the classifier derived in a developmental study.

The developmental study should report a completely specified classifier. The study should not identify only the genes that are differentially expressed between responders and nonresponders, but it should also combine those genes in a defined manner and use the data to determine the weights, cut-off values, and any other parameters needed for subsequent application of the classifier. In the validation study the completely specified classifier is applied without change to a new set of patients.

The developmental study is analogous to a phase II clinical trial in the sense that it attempts to optimize the study agent and identify the patient population for which it is promising. The developmental study should provide an estimate of whether the classifier is promising enough to warrant a validation study because good validation studies are expensive. Naively using the same data to develop a genomic classifier and to test it is unsatisfactory. Such estimates are called "resubstitution" estimates and are extremely biased.

Several methods have been developed for using the data from the developmental study to determine whether the classifier is sufficiently promising to warrant external validation. The simplest method is called the "split-sample" method and consists of partitioning the data in the developmental study into two parts.[14] The separation is often done randomly, with either half of the cases in each group or two thirds of the cases in the "training data" used for developing the classifier and one third of the cases in the test set. The cases in the test set should not be used in any way until a single completely specified model is developed using the training data. At that time the classifier is applied to the cases in the test set. For example, with an expression profile classifier, the classifier is applied to the expression profiles of the cases in the test set and each of them is predicted (ie, classified) as a responder or nonresponder to the therapy. The patients in the test set have received the treatment in question and so one can count how many of those predictions were correct and how many were incorrect. Hence, one has an estimate of prediction accuracy of the classifier developed in the training set while avoiding the bias of using exactly the same data for developing the classifier and evaluating it. This split-sample method was used effectively by Rosenwald and coworkers[15] in developing a classifier for predicting outcome for patients with large B-cell lymphoma receiving cyclophosphamide/doxorubicin/vincristine/prednisone chemotherapy.

There are other, more complex forms of dividing data into training and testing portions. Many of these alternatives are called cross-validation methods and utilize data more efficiently than the simple division described above.[14,16] The split-sample method and the cross-validation methods are useful for providing a preliminary phase II–type estimate of the promise of the classifier for predicting patient outcome. Unfortunately, the cross-validation methods are often used incorrectly, resulting in very biased estimates of predictive accuracy. Cross-validation generally partitions the data into a large training set and a small test set. A model is developed based on the training set and then applied to the cases in the test set to estimate the error rate. This is repeated for numerous training-test partitions and the prediction error estimates are averaged. In order to preserve the key dictum of not using the same data to develop a model and to evaluate the model, it is important that for each training-test partition the data in the test set are not used in any way. Hence a model should be developed from scratch in each training set.[14] This means that multiple classifiers are developed in the process of doing cross-validation and those classifiers will in general involve different genes. It is invalid to select the genes beforehand using all the data and then simply to cross-validate the model-building process for that restricted set of genes. Simon and coauthors[13] as well as Ambroise and McLachlan[17] demonstrated that such preselection results in severely biased estimates of prediction accuracy. This error is made in many developmental classifier studies.

For a variety of reasons, simply using a split-sample or cross-validation analysis does not make a developmental study a validation study. There are many factors that may influence the predictive accuracy of a classifier that are not represented by artificially subdividing the cases from a single study. These factors include differences in patients from different centers, the nature of their diseases and prior treatments, differences in tissue handling, and differences in assay performance over time and location. Developmental studies are often conducted based on specimens available at one or a very limited number of centers, and the results may not be applicable to patients more generally. Developmental studies also often have all assays performed at one time in one research laboratory and may not reflect important sources of variation involved in real-world sample collection, tissue handling, and assay performance. Following the performance of a successful developmental study, it is often appropriate to address whether the original assay platform is suitable for broad application of the classifier. If not, then a recalibration of the classifier for its new platform is necessary before conducting the validation study. Dobbin and colleagues[18]

reported that in order to ensure good interlaboratory reproducibility in using the Affymetrix GeneChip system, a pilot study and development of a common protocol was necessary. In classifying the risk of recurrence for patients with node-negative and estrogen receptor–positive breast cancer receiving tamoxifen, Paik and colleagues[12] utilized DNA microarray gene expression profiling to identify the informative genes, but then transferred to a reverse transcription polymerase chain reaction platform based on primers for use with paraffin-embedded formalin-fixed tissue. They performed detailed studies on sources of variation of the assay in order to assure reproducibility of results.

Validation studies also differ from artificially subdivided developmental studies in that they should also address clinical utility of the classifier, not just predictive accuracy. This aspect is further discussed below.

Does the Study Address a Therapeutically Meaningful Set of Patients?

The field of oncology already has too many prognostic factors.[6,19] Most prognostic factor studies are conducted based on convenience samples of available specimens. Consequently they often include a heterogeneous group of patients who have received a variety of treatments.[20] For example, many prognostic factor studies in breast cancer include node-negative and -positive, estrogen receptor–negative and –positive patients, those who received cytotoxic chemotherapy, and those who received tamoxifen alone. Showing that a new classifier is prognostic for such a mixed group generally has no therapeutic value and such classifiers are rarely used.[21] It does not matter whether one shows, from a multivariate analysis, that the new classifier is more statistically significant than standard prognostic variables; therapeutic strategies have often been developed based on the established variables.

Although we usually do not need any more prognostic factors, we do need predictive factors—that is, biological measurements and classifiers that identify which patients respond to specific treatments. Predictive factors are needed because we often overtreat the majority of patients in hope of benefiting the minority. For example, if 85% of node-negative, estrogen receptor–positive women with breast cancer are cured with tamoxifen alone following local treatment, and if adding cytotoxic chemotherapy increases the cure rate to 90%, then unless we tailor the use of chemotherapy, we will be overtreating 85% of patients for the 5% who do benefit. Similarly, many molecularly targeted cancer drugs are very expensive and benefit only a small proportion of patients. Unless we can identify which patients have tumors that are driven by the pathways inhibited by such drugs, there will be pressure to overtreat patients, with serious detrimental consequences for both the patients and society.

Developing classifiers for predicting which patients respond to new drugs can dramatically improve the efficiency of clinical trials for establishing the effectiveness of the drugs. This was shown theoretically by Simon and Maitournam.[22,23] Targeted development of trastuzumab (Herceptin, Genentech) is a practical example of the effectiveness of this approach and recent experience with trastuzumab in women with node-positive breast cancer has shown how the use of classifiers for selecting the right drug for the right patient can dramatically mitigate the trade-off between effectiveness and toxicity in cancer therapeutics; that is, a much greater proportion of patients who receive a potentially toxic drug actually benefit.

It is very desirable for classifier development and validation to use patients who received a treatment in a single clinical trial because it helps ensure that the classifier developed is a therapeutically relevant predictive classifier, not just a prognostic factor. Both developmental studies and validation studies should address predictive classifiers for therapeutically meaningful sets of patients. Often, however, they do not. This is a very serious deficiency that cannot be overcome with complex data analysis.

Does the Study Address Predictive Accuracy or Clinical Utility?

Suppose that an excellent developmental study identifies a classifier that, based on a split-sample analysis, seems to predict accurately which patients will respond to a specified chemotherapy. Does that classifier have clinical utility? The answer depends on a variety of factors including the status of the treatment, the other treatments available for those patients, the availability of other more easily measured predictive factors, and the clinical relevance of tumor response for that stage of that disease. In general, establishing clinical utility requires demonstrating that a clinically meaningful measure of patient benefit is improved based on using the new classifier compared to not using the classifier. The design of a validation study should be based on establishing clinical utility, and this generally cannot be accomplished by artificially splitting the data in a retrospective set of patients used to develop a classifier.[5,6,24,25]

Is There Potential Bias in Tissue Handling and Assay Performance?

The results of some genomic and proteomic assays are distorted by differences in tissue handling and assay drift. In developing a classifier of responders versus nonresponders, it is essential that the specimens for the responders and

nonresponders be handled and assayed in the same way. The same is true for developing classifiers to be used for early cancer detection. It is not adequate to obtain the samples of one class from one institution and those of the other class from a different institution. It is also not adequate to assay the samples of one class at one time and those from the other class at another time. If there are too many samples to analyze on the same day, then the assays should be interleaved in time, each day analyzing some from both classes.

Does the Developmental Study Use Cluster Analysis?

Cluster analysis is a body of exploratory algorithms used to group genes or samples in expression profiling studies. It is not an appropriate approach for developing predictive classifiers because it does not properly utilize the patient outcome data. If all genes are used in clustering the samples, the clusters of samples corresponding to patients are created without reference to the outcome data. This is generally an ineffective way of developing a classifier of outcome. Often the samples are clustered with genes whose selection is based on their correlation with patient outcome. This approach generally produces misleading results. Since all of the data are used to select the genes, the resulting clusters of samples will have to separate the patients with regard to outcome, even if there is no real relationship between expression profile and outcome. With 20,000 genes, there will be 1,000 genes on average having a statistically significant ($P<.05$) correlation with outcome just as the false-positives of a huge number of significance tests. If the samples are then clustered with these genes, the clusters will spuriously differ with regard to outcome. Unfortunately, this is a very common method used in the oncology literature. Cluster analysis is thus frequently used in a misleading way to demonstrate the discriminatory power of selected genes. Consequently, use of cluster analysis in classifier development is often an indicator that the study is statistically deficient.

Does the Developmental Study Use Multivariate Analysis?

Multivariate analysis is often used in developmental studies to support the claim that the new classifier is more important than standard prognostic/predictive factors. Often this is done because the patients are too heterogeneous and not therapeutically relevant. A multivariate analysis, however, is an inadequate solution to the problem. If the cases selected are too heterogeneous to be therapeutically relevant, it is better to analyze homogeneous subsets separately than to perform a multivariate analysis. If there are established more easily measured prognostic factors that can be used to classify therapeutically homogeneous subsets of patients, it is much better to evaluate whether the new classifier is predictive of outcome within the levels of the standard factors than it is to do a multivariate analysis.[19,20,26,27] Multivariate analysis does not address predictive accuracy, which is the endpoint of concern in developmental studies. Validation studies should generally be designed to evaluate clinical utility of the new classifier relative to that achievable with standard measures.

Is the Sample Size Sufficient?

There are no established standards for sample size for studies that develop classifiers or predictive indices based on high-dimensional assay data. As a general rule, Dobbin and Simon[28] (also R.S. and K. Dobbin, PhD, unpublished data, 2006) recommend a minimum of 20 responders and 20 nonresponders in a training set for developing a classifier identifying the patients likely to respond. If the study is developing a classifier for risk of recurrence, there should be at least 20 patients who recur in the training set. The test set should be at least as large as the training set. For developmental studies using a split-sample approach, the classifier is developed in the training set and applied to the patients in the test set. The test set provides an estimate of the classification error rate, which is the number of responders classified as nonresponders plus the number of nonresponders classified as responders divided by the number of patients in the test set.

The study should also provide the sensitivity, specificity, and positive and negative predictive values in the test set. The positive predictive value is the number of responders divided by the number of patients predicted to be responders, and the negative predictive value is the number of nonresponders divided by the number of patients predicted to be nonresponders in the test set. The sensitivity is the number of patients predicted to respond divided by the number of responders in the test set and the specificity is the number of patients predicted to not respond divided by the number of nonresponders in the test set. Confidence intervals can be computed for all of these measures in the test set, and one can also compute the statistical significance of the test set error rate to establish that it is better than could be achieved without using the classifier at all. Most of these quantities (except for the confidence intervals) can also be computed if a cross-validation procedure is used rather than a simple split-sample method.[29]

Conclusion

Two of the greatest needs in oncology therapeutics development are better molecular targets and better methods for matching the right treatment to the right patient. Uti-

lizing genomic technology for identifying the key molecular targets is currently difficult. For example, expression profiling has not yet been broadly effective in identifying good therapeutic targets. In an expanding tumor cell population, there are many genes that are overexpressed and underexpressed relative to normal tissue. Many genes are differentially expressed in a subset of the tumor cells, particularly in a genetically unstable tumor expanding and invading under host selection pressures. It has been difficult to use expression profiling to discover the initially mutated genes that are essential to oncogenesis and that drive the growth, invasion, and development of secondary genomic and epigenetic effects. Progress in cancer therapeutics is primarily limited by lack of identification of these key molecular targets.[30] The pharmaceutical and biotechnology industries are very effective at developing inhibitors of defined targets, and there is a large infrastructure available to perform good clinical trials to evaluate such inhibitors. The numerous disappointments in drug development over the years are largely attributable to lack of stringency in credentialing therapeutic targets.

The current level of technology is sufficient, however, to develop effective tools for utilizing currently available therapy more appropriately and for more efficiently developing new drugs that inhibit the current molecular targets. Genomics and technology development are making available powerful tools for achieving this objective, but there are difficulties in effectively utilizing these tools. There is a voluminous oncology literature on prognostic factors that are not used and have not benefited therapeutic decision making. Properly developing and validating therapeutically relevant predictive classifiers is much more difficult and costly than the conventional prognostic analysis of a convenience sample of available specimens. The development and validation of therapeutically relevant predictive classifiers demands effective collaboration among experienced investigators with backgrounds in oncology medicine, biostatistics, and tumor biology. It also demands effective collaboration between academia and industry, and effective interaction with the scientists at the FDA. Such classifiers, however, can be of great value to patients, can better assure that more patients actually benefit from administered treatment, and can help alleviate our healthcare financing crisis.

References

1. Gasparini G, Pozza F, Harris AL. Evaluating the potential usefulness of new prognostic and predictive indicators in node-negative breast cancer patients. *J Natl Cancer Inst*. 1993;85:1206-1219.
2. Hayes DF, Bast RC, Desch CE, et al. Tumor marker utility grading system: a framework to evaluate clinical utility of tumor markers. *J Natl Cancer Inst*. 1996;88:1456-1466.
3. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. Reporting recommendations for tumor marker prognostic studies (REMARK). *J Natl Cancer Inst*. 2005;97:1180-1184.
4. Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y. *Design and Analysis of DNA Microarray Investigations*. New York: Springer; 2003.
5. Simon R. A roadmap for developing and validating therapeutically relevant genomic classifiers. *J Clin Oncol.* 2006;23:7332-7341.
6. Pusztai L, Hess KR. Clinical trial design for microarray predictive marker discovery and assessment. *Ann Oncol*. 2004;15:1731-1737.
7. Simon R. Development and validation of therapeutically relevant multi-gene biomarker classifiers. *J Natl Cancer Inst*. 2005;97:866-867.
8. Group BDW. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharm Therap*. 2001;69:89-95.
9. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med*. 1989;8:431-440.
10. Dudoit S, Fridlyand J. Classification in microarray experiments. In: Speed T, ed. *Statistical Analysis of Gene Expression Microarray Data*. New York: Chapman & Hall/CRC; 2003:93-158.
11. 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*. 2003;100:6958-6963.
12. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med*. 2004;351:2817-2826.
13. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the analysis of DNA microarray data: class prediction methods. *J Nat Cancer Inst*. 2003;95:14-18.
14. Simon R. Diagnostic and prognostic prediction using gene expression profiles in high dimensional microarray data. *Br J Cancer*. 2003;89:1599-1604.
15. 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*. 2002;346:1937-1947.
16. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*. 2005;21:3301-3307.
17. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci*. 2002;99:6562-66.
18. Dobbin K, Beer DG, Meyerson M, et al. Inter-laboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. *Clin Cancer Res*. 2005;11:565-572.
19. Hayes DF, Trock B, Harris AL. Assessing the clinical impact of prognostic factors: when is "statistically significant" clinically useful? *Breast Cancer Res Treatment*. 1998;52:305-319.
20. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. *Br J Cancer*. 1994;69:979-985.
21. Bast RC, Ravdin P, Hayes DF, et al. 2000 update of recommendations for the use of tumor markers in breast and colorectal cancer: clinical practice guidelines of the American Society of Clinical Oncology. *J Clin Oncol.* 2001;19:1865-1878.
22. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clin Cancer Res*. 2004;10:6759-6763.
23. Maitournam A, Simon R. On the efficiency of targeted clinical trials. *Stat Med.* 2005;24:329-339.
24. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *J Clin Oncol*. 2005;23:2020-2027.
25. Simon R. When is a genomic classifier ready for prime time? *Nat Clin Pract Oncol*. 2004;1:2-3.
26. Kattan MW. Judging new markers by their ability to improve predictive accuracy. *J Natl Cancer Inst*. 2003;95:634-635.
27. Kattan MW. Evaluating a new marker's predictive contribution. *Clin Cancer Res*. 2004;10:822-824.
28. Dobbin K, Simon R. Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*. 2005;6:27-38.
29. Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. *J Computational Biol*. 2002;9:505-511.
30. Simon R. Targets for treatment success. *Nat Clin Pract Oncol*. 2005;3:1.