At the current time, analysis of circulating tumor cells should not be used routinely until it can be clearly shown that the information it generates cannot be derived from other measures of tumor burden and response. Further, this assay must be shown to influence therapeutic choices that will improve patient outcome.

References


The Importance of Validation in Genomic Studies of Breast Cancer

Jennifer Pittman, PhD

Genomic information is increasingly being used to explore the biological causes and effects of cancer as well as to define clinically relevant risk factors in disease prognosis (see Abstract 1–23).1-5 Microarray technology represents a potentially invaluable tool for the biosciences and for cancer research in particular. Characteristics of breast cancer including molecular subtype and response to treatment can be correlated with changes in the mRNA levels of many genes in diseased cells, and gene array studies are poised to exploit this information to enable both predictive and preventative medicine.6,7 The promise of scientific and medical advancements arising from genomic knowledge is enormous.

Most biological phenomena, however, involve multiple genes that interact in complicated ways with each other and with the environment.8 Resolving this complexity will require nonexperimental as well as experimental methods—an interface of molecular biology, clinical epidemiology, and statistics. In this article, an overview of some of the challenges in creating such an interface will be discussed, with an emphasis on the need for careful study design and properly validated statistical models that can integrate diverse types of data and combine them to produce informed predictions of patient outcomes.

THE IMPORTANCE OF VALIDATION

The growing avalanche of microarrays and other forms of “-omic” data (e.g., proteomic, metabolomic) has driven an explosion of high-throughput, “discovery-based” research during the past decade.5 Among the results have been claims of the discovery of molecular markers successful for diagnosing cancer.
(the original results of which have not been reproduced) and the identification of expression profiles predictive of prognosis in breast cancer (the definition of which may have been compromised by poor study design). The consequence of little or no validation can be chance results and erroneous conclusions, which fuel inflated expectations and then lead to disappointment when results cannot be reproduced.

The sources of uncertainty in genomic studies are numerous, and include experimental design issues, preprocessing (image analysis, probe summaries, normalization, and filtering), data quality (laboratory, platform, and batch effects), and data analysis (methods, integration of diverse data types, and model assessment). Each of these sources of uncertainty in gene array studies is deserving of careful consideration.

Experimental Design
A brief overview of study design issues relating to validation will be given here; additional details can be found in the cited references. The objective of experimental design is to make the analysis of the data and the interpretation of the results as simple and as powerful as possible given the purpose of the experiment and the constraints of the experimental material or data. Given the cost of large-scale gene expression experiments, careful study design is particularly important. Replication of samples (biological, across samples; technical, one sample on several chips) is needed to ensure a valid estimate of error for assessing the significance of the results. Replicates allow summary statistics and averages, which are less variable than their component terms, to be generated, and data from replicates can be analyzed using formal statistical methods. The temptation to overinterpret study results can be offset with careful statements of the characteristics of the study population, sampling design, methods of data analysis, and results.

Data Preprocessing
Important to keep in mind is that the parameter being measured, i.e., the signal intensity of indirectly labeled probes, is many steps removed from the parameter being inferred, i.e., gene expression. Uncontrolled experimental variables are introduced at each step (e.g., labeling reactions, hybridization) that are numerous and poorly understood. As a consequence, a comparison of the same experiment performed twice a few weeks apart reveals wider variation than that seen when a single sample is tested by repeated hybridization, and estimates of “fold change” are biased below the true values because of concentration saturation. Different algorithms have been proposed for doing background correction, normalization, and filtering of expression data, all of which were designed to reduce systematic variation within and across replicate arrays and extract accurate values of expression. However, the effect of these methods on the results of the subsequent analyses is unclear, as few systematic studies of these effects have been done. Further study of and increased familiarity with low-level normalization and analysis methods could allow an accounting of the possibility that information reduction concealed too much about the biological issue of interest when scientific conclusions were drawn.

Not only are different algorithms available for extracting gene expression values, but different technologies are available for measuring gene expression. None of the hybridization-based methods are capable of absolute quantification of mRNA, and none are accepted as a gold standard, even for relative measurements. Various investigators have discovered very low concordance between the results of different technologies, and problems have been reported with oligonucleotide probe performance (personal communication with Wendell Jones, October 2004). Until these issues are resolved, or well understood, various scientific publications have endorsed the use of an independent verification method (northern or western blots, reverse transcriptase–polymerase chain reaction, in situ hybridization) as a follow-up to experimental results.

Model Assessment
Statisticians know that the assessment of the “general” or “real-world” performance of an analytical method, i.e., its ability to predict given independent test data, is extremely important, as it guides not only model choice but gives us a measure of the quality of the chosen model. Training error (the average loss over the training sample) is not a good estimate of test error (the expected prediction error over an independent test sample). The training error decreases with model complexity, eventually dropping to zero, but a model with zero training error is overfit to the training data and typically does not generalize well. This is a reflection of the bias–variance tradeoff: as model complexity increases, the model becomes more adaptive and the bias, or difference between the average of the estimate and the true mean, decreases. At the same time, the variance in the estimate increases, thus increasing estimation error. To assess test error, training data are needed on which to fit the model and validation or test data are needed with which to estimate both predictive error and generalization error.

The true predictive accuracy of a model for a given outcome can be assessed from a cross-validation study in which the analysis is repeatedly performed while holding out a group of samples at each reanalysis and predicting the outcome for the holdout group. Importantly, the entire model building process must form part of each reanalysis; no preselection of predictor variables, or prespecification of aspects of the model, can be made based on an examination of all of the data (see Abstract__). Notably, if the sample size is small, as is true for many expression studies, the cross-validated error estimates will have large variance and cross-validation will fail to estimate the effect on generalization error of complications due to exogenous sources of variation including intrinsic noise (in sub-
jects across studies), intermediate noise (lab procedures), and measurement error (instrumentation, array manufacture, scanning, or in silico processing). For these reasons, an independent validation set for estimating generalization error is strongly preferred.

**Interpretation of Results**

Microarray data analyses are subject to both the "curse of dimensionality," where the number of features is overwhelming, and the "curse of dataset sparsity," where the number of samples is limited. This situation results in a low sample-per-feature ratio and a non-uniqueness of feature sets, complicating the ability to interpret the results. Significant genomic features or patterns can be useful only if they can be put in context and followed up with more detailed studies. For example, in our clinicogenomic study of breast cancer recurrence (see Abstract ), the expertise of geneticists and oncologists was invaluable in interpreting statistical results from a biological perspective, identifying potentially relevant genes and pathways providing clues to the pathophysiology underlying the disease. Although further work is necessary to fully appreciate the significance of their observations, those observations do suggest that characteristics of the tumor that predict lymph node metastasis, and ultimately disease recurrence, relate to the involvement of processes associated with immunologic response to the tumor. Unfortunately, many approaches leave open the question of what significant features or patterns mean from a biological perspective. Extracting knowledge from discovered patterns is a valuable in interpreting statistical results from a biological perspective, identifying potentially relevant genes and pathways providing clues to the pathophysiology underlying the disease. Although further work is necessary to fully appreciate the significance of their observations, those observations do suggest that characteristics of the tumor that predict lymph node metastasis, and ultimately disease recurrence, relate to the involvement of processes associated with immunologic response to the tumor. Unfortunately, many approaches leave open the question of what significant features or patterns mean from a biological perspective. Extracting knowledge from discovered patterns is a serious scientific bottleneck, underscoring the need for effective interdisciplinary communication and collaboration in an integrative genomics context.

**PREDICTION, VALIDATION, AND THE PROMISE OF GENOMICS**

Despite the caveats of uncertainty, the great promise of scientific and medical advancement that genomic data represents is without question. For example, genomic data, particularly gene expression profiles, clearly have the capacity to significantly improve clinical prognosis (see Abstract 1–23). Scientists should embrace discovery-based research and its results, but only if they are rigorously validated. Investigators in some scientific disciplines, such as toxicogenomics, have called for standards in genomic analysis with respect to data generation and exchange, but the rules of evidence to assess validity about diagnosis and prognosis remain underdeveloped and not routinely applied.

Any method of modeling and analysis must show reproducible results before being implemented as novel prognostic or preventative tools in clinical practice. Predictive analysis and proper study validation are necessary to properly assess model performance in terms of generalization error and areas of misspecification, as well as local lack of fit and overfit. Appropriate experimental design, execution, and proper analytic and statistical methods are necessary and powerful tools in making the promise of genomics a reality.

**References**


