Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting

Alain Dupuy, Richard M. Simon

Background Both the validity and the reproducibility of microarray-based clinical research have been challenged. There is a need for critical review of the statistical analysis and reporting in published microarray studies that focus on cancer-related clinical outcomes.

Methods Studies published through 2004 in which microarray-based gene expression profiles were analyzed for their relation to a clinical cancer outcome were identified through a Medline search followed by hand screening of abstracts and full text articles. Studies that were eligible for our analysis addressed one or more outcomes that were either an event occurring during follow-up, such as death or relapse, or a therapeutic response. We recorded descriptive characteristics for all the selected studies. A critical review of outcome-related statistical analyses was undertaken for the articles published in 2004.

Results Ninety studies were identified, and their descriptive characteristics are presented. Sixty-eight (76%) were published in journals of impact factor greater than 6. A detailed account of the 42 studies (47%) published in 2004 is reported. Twenty-one (50%) of them contained at least one of the following three basic flaws: 1) in outcome-related gene finding, an unstated, unclear, or inadequate control for multiple testing; 2) in class discovery, a spurious claim of correlation between clusters and clinical outcome, made after clustering samples using a selection of outcome-related differentially expressed genes; or 3) in supervised prediction, a biased estimation of the prediction accuracy through an incorrect cross-validation procedure.

Conclusions The most common and serious mistakes and misunderstandings recorded in published studies are described and illustrated. Based on this analysis, a proposal of guidelines for statistical analysis and reporting for clinical microarray studies, presented as a checklist of "Do's and Don'ts," is provided.

J Natl Cancer Inst 2007;99:147–57

DNA microarray technology has found many applications in biomedical research. In oncology, it is being used to better understand the biological mechanisms underlying oncogenesis, to discover new targets and new drugs, and to develop classifiers (predictors of good outcome versus poor outcome) for tailoring individualized treatments (1–4). Microarray-based clinical research is a recent and active area, with an exponentially growing number of publications. Both the reproducibility and validity of findings have been challenged, however (5,6). In our experience, microarray-based clinical investigations have generated both unrealistic hype and excessive skepticism. We reviewed published microarray studies in which gene expression data are analyzed for relationships with cancer outcomes, and we propose guidelines for statistical analysis and reporting, based on the most common and serious problems identified.

Methods

Studies were retrieved from a search of the Medline bibliographic database on the Pubmed Web site of the National Library of Medicine, followed by hand screening of abstracts and articles. The detailed process of selection is presented in Supplementary Note 1 (available online). The inclusion criteria were as follows: the work was an original clinical study on human cancer patients, published in English before December 31, 2004; it analyzed gene expression data of more than 1000 spots; and it presented statistical analyses relating the gene expression profiling to a clinical outcome. Two types of outcome were considered: 1) A relapse or death occurring during the course of the disease. 2) A therapeutic response.

Affiliations of authors: Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, Bethesda, MD (AD, RMS); Université Paris VII Denis Diderot, Paris, France (AD); Assistance Publique-Hôpitaux de Paris, Service de Dermatologie, Hôpital Saint-Louis, Paris, France (AD).

Correspondence to: Richard M. Simon, DSc, National Cancer Institute, 9000 Rockville Pike, MSC 7434, Bethesda, MD 20892 (e-mail: rsimon@nih.gov).

See "Notes" following "References."

DOI: 10.1093/jnci/djk018

CONTEXT AND CAVEATS

Prior knowledge

The use of microarray technology has generated great excitement for its potential to identify biomarkers for cancer outcomes, but the reproducibility and validity of findings based on microarray data have come under widespread challenge.

Study design

This is a systematic review of microarray studies in which gene expression data were analyzed for relationships with cancer outcomes.

Contribution

Common methodologic errors committed in statistical analysis of the relationship of gene expression data to cancer outcomes were identified and explained. A set of useable guidelines for statistical analysis and reporting of clinical microarray studies were created for the cancer research community.

Implications

The new guidelines could serve as an accessible and common basis for discussion among all cancer researchers involved in microarray investigations.

Limitations

Technical procedures for generating reproducible gene expression data are not addressed here.

Exclusion criteria were as follows: 1) the study focused the outcome-related analysis on one or a few individual genes rather than on a gene expression signature and 2) the study on therapeutic response dealt exclusively with before–after comparisons of gene expression profiles.

The bibliographic selection process yielded 90 papers. Descriptive characteristics of these papers were recorded: the journal, with its 2004 impact factor; the year of publication; the type of cancer studied; the number of patients with outcome information; the type of clinical outcome considered; and the type of analysis (Table 1; more details are provided in Supplementary Table 1, available online).

The statistical analysis was examined in detail for the 42 articles that were published in 2004. In each article, the outcome-related analyses that agreed with the above criteria were identified. For instance, a search for differentially expressed genes between patients who died within 5 years and patients who survived was selected for examination, whereas the same type of analysis, in the same paper, comparing two histologic subtypes was not selected because the classes were not defined by the clinical outcome. Some articles presented several outcome-related analyses, and thus, the flaws identified did not necessarily invalidate all the results presented in the paper. We did not rate the overall quality of the statistical analysis; we abstracted details about the analysis and then tabulated the problems identified. Outcome-related analyses were classified into the three broad categories of statistical analysis of microarray data: finding genes correlated with outcome, class discovery, and supervised prediction.

Outcome-related gene finding generally involved statistical methods to identify genes that were differentially expressed according to two categories of outcome (e.g., responders and nonresponders to a particular treatment). Finding genes whose expression differ between two classes has often been called "class comparison." Here we use the term "outcome-related gene finding" to emphasize that only outcome-related analyses were addressed and because, in some cases, finding genes was accomplished by correlating expression level to survival or disease-free survival directly, rather than dichotomizing the outcome into discrete classes. Outcome-related gene finding is usually done to obtain clues about the biological mechanisms that might be related to prognosis or response to therapy. It does involve making an inference about each gene whose expression is measured on the array, however, and therefore, we examined the methods used by the authors to control the number of positive claims that a gene was outcome related.

Class discovery generally uses cluster analysis methods for grouping specimens that have similar gene expression profiles.

Table 1. Descriptive characteristics of the 90 reviewed studies

Study characteristic	No. of studies, n (%)
Type of cancer studied	
Hematologic malignancies	24 (27)
Lung and pleura	12 (13)
Breast	12 (13)
Hepatodigestive system	9 (10)
Genitourinary*	8 (9)
Genital (female)	6 (7)
Head and neck	5 (6)
Brain	4 (4)
Melanoma	2 (2)
Other	8 (9)
No. of patients with outcome information	
<15	11 (12)
15–24	26 (29)
25–49	22 (24)
50–99	26 (29)
≥100	5 (6)
Type of clinical outcome addressed†	
Follow-up data	69 (77)
Type of event	
Death	34 (38)
Relapse	25 (27)
Both	10 (11)
Response to treatment	26 (29)
Type of treatment	
Chemotherapy	15 (17)
Radiotherapy	5 (6)
Biological therapies	6 (7)
Type of analysis	
Outcome-related gene finding	48 (53)
Class discovery	60 (67)
Supervised prediction	57 (63)
Outcome-related gene finding only	5 (6)
Class discovery or supervised prediction	85 (94)
Journal impact factor (2004)‡	
<3	7 (8)
3–<6	15 (16)
6–<10	35 (39)
≥10	33 (37)

* Encompasses cancers of the urinary system (both sexes) and male genital cancers.

† Some studies had both types of outcomes.

‡ Journal impact factor from Journal Citation Reports, Thomson Scientific, Philadelphia, PA.

The specimens are grouped based on the pairwise similarities or dissimilarities (i.e., distances) of their corresponding expression profiles. Clustering of specimens based on similarities of expression profiles is not in itself an outcome-related analysis. Some of the papers we examined attempted to establish clinical relevance of their clusters by comparing outcomes for groups of patients whose specimens were in different clusters. We considered such analyses to be outcome-related analyses, and, in these cases, we recorded the clustering method used. The most popular method is hierarchical clustering (7), which produces a tree representation of nested clusters. This representation, called a dendrogram, shows the individual specimens at the bottom level and the single supercluster containing all specimens at the top. The dendrogram itself does not define a specific set of disjoint clusters that can be correlated with outcome, and therefore, we recorded how the authors "cut" the dendrogram to obtain distinct clusters. In most cases, the cutting of the dendrogram was based on subjective visual analysis. We also recorded whether the authors used any of the available statistical methods for evaluating the robustness of the clusters to variation in expression values (8,9). Cluster analysis is generally considered to be "unsupervised" in the sense that clusters or dendrograms are based only on the similarities among the expression profiles and not on any external class variable. The quantitative similarity or dissimilarity of two expression profiles depends, however, on the genes included in the calculation of the similarity function. Consequently, we focused on whether clinical outcome information had been used in earlier steps for limiting the expression profiles used for clustering so that they contained only outcome-related expressed genes.

The third type of analysis we addressed is supervised prediction. This generally involves building and validating a classifier that can be used in the future to accurately predict the outcome of similar patients based on their expression profiles. For some of the papers, the outcome classes predicted were response or no-response to a defined treatment. Other studies measured survival, disease-free survival, or time to disease progression and categorized patients into those with survival times beyond a landmark (e.g., 5 years) versus those who had died with shorter survival times (excluding those still alive with shorter survival times). Some papers defined the classes as "alive" or "dead" without consideration of the survival times. We recorded key features of both classifier building and classifier validation steps. Many different types of classifiers are possible—linear discriminants, decision trees, nearest neighbor classifiers, and neural network classifiers, among others. We recorded the type of classifier used. Many approaches to supervised classification involve using only genes that are correlated with outcome. We also recorded the method used by the authors for gene selection in building the classifier. Because there is no consensus in the statistical and machine learning communities about what types of classifiers or variable selection methods are best, we did not critique the papers on these aspects. We did, however, focus on the methods used for validating the classifier. The fundamental rule here is that the samples used for the validation step must not have been used for building the classifier. Two types of methods can be used to ensure that this principle is not violated: the cross-validation procedure and the split-sample procedure. For supervised prediction, we recorded details about the type of validation procedure and how the classifier performance was assessed, tested, and presented. We focused on whether the principle of separating the classifier development and its validation was followed.

These three types of statistical analysis, together with the main pitfalls we found, are further described in the "Results" and "Discussion" sections of this paper.

Results

We reviewed 90 microarray studies for cancer outcome. They were published between 2000 and 2004. The general characteristics of the studies are presented in Table 1. Death or relapse was an outcome in 64 studies (71%) and a therapeutic response in 21 studies (23%); both death or relapse and therapeutic response were the outcomes in five studies (6%).

Statistical analyses fell into three distinct types: outcome-related gene finding, class discovery, and supervised prediction. Outcome-related gene finding was used to identify genes correlated with the outcome. It was used in 48 (53%) of the studies. Class discovery was mainly based on various forms of cluster analysis. Here the aim was to identify groups of patients with similar expression profiles and then attempt to demonstrate that the resulting clusters were related to outcome. Class discovery was used in 60 (67%) of the studies. Finally, supervised prediction used both the outcome information and the expression data to develop and evaluate a classifier that could be used in the future to predict outcome in new patients, based only on their expression profiles. Supervised prediction was used in 57 (63%) of the studies. Overall, 85 studies (94%) used either class discovery or supervised prediction or both.

We recorded detailed information on these outcome-related analyses in the 42 (47%) studies published in 2004. Table 2 presents, for each type of outcome-related analyses in those studies, the methods used and the way the findings were reported.

For outcome-related gene finding, the most common and serious flaw was an inadequate, unclear, or unstated method for controlling the number of false-positive differentially expressed genes. This flaw was present in 9 of the 23 studies published in 2004 that reported results of outcome-related gene-finding analyses. Most of the studies in the group of 23 identified genes whose expression was correlated with outcome by performing statistical significance tests comparing, for each gene represented on the array, the gene's average expression in the poor outcome group to its average expression in the good outcome group. Some studies fitted a proportional hazards survival model to expression levels for each gene, one gene at a time, to obtain a P value for the correlation of the expression of that gene to outcome. In either case, if the threshold used for claiming statistical significance is the traditional P less than .05 level, then one expects an average of 500 false-positive genes to be claimed as correlated with outcome for every 10 000 analyzed on the array. We considered use of this threshold to be inadequate control of the number of false positives. An approach to controlling false positives that we did consider adequate is the use of a reduced statistical significance threshold (P<.001) for analysis of individual genes. This threshold limits the number of false positives to 10 per 10 000 genes analyzed on the array.

Some authors used the Benjamini–Hochberg method (10) or similar methods to control the false discovery rate (i.e., the

Table 2. Statistical analysis and reporting in microarray studies for cancer outcome published in 2004*

Type of analysis/reporting	No. of studies
Outcome-related gene finding	23
Statistics used for comparison	
Not mentioned	2
SAM *t* statistic†	6
Golub's discrimination score‡	4
t statistic	3
Wilcoxon–Mann–Whitney statistic	3
Hazard ratio coefficient	2
Combination or other	3
Assessment of statistical significance	
Not mentioned	1
Parametric	5
SAM†	6
Wilcoxon test	3
Permutation test	5
Combination	3
Method for controlling the number of false positives	
None mentioned	1
Inadequate§	3
Uncertain adequacy§	5
Lowering *P* value threshold (<.05)	6
SAM†	6
Other	2
Class discovery‖	28
Type of analysis‖	
Hierarchical clustering	28
K-means clustering	2
Nonclustering methods	6
Dataset clustered	
Whole dataset	18
Selection of outcome-related differentially expressed genes	13
Test for validation of cluster–outcome correlation¶	
None	12
Log-rank test or hazard ratio significance	12
Chi-square test or Fisher's exact test	4
Test not specified	2
Supervised prediction	28
Main model used for classification	
Weighted voting	7
Nearest centroid	4
Nearest shrunken centroid	4
Proportional hazards	3
K-nearest neighbor	2
Artificial neural networks	2
Other	6
Statistics used for feature selection	
Golub's discrimination score	9
t test	6
PAM#	4
Hazard ratio	3
Other or multiple	6
Validation procedure	
Use of a separate test set	15
Origin of test set	
Random split on initial dataset	10
External dataset or new series of samples	5
Size of test set, median (range)	25 (4–96)
Ratio test/training set size, median (range)	0.9 (0.5–3.8)
Preliminary use of outcome information from test set samples	
Yes	0
No	15
Use of expression data from test set samples for class definition	
Yes	2
No	13
Cross-validation procedure	13
Preliminary use of outcome information from test samples	
Yes	12
No	1
Use of expression data from test samples for class definition	
Yes	2
No	11
Presentation of classifier performance	
Prediction of a risk group (four studies)	
Survival curves for predicted groups	4
Prediction of a binary outcome (death, recurrence, response) (24 studies)	
None	1
Prediction accuracy or misclassification rate	23
Sensitivity and specificity, or equivalent**	21
Odds ratio	2
Survival curves for predicted groups	9

* A more detailed table is available as Supplementary Table 2 (available online). SAM = significance analysis of microarrays; PAM = prediction analysis of microarrays.

† (22).

‡ See reference (23).

§ Control for falsely differentially expressed genes was considered inadequate if genes were selected at a .05 *P* value threshold. Adequacy was considered unclear if genes were selected at a .05 *P* value threshold, with additional selection based on fold change.

‖ The descriptive characteristics included in the table for class discovery refer exclusively to hierarchical clustering.

¶ Several methods could be used in the same article.

(24).

** Any presentation (graphical display, positive and negative apparent predictive values, contingency table) allowing calculation of sensitivity and specificity.

expected proportion of false positives among the genes claimed to be correlated with outcome). We considered use of such methods to be adequate approaches to controlling for multiple testing. We did not impose a judgment of what level the false discovery rate should be limited to, but levels less than 10%–20% are desirable. A 10% false discovery rate means that for every 10 findings that a gene is correlated with outcome, one is expected to be false. We also considered the approach adequate if the authors used the more powerful multivariable methods such as significance analysis of microarrays or the multivariable permutation test for identifying genes correlated with outcome while controlling the false discovery rate (11,12). The false discovery rate has achieved broad acceptance as the appropriate criterion for controlling the multiple testing problem in microarray investigations.

For class discovery, the most common and serious flaw was a spurious claim that the expression clusters were meaningful for distinguishing different outcomes, when the clustering itself was

based on genes selected for their correlation with outcome. This flaw was present in 13 of the 28 studies published in 2004 reporting class discovery analyses. If one uses a null dataset in which the expression of all genes have the same distribution in the two outcome classes, for example, dead and alive, clustering the data with regard to all the genes will not identify a cluster grouping the dead and another grouping the alive patients. If there are 10 000 genes on the array, however, then we can expect to find about 500 genes for which the expression levels of the alive group is significantly different than those of the dead group at the *P* less than .05 level. If we cluster the samples using expression levels only for those 500 or so genes, the procedure will result in a cluster grouping a majority of dead and another grouping a majority of alive (Fig. 1). In this case, the correlation between clusters and outcome is a consequence of the selection of spuriously outcome-related differentially expressed genes. It is not independent evidence that outcome can be predicted based on expression levels.

For supervised prediction, many different classification algorithms were used. Although previous studies have indicated that simpler classifiers such as diagonal linear discriminant analysis and nearest neighbor methods perform as well or better as more complex algorithms (13), we did not consider selection of an inappropriate classification algorithm to be a flaw in any study. For supervised prediction, the most common and serious flaw was a biased estimation of the prediction accuracy for binary outcomes. This flaw was present in 12 of the 28 studies published in 2004 reporting supervised prediction analyses.

The most straightforward approach for properly evaluating a classifier is to base the evaluation on a separate test set of cases. In a split-sample procedure, the initial dataset is randomly split into two subsets. One part, the training set, is used for developing the classifier. The other part, the test set, is used to evaluate the fully specified classifier developed from the training set. The split-sample validation procedure is illustrated in Fig. 2. The fundamental principle of classifier validation, whatever the classifier type, is that the samples used for validation must not have been used in any way before being tested. Most importantly, the outcome information of the tested samples must not have been used for developing the classifier or in steps before classifier development.

In a procedure using separate training and test sets, classifier development and evaluation of the prediction are distinguished in a physical way. Another procedure for developing and evaluating a classifier is cross-validation. In a cross-validation procedure, the two conceptually distinct steps of classifier development, on the one hand, and validation of the prediction, on the other, may seem intertwined. However, the fundamental principle of not using a sample before testing it still holds. Cross-validation is an iterative process. In each iteration, part of the initial dataset is left apart to be tested. The other part is used as a temporary training set. A detailed description of a correct cross-validation procedure is presented in Fig. 2. Cross-validation methods were widely used in the reviewed studies, mainly in the "leave-one-out" form, in which one sample is left out for testing at each iteration. Unfortunately, cross-validation was sometimes used improperly, resulting in a biased estimation of prediction accuracy. The most common forms of misuse involved using the outcome data to select genes using the full dataset, rather than performing gene selection from scratch within each loop of the cross-validation. This problem can also exist with other validation methods that have been proposed such as bootstrap resampling (5) or multiple training test partitions (5,14).

At least one of the three major flaws described above was present in 21 (50%) of the 2004 publications. Flaws in class discovery analysis or supervised prediction were present in 19 of them. The presence of at least one of these three flaws was inversely correlated with the journal impact factor ($P = .005$, Wilcoxon two-sample test). Articles presenting these types of flawed analyses were nevertheless highly prevalent in high–impact factor journals. For example, in journals of impact factor between 6 and 10, they were present in 10 out of 20 articles (50%). Selected commented examples of these serious flaws extracted from the studies are provided in Supplementary Note 2 (available online). Other inappropriate or incorrect analyses present in the studies are mentioned in Supplementary Table 2 (available online). Details of the distribution of major flaws according to journal impact factor are given in Supplementary Table 3 (available online).

Discussion

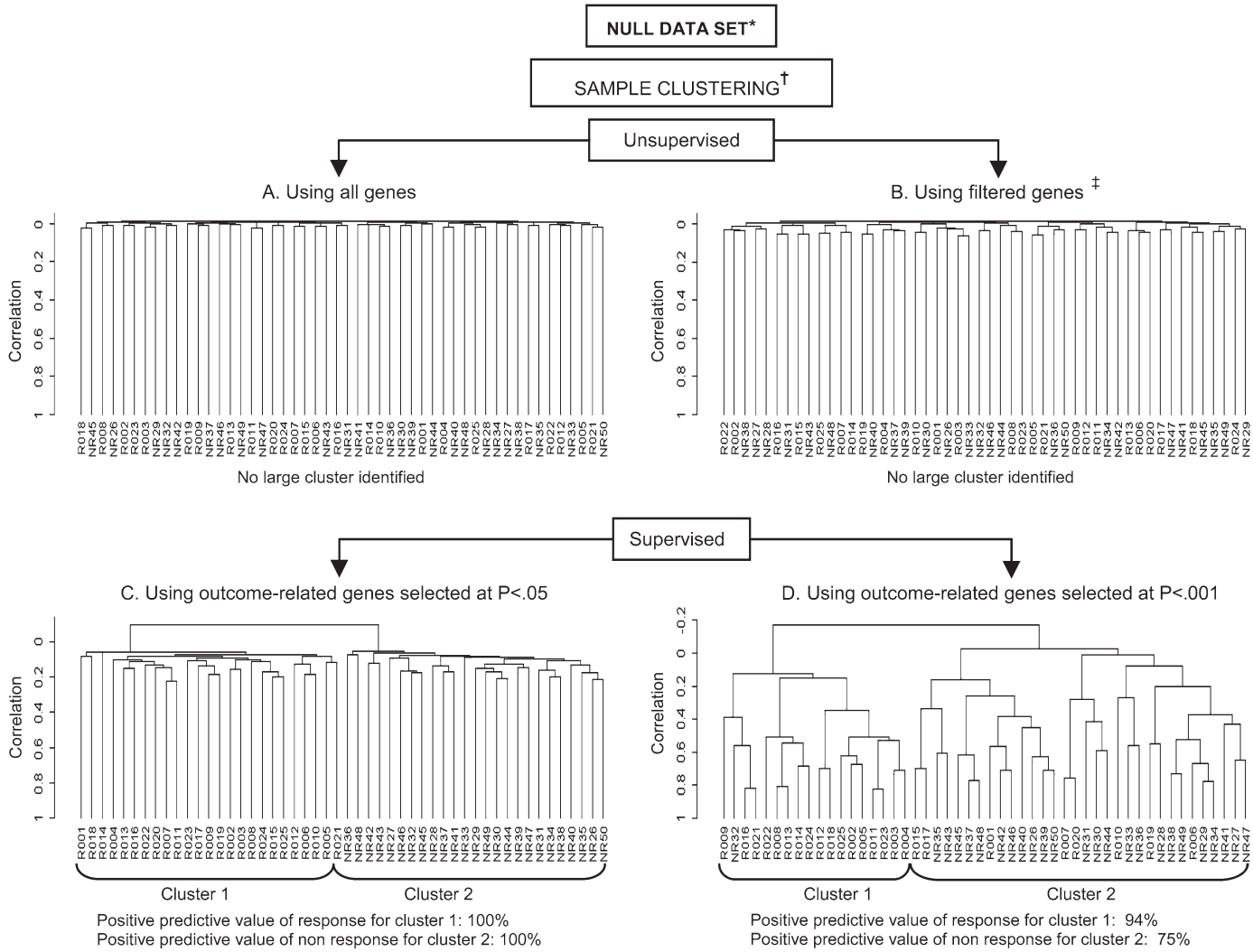
Our review of microarray studies for cancer outcome published in 2004 showed that half of them presented basic flaws in statistical analysis. Although our study selection may not have been exhaustive, a broad range of journals and cancer types are represented, and therefore, it is unlikely that our selection process was biased toward statistical analyses of poorer quality.

Need for Clear Objectives

In our review, we assessed the studies only for their outcome-related analyses. We did not present the type of objective in our tabulations because many studies had ill-defined objectives. In contrast to hypothesis-driven research, microarray investigation has been defined as discovery-based research (8). However, even for discovery-based research, clear objectives are needed for determining an effective study design and for selecting an appropriate analysis strategy.

Study objective should influence patient selection. In cancer studies, selecting a heterogeneous group of patients presenting with different stages of disease and receiving a variety of treatments usually leads to substantial difficulties in interpreting the results of outcome-related analyses. The main problem lies in the possibility of confounding patient outcome by stage and treatment. For example, cluster analyses of very heterogeneous sets of patients frequently demonstrate that advanced cancers have similar expression profiles that differ from early cancers, a result of little relevance regarding outcome analysis. For supervised prediction, the development of a classifier should be guided by the specific therapeutic decision context. There is a vast literature of unused "prognostic factors" that have no therapeutic relevance. The therapeutic context should be reflected in the patients included in the classifier development study.

The choice of analysis methods should be made according to the objective of the study. Microarray study objectives are often categorized as class comparison (or gene finding), class prediction



E. Simulation study of clustering null data sets* (sample clustering using partition around medoids with 2 clusters)

Simulation	Criteria for selecting genes used for sample clustering	Number of simulated data sets	Number of genes Mean (SD)	Positive predictive value of response for cluster 1 Mean (SD)	Positive predictive value of nonresponse for cluster 2 Mean (SD)
1	None	500	10000 (0)	55.7% (5.1)	55.3% (4.5)
2	Filtering	500	6785 (47)	55.2% (4.1)	55.4% (4.5)
3	DEG 0.05§	500	498 (21)	99.7% (1.3)	99.8% (1.3)
4	DEG 0.001§	500	9.6 (3.0)	86.6% (1.1)	86.9% (1.0)

Fig. 1. A–D) Dendrograms of hierarchical sample clustering of a null dataset* in four situations according to the type of gene selection (mentioned above each dendrogram) performed before clustering. In cases shown in **A** and **B**, unsupervised clustering failed to identify clusters correlated with the clinical outcome. This is the expected result for clustering a null dataset with a randomly allocated outcome. In cases shown in **C** and **D**, clustering was priorily supervised by using the response information to select the genes. The two identified clusters correlate convincingly with the clinical outcome, demonstrating that clustering was actually outcome driven through the prior selection of outcome-related differentially expressed genes. A claim for having discovered clinically meaningful clusters by correlating cluster and clinical outcome would therefore be spurious. **E)** Results of a simulation study using 500 different null datasets. Simulations 1–4 reproduced the type of gene selection used in **A–D**, respectively. In simulations 1 and 2, when clustering is unsupervised, no correlation between clusters and outcome categories is evidenced. In simulations 3 and 4, when clustering has been supervised by selecting outcome-related differentially expressed genes, a spurious correlation between cluster and outcome is evidenced. *The null dataset incorporates 10 000 genes and 50 samples. Gene expression data values originate from a normal distribution (mean = 0; standard deviation = .05). The outcome is binary: response or nonresponse to a treatment. Half of the samples were randomly allocated to being from a responder, the other half to being from a nonresponder. †Hierarchical clustering using centered Pearson correlation metric and average linkage. ‡A gene was filtered out if less than 20% of its expression data values had at least 1.5-fold change in either direction from the gene's median value; § = Differentially expressed genes (DEG) using a .05 or .001 *P* value threshold for *t* test between outcome-defined classes.

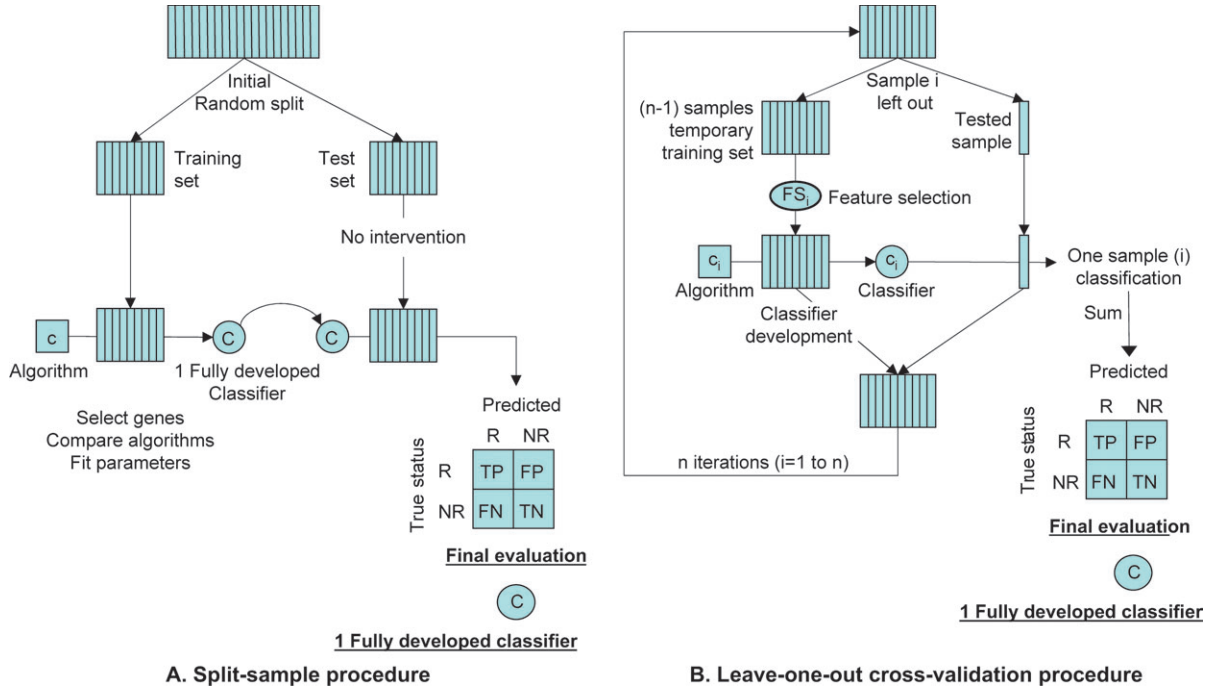


Fig. 2. Developing and validating a classifier. The classifier is for a binary outcome such as response (R) or nonresponse (NR) to a treatment. The final result is composed of two parts: a contingency table presenting the results of the validation (TP = true positive; TN = true negative; FP = false positive; FN = false negative) and the fully parametrized classifier that can be used to predict the outcome in new samples with unknown status. **A**) Split-sample procedure. A random split divides the initial dataset into a training set and a test set. The classifier is developed on the training set. Once the classifier has been fully specified, the test set is accessed once and only for estimating the prediction accuracy of the classifier. The two steps of developing (on the training set) and evaluating (on the test set) are physically and temporally distinct. The information from the test set samples has never been used in any prior step of data handling. **B**) Leave-one-out cross-validation procedure. From the initial n-sample dataset, one sample is withdrawn, leaving a temporary (n − 1)-sample training set and one left-out test sample. On the training set, a group of outcome-related genes is selected (FS = feature selection). The expression data values from these selected genes are used to parametrize the classification algorithm (c in square = classification algorithm before parametrization; C in circle = parametrized classifier). The parametrized classifier is then used to classify the previously left-out test sample as "responder" or "nonresponder." These steps are reiterated n times, until each sample has in turn been left out once for testing. Subscripts i denote that the genes selected (feature selection, FSi) or the parameters of the classifier (Ci) are different at each iteration of the cross-validation.

(prediction of clinical outcome), or class discovery (grouping samples or genes with similar expression profiles).

Class Discovery

The place of class discovery in outcome-related analyses is limited. Regarding the correlation of gene expression and clinical outcome, there are two basic questions: which genes have expression levels correlated with outcome and whether an expression profile can predict the clinical outcome. The former leads to gene-finding methods and the latter to supervised prediction methods, which use the clinical outcome information to optimize the predictive accuracy. Class discovery methods per se are best suited for grouping genes into subsets with similar expression patterns over the samples to elucidate pathways.

Finding Outcome-Related Genes

If the goal is to identify which genes have expression levels that are correlated with outcome, the methods of class comparison are appropriate if the outcomes are grouped into discrete classes. Although many methods of class comparison are available, it is very important that the method being used control the number of false positives because usually thousands or tens of thousands of genes are being evaluated. If the outcome is survival, disease-free survival, or progression-free survival, it is best not to group the cases into discrete outcome classes as this reduces the information available and may invite improper handling of censored values.

Supervised Prediction

If the goal is to predict patient outcome, supervised prediction methods should be used. Supervised methods utilize the outcome data in developing a classifier. In many studies, most genes are not correlated with outcome. Consequently, the expression profiles as a whole are not effective in predicting outcome because the information in the informative genes is swamped by the number of noninformative genes. Classifiers based on combining information from the informative genes that are correlated with outcome give more accurate predictions, and supervised methods can identify those genes (15).

In using supervised methods, however, it is essential to strictly observe the principle that the data used for evaluating the predictive accuracy of the classifier must be distinct from the data used for selecting the genes and building the supervised classifier. Preliminary use of outcome information from tested samples was common in the reviewed studies using a cross-validation procedure. Two distinct practical ways of violating the principle of strict separation between classifier development and evaluation of the prediction are illustrated in Supplementary Fig. 1, A and B (available online). The most straightforward way of violating this

Table 3. Guidelines for statistical analysis and reporting of microarray studies for clinical outcomes*

Checklist			Comment
Objective			
1	Do	Define and state the objective of the study.	
2	Do	State the criteria for selecting the patients.	An excessively heterogeneous set of patients is often used.
Acquisition of data			
3	Do	Describe the biotechnical characteristics of the array experiment.	See Minimum Information About Microarray Experiment checklist (25).
4	Do	Make the raw dataset publicly available.	Allows reproducibility of analysis to be verified. Allows other investigators to use it as an external set for validation.
Statistical analysis: general options			
5	Do	Be aware that many aspects of statistical analysis and reporting of microarray studies are not covered in this checklist.	
6	Don't	Consider that all the items included in these guidelines are commandments.	Most are, however. Violations should be justified.
7	Do	Describe in sufficient detail all the statistical methods used.	
8	Do	Provide detailed information about the experimental design and criteria used for selection of cases.	Randomized clinical trials are preferable.
9	Don't	Transform time-to-outcome data into a binary outcome variable if the goal is to predict groups with different survival probabilities.	Use statistical methods suited for time-to-event data, unless you can ensure the absence of bias due to transformation. See text and Supplementary Fig. 2 (available online).
Outcome-related gene finding†			
10	Don't	Use only fold changes between groups to select the differentially expressed genes.	This does not take into account the variance of the genes' data values.
11	Don't	Use a .05 *P* value threshold to select the differentially expressed genes.	A set of 10000 genes will yield on average 500 false-positive genes if this threshold is used.
12	Do	Use a method for controlling the number of falsely differentially expressed genes.	Lowering the *P* value threshold for selection (e.g., to .001) is the simplest method. Others are available.
13	Do	Use a permutation test to assess the probability of finding the same number of differentially expressed genes as the one you found from your dataset.	The result should be significant at .05 *P* value level.
Class discovery			
14	Don't	Use class discovery methods if you are interested in classifying new samples in the future.	Supervised prediction should be used for this purpose. It utilizes the outcome information to optimize predictive accuracy. See text.
15	Don't	Use a selection of outcome-related differentially expressed genes if you intend to correlate cluster-defined classes with the outcome.	Supervised clustering leads to a spurious correlation between cluster and outcome. See text and Fig. 1.
16	Don't	Select the clustering method that gives the best result.	Class discovery should not be result driven.
17	Do	Use methods for testing the reproducibility of cluster finding.	Assessing the reproducibility of cluster finding without using external information makes class discovery more convincing. See text.
18	Don't	Use conventional statistical tests for computing the statistical significance of genes that are differentially expressed between two clusters.	These tests assume independence between class definition and expression profile data, which is not the case for cluster-defined classes.
Supervised prediction			
19	Do	Frame a therapeutically relevant question and select a homogeneous set of patients accordingly.	Classifiers developed outside a specific therapeutically relevant context are unlikely to be useful and utilized. See text.
20	Don't	Violate the fundamental principle of classifier validation, i.e., no preliminary use of the tested samples.	Most of the "Don't" items on validation procedures are illustrations of how this principle can be violated. See text and Fig. 2 and Supplementary Fig.1 (available online).
21	Don't	Attempt to predict cluster-defined classes.	Classes should be defined independently from the expression profile data.
Evaluating the prediction on a separate test set			
22	Don't	Use any information from the test set for developing the classifier.	The test set is to be used exclusively for evaluating the classifier performance. See text and Fig. 2.
23	Do	Access the test set only once and only for testing the samples with the fully specified classifier developed from the training set.	The test set must not be used to choose the best classifier. See text and Fig. 2.
24	Do	Use the same outcome definition as the one used in the training set.	

(Table continues)

Table 3 (continued).

Checklist			Comment
Evaluating the prediction with a cross-validation procedure			
25	Don't	Use all the samples from the dataset to develop the classifier and test them.	The resubstitution estimate is not a cross-validation procedure. See text and Fig. 2.
26	Don't	Use the same feature selection for all iterations.	This inflates the estimate of the prediction accuracy. See text and Fig. 2.
27	Don't	Perform a cross-validation procedure on a selection of outcome-related differentially expressed genes.	Idem. Invalid although commonly done.
28	Do	Report the estimates for all the classification algorithms if several have been tested, not just the most accurate.	
29	Don't	Consider that testing a few additional independent samples adds value to a correctly cross-validated estimate of the classifier prediction accuracy.	However, this may be valuable if the additional samples are in sufficient number and are representative of the samples in which the classifier might be used in the future. See text.
30	Do	Report the fully specified classifier with its parameters.	So it can be used by others. Parameters are obtained from the whole training set in a separate test set procedure and from the whole dataset in a cross-validation procedure.
31	Do	Report the correctly validated sensitivity and specificity or positive and negative apparent predictive values (for a binary outcome).	Receiver-operating characteristic curves may also be used. See text.
32	Don't	Use an odds ratio to assess the performance of the prediction (for a binary outcome).	The odds ratio is a measure of association, not of prediction accuracy. See text and Supplementary Fig. 3 (available online).
33	Do	Report the statistical significance of the prediction accuracy and, even better, of the sensitivity and specificity (for a binary outcome).	It states the probability of obtaining a prediction accuracy as high as actually observed if there was no relationship between the expression data and the outcome. See text.
34	Don't	Use a Fisher's exact test or chi-square test to assess the statistical significance of the prediction accuracy for a binary outcome.	They do not test the statistical significance of the prediction. See text and Supplementary Fig. 3 (available online).
35	Do	Pay attention to the imbalance between outcome categories when interpreting the prediction accuracy of a binary outcome.	90% prediction accuracy may be inadequate if outcome categories are highly imbalanced. See text and Supplementary Fig. 3 (available online).
36	Don't	Use the log-rank test for testing the difference in survival between cross-validated groups.	The test is invalid because of a dependency among cases after cross-validation.
37	Don't	Use standard regression models, e.g., logistic regression or proportional hazards model, with cross-validated predicted groups.	Idem.
38	Don't	Assess the utility of the prediction based on the value of the regression coefficient or on its *P* value from multivariable regression models.	Regression coefficients are poor measures of prediction accuracy, and the test of statistical significance simply assesses if the coefficient is different from 0. See text.
39	Do	Assess the added value of the classifier by examining its performance within the levels of the standard prognostic factors.	Other approaches can be used. See text.
40	Do	Assess the utility of the classifier in a clinical context, for the therapeutically relevant question, and plan, if appropriate, further studies for external validation.	

* The items are described in a context of a microarray investigation for a clinical outcome in cancer patients but may apply to other situations.

† Described for a binary outcome (class comparison).

principle is to use a resubstitution estimate, with no cross-validation attempt, as illustrated in Supplementary Fig. 1, C (available online). The consequences of using information from tested samples before they are actually tested should not be underestimated—the practice leads to an overly optimistic and markedly inflated prediction accuracy (9).

Cross-validation provides an estimate of the prediction error to be expected for the classifier developed using the full dataset. Some authors have criticized microarray classifiers because different studies analyzing the same outcome report different genes used in the classifiers. The true test of a classifier, however, is whether it predicts accurately for independent data, not whether a repeat of the development process on independent data results in a similar gene set. Because the expression levels of different genes are correlated and because statistical power to select individual genes is limited, the set of genes selected may vary substantially among studies and even among different iterations of the cross-validation process in a single study. In some cases, it may be useful to summarize the stability found, but using the stability results to refine the genes used in the classifier amounts to defining a different classification algorithm.

Some studies presented a dual-validation procedure. Validation of the classifier was achieved both with a cross-validation procedure and by using "additional independent samples." This practice almost invariably brought more confusion than clarity. The so-called test set was generally inadequate because it was based on too few samples, or too few patients, in one of the outcome categories.

The important rule is that the test set should be large enough and composed of patients representative of the set of patients for which the classifier might be used in the future.

Although the use of a separate test set may appear to be a gold standard to some, the gold standard should rather be to properly validate the classifier performance, and this can be achieved through cross-validation as well. Using a separate test set is most useful when one does not have an a priori, well-defined algorithm for developing the classifier. All the comparisons between different algorithms can be made on the training set, provided that only one fully specified classifier is finally chosen for evaluation on the test set. By contrast, comparing different algorithms in a cross-validation procedure could be misleading if only the best results are reported. More complex procedures, such as embedding cross-validation for model selection in each iteration of the cross-validation (for estimating the prediction accuracy), can be used for choosing the best-performing model in the absence of a separate test set.

Any presentation of a correctly validated prediction accuracy should be accompanied by an assessment of its statistical significance to determine the probability of obtaining a prediction accuracy as high as actually observed if there were no relationship between the expression data and the outcome. For a binary outcome, the *P* values of a chi-square test, a Fisher's exact test, or an odds ratio are not suited for this purpose because they all test association, not prediction accuracy (13). This is illustrated in Supplementary Fig. 2, panel A, examples 1 and 2 (available online). In addition, if the predicted groups are obtained from a cross-validation procedure, conventional statistical tests are invalid because of a dependency among predictions (16), and permutation testing should be used (15).

However, the prediction accuracy with its statistical significance alone is insufficient if one is to obtain a complete picture of the classifier's predictive ability and its potential clinical utility. The number of true and false positives and true and false negatives should be presented, allowing the calculation of sensitivity and specificity or positive and negative apparent predictive values. Sensitivity and specificity are clinically more meaningful than global accuracy because they yield information on how the classifier behaves in each outcome category. Receiver-operating characteristic curves plotting the sensitivity and specificity obtained for multiple cutoff points also provide valuable information as to the performance of a classifier.

From a clinical perspective, the prevalence of each outcome category is important. A given classifier, with its sensitivity and specificity, yields different prediction accuracies in settings with different prevalences for the outcome categories. This is illustrated in Supplementary Fig. 2, panel A, examples 3 and 4 (available online). The most directly useful indicators are probably the positive and negative apparent predictive values because they take into account the prevalence of the outcome categories and can be directly linked to the specific clinical purpose of the prediction. For instance, to ensure that patients are not denied access to a curative but toxic treatment, the priority lies in maximizing the negative apparent predictive value. Imbalance between outcome categories should always be considered because there are situations in which an observed prediction accuracy of 90% may not be as valuable as it first appears, as shown in Supplementary Fig. 2, panel A, examples 5 and 6 (available online). Because the prevalence of the outcome categories in the reported study may not reflect the prevalence in clinical practice, authors should be careful to label predictive values as "apparent" if they are reported.

Most studies with survival or disease-free survival outcomes were analyzed as binary outcome classification. Classifier performance should be presented in keeping with the way the classifier has been trained. The performance of a classifier trained to predict a "dead or alive" binary outcome should be presented in a contingency table of these two categories for both true and predicted status. Using survival analysis to assess prediction accuracy can sometimes be misleading. First, survival analysis may indicate a statistically significant difference in survival even if the classifier poorly predicts the binary outcome. Second, the transformation of time-to-event data into binary outcome data may induce distortions that result in predicted groups with a spurious but statistically significant difference in survival. Examples of such biases are presented in Supplementary Fig. 2, panel B (available online), with a more detailed explanation given in Supplementary Note 3 (available online).

The willingness to demonstrate that predicted groups have different disease-free or overall survival probabilities is logical for outcomes such as relapse or death. With such an objective, the classifier should be trained to predict risk groups rather than a binary outcome, i.e., using directly the information from time-to-event data rather than transforming it into binary variables. These methods are available and have been implemented in freely available software packages for microarray analysis (17–19).

Developmental studies should also begin to address clinical utility by demonstrating clear evidence of the classifier's ability to improve the prediction accuracy of standard prognostic factors. Claims for the utility of classifiers based on a *P* value below a threshold of statistical significance in a multivariable model are spurious, however. Clinical relevance should be addressed by examining outcome of the new system within the levels of the standard system or by comparing predictive abilities of the standard system with and without the new one (20). Such a comparison might be biased, however, if the standard prognostic system has been used for selecting patients.

Guidelines

The above comments are the justification for a proposal of guidelines for the statistical analysis and reporting of microarray studies for clinical outcomes. Such guidelines seem necessary for a variety of reasons. First, microarray studies are a fast-growing area for both basic and clinical research with an exponentially growing number of publications. Second, as demonstrated by our results, common mistakes and misunderstandings are pervasive in studies published in good-quality, peer-reviewed journals. Third, although there is obviously a need for them, no practical guidelines are currently available. Textbooks on statistical analysis for microarrays, some of them accessible to nonstatisticians, have been published (21). However, we believe that comments and guidelines rooted in what is actually presented in published studies could be of substantial added value for authors, reviewers, editors, and readers and closely meet their needs. Most of the comments will apply to a broad range of scientific studies, including those outside of cancer.

To make these guidelines as practical as possible, we present them as a checklist of "Do's and Don'ts," in Table 3. We believe that following these guidelines should substantially improve the quality of analysis and reporting of microarray investigations. This list is intended, however, to be neither a "Thou Shalt Not" prescription list nor a self-sufficient toolbox to conduct statistical analyses. Rather it should be viewed as a simple and common basis for discussion among people involved in conducting, analyzing, reporting, and interpreting microarray investigations.

References

(1) Butte A. The use and analysis of microarray data. Nat Rev Drug Discov 2002;1:951–60.

(2) Perez EA, Pusztai L, Van de Vijver M. Improving patient care through molecular diagnostics. Semin Oncol 2004;31 Suppl10:14–20.

(3) Pusztai L, Symmans FW, Hortobagyi GN. Development of pharmacogenomic markers to select preoperative chemotherapy for breast cancer. Breast Cancer 2005;12:73–85.

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

(5) Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 2005;365:488–92.

(6) Ntzani EE, Ioannidis JP. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. Lancet 2003;362:1439–44.

(7) Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95:14863–8.

(8) Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. Nat Rev Cancer 2004;4:309–14.

(9) Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003;95:14–8.

(10) Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med 1990;9:811–8.

(11) Korn EL, Troendle JF, McShane LM, Simon R. Controlling the number of false discoveries: application to high-dimensional genomic data. J Stat Plan Inference 2004;124:379–98.

(12) Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 2003;100:9440–5.

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

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

(15) Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. J Comput Biol 2002;9:505–11.

(16) Lusa L, McShane LM, Radmacher MD, Shih JH, Wright G, Simon R. Appropriateness of some resampling-based inference procedures for assessing performance of prognostic classifiers derived from microarray data. Stat Med 2006; [Epub ahead of print].

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

(18) Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol 2004;2:E108.

(19) Simon R, Lam A. BRB-ArrayTools User's Guide Release 3.5, National Cancer Institute. Available at http://linus.nci.nih.gov/brb.

(20) Kattan MW. Judging new markers by their ability to improve predictive accuracy. J Natl Cancer Inst 2003;95:634–5.

(21) Simon R, Korn EL, McShane LM, Radmacher MD, Wright G, Zhao Y. Design and analysis of DNA microarray investigations. New York (NY): Springer; 2004.

(22) Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001;98:5116–21.

(23) Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.

(24) Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 2002;99:6567–72.

(25) Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 2001;29:365–71.

Notes

A. Dupuy received grants from Association pour la Recherche contre le Cancer, Bourse Lavoisier du Ministère des Affaires Etrangères, and Fondation René Touraine.

The funding sources had no role in study design, collection of data, analysis and interpretation of data, writing the report, or the decision to submit the paper to publication.

Manuscript received March 13, 2006; revised October 24, 2006; accepted December 1, 2006.