STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:1102–1113
Published online 5 June 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2598

Appropriateness of some resampling-based inference procedures for assessing performance of prognostic classifiers derived from microarray data

Lara Lusa[1,2,‡], Lisa M. McShane[3,*,†], Michael D. Radmacher[4,§], Joanna H. Shih[3,¶], George W. Wright[3,‖] and Richard Simon[3,**]

[1]*Department of Experimental Oncology, Istituto Nazionale per lo Studio e la Cura dei Tumori, Milano, Italy*
[2]*Molecular Genetics of Cancer Group, Fondazione Istituto FIRC di Oncologia Molecolare (IFOM), Milano, Italy*
[3]*Biometric Research Branch, National Cancer Institute, Bethesda, MD 20892-7434, U.S.A.*
[4]*Center for Biostatistics, The Ohio State University,Columbus, OH 43210, U.S.A.*

SUMMARY

The goal of many gene-expression microarray profiling clinical studies is to develop a multivariate classifier to predict patient disease outcome from a gene-expression profile measured on some biological specimen from the patient. Often some preliminary validation of the predictive power of a profile-based classifier is carried out using the same data set that was used to derive the classifier. Techniques such as cross-validation or bootstrapping can be used in this setting to assess predictive power, and if applied correctly, can result in a less biased estimate of predictive accuracy of a classifier. However, some investigators have attempted to apply standard statistical inference procedures to assess the statistical significance of associations between true and cross-validated predicted outcomes. We demonstrate in this paper that naïve application of standard statistical inference procedures to these measures of association under null situations can result in greatly inflated testing type I error rates. Under alternatives of small to moderate associations, confidence interval coverage probabilities may be too low, although for very large associations coverage probabilities approach their intended values. Our results suggest that caution should be exercised

*Correspondence to: Lisa M. McShane, Biometric Research Branch, EPN 8126, National Cancer Institute, 6130 Executive Blvd., Bethesda, MD 20892-7434, U.S.A.
†E-mail: Lm5h@nih.gov
‡E-mail: lara.lusa@ifom-ieo-campus.it
§E-mail: radmacher.2@osu.edu
¶E-mail: jshih@mail.nih.gov
‖E-mail: wrightge@mail.nih.gov
**E-mail: rsimon@mail.nih.gov

Contract/grant sponsor: Ministero dell'Istruzione, dell'Università e della Ricerca; contract/grant number: PRIN 2003133820
Contract/grant sponsor: Associazione Italiana per la Ricerca sul Cancro
Contract/grant sponsor: Istituto Superiore di Sanità

Received 4 May 2005
Accepted 12 April 2006

in interpreting some of the claims of exceptional prognostic classifier performance that have been reported in prominent biomedical journals in the past few years.

KEY WORDS: cross-validation; microarray; gene-expression; classification; resampling; molecular profiling

1. INTRODUCTION

A frequent goal in gene-expression microarray clinical studies is to develop a multivariate classifier of disease outcome [1–9]. In these studies, gene-expression microarray assays are performed on tissue or other biological material from patients for whom clinical outcomes such as survival are known. The results of the microarray assays are thousands of gene-expression measures, comprising a 'profile', for each of the patient samples assayed. Patient disease outcome might be classified, for example, as 'good' (long survival) or 'bad' (short survival). Mathematical methods are applied to the expression profile data to develop a multivariate classifier to predict disease outcome. For example, van't Veer *et al.* [1] conducted gene-expression microarray analyses on breast tumours and used the data from 78 of the lymph node-negative tumours to build a 70-gene classifier of clinical outcome; they reported it had excellent ability to distinguish between breast cancer patients who did *versus* did not develop distant metastases within 5 years. Beer *et al.* [3] developed a 50-gene risk index using gene-expression profiles from 86 primary lung adenocarcinomas and demonstrated that their risk index could separate patients into subgroups with distinct overall survival probabilities.

A common approach to assessing classifier performance is to compute prediction accuracy, defined as the proportion of samples correctly classified. Ideally, classifier performance would be assessed on a completely independent set of patient specimens, but often sufficiently large numbers of suitable specimens with associated clinical outcome data are not available. The alternative is to estimate prediction accuracy using the same data from which the classifier was derived; however, this requires proper application of resampling methods such as cross-validation [10] or bootstrapping [11, 12] in order to avoid seriously overestimating the prediction accuracy. Alternatively, some authors have chosen to evaluate classifier performance by estimating the association of the true (known) classes with the classes predicted from the cross-validated classifier and performing a test of the statistical significance of that association. For example, this is an approach that was taken in the study by van't Veer *et al.* [1] that has received much attention. In the simplest case of a two-class prediction problem, this measure of association might take the form of an odds ratio calculated from a 2×2 table, as the one displayed in Table I, with one dimension of the table representing the true class (Class 1 *versus* Class 2) and the other dimension representing the cross-validated classifier-predicted class (Class 1 *versus* Class 2).

A similar approach is to perform a logistic regression analysis considering the true class as the dependent binary variable and the cross-validated predicted class as an independent variable in the model, and then test the regression coefficient. Cox proportional hazards regression could also be used to examine the association between the predicted class indicators and actual survival times. A potential advantage of the logistic and Cox regression approaches is their flexibility to allow adjustment for other covariates, but here we consider only the simple 2×2 table approach with no additional covariates. Interested readers are referred to a technical report by the authors (ftp://linus.nci.nih.gov/pub/techreport/Lusa_manuscript_SIM-5-03-05.pdf)

DOI: 10.1002/sim

Table I. 2×2 table for estimation of odds ratio.

		True-class	
		Class 1	Class 2
CV-class	Class 1	a	b
	Class 2	c	d

that examines the performance of the Cox regression approach using cross-validated class indicators as predictors in the regression model.

The questions we explore in this paper are whether standard statistical inference procedures applied to the odds ratio measuring the association between true and cross-validated predicted classes are valid. As we will demonstrate through a series of selected simulation studies, these naïve inference procedures for testing the significance of measures of association can suffer from severely inflated type I errors and poor confidence interval coverage. Furthermore, our simulations will clearly demonstrate the difficulty in interpreting measures of association such as the odds ratio for purposes of gauging performance of a classifier.

2. METHODS

2.1. Class prediction method

Many methods have been proposed for constructing multivariate predictors of class membership using microarray data, including linear and quadratic discriminant analysis, logistic regression, decision trees, support vector machines, and numerous others [13, 14]. Interestingly, simple classification methods such as diagonal linear discriminant analysis have been shown to work well for microarray data [13], where a very large number of measured characteristics, compared to the number of subjects, are available. For purposes of our simulation studies, we use diagonal linear discriminant analysis, but we expect the results would be similar if we were to use other class prediction methods.

In brief, diagonal linear discriminant analysis is performed as follows. Suppose we have a collection of n subjects. Some of these subjects are known to belong to Class 1 (e.g. poor prognosis), and the rest belong to Class 2 (e.g. good prognosis). Let $x_{ij} =$ measurement of the jth characteristic (e.g. gene-expression value) on the ith subject where these measurements collectively form the gene-expression profile for subject i. Apply a feature selection step to reduce the number of candidate predictor variables to a limited set of G genes that are the most informative about the class distinction. For subject i we denote the set of selected features by $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iG})$. For example, feature selection might be accomplished using univariate two-sample t-tests to test, for each gene, if its mean expression level differs between the two prognosis classes. Let $\bar{x}_j^{(1)}$ and $\bar{x}_j^{(2)}$ denote the mean expression of gene j in Classes 1 and 2, respectively. The value s_j^2 denotes the pooled estimate of the within class variance for gene j. The diagonal linear discriminant rule assigns a new sample, represented by a vector $\mathbf{x}^*$ of expression measurements, to Class 1 if

$$\sum_{j=1}^{G} \left[\frac{(x_j^* - \bar{x}_j^{(1)})^2}{s_j^2} \right] \leqslant \sum_{j=1}^{G} \left[\frac{(x_j^* - \bar{x}_j^{(2)})^2}{s_j^2} \right]$$

DOI: 10.1002/sim

and otherwise the new sample is assigned to Class 2. In this formula, x_j^* denotes the expression for gene j in the new sample to be classified.

2.2. Cross-validation

The entire linear discriminant analysis procedure, including the feature selection step, is subjected to cross-validation in order to obtain cross-validated class predictions. K-fold cross-validated class predictions are obtained by dividing the data into K parts. One of the K parts is set aside (test set) and a prediction rule is built on the remaining data (training set). The procedure is repeated until all specimens are included in a test set exactly once and their class membership is predicted using the prediction rule developed on the training set that excludes that test set. A special case of K-fold cross-validation is leave-one-out (LOO) cross-validation in which there are n test sets, each consisting of a single subject. Leave-one-out cross-validation has been described as a logical choice for relatively small sample sizes [15] and has been frequently used in microarray studies. At the completion of the cross-validated classification process, each subject has an associated true class membership and a cross-validated predicted class membership.

2.3. Cross-validated odds ratio as a measure of association

One measure of the association between the cross-validated classifier-predicted class (CV-class) and the true class is the odds ratio formed from a 2×2 table such as the one displayed in Table I.

The usual estimate of the log odds ratio is the logarithm of the cross-product ratio, or $\log(ad/bc)$, and its standard error is calculated as $(1/a + 1/b + 1/c + 1/d)^{1/2}$. Results in this paper are based on the commonly used modifications of the log odds ratio and its standard error, $\hat{\psi} = \log((a+0.5)(d+0.5))/((b+0.5)(c+0.5))$ and $\text{s.e.}(\hat{\psi}) = (1/(a+0.5)+1/(b+0.5)+1/(c+0.5)+1/(d+0.5))^{1/2}$, which avoid problems handling zero cell counts. Ignoring for the moment the fact that the CV-class designations are data derived, a typical test of no association (odds ratio equals one or log odds ratio equals zero) would be based on the statistic $z = \hat{\psi}/\text{s.e.}(\hat{\psi})$ which, under standard conditions, is assumed to have an approximate standard normal distribution under the null hypothesis that the true odds ratio is equal to 1. A 95 per cent confidence interval is given by $\hat{\psi} \pm 1.96 \times \text{s.e.}(\hat{\psi})$. Calculations similar to these were performed in the papers by van't Veer *et al.* [1] and van de Vijver *et al.* [2].

2.4. Data simulation

To simulate data under the null case, gene-expression profiles for each patient were generated independently of the patients' survival times. Expression measurements for each of 10 000 genes were generated independently from the standard normal distribution. Survival times were generated independently from an exponential distribution with parameter lambda $= -\log(0.5)/10$ which translates to an overall survival probability of 50 per cent at 10 years.

Under the alternative case, survival times were generated to depend on gene-expression profiles. For each patient, 9900 genes were generated independently from a standard normal distribution, while the remaining 100 genes were generated independently from a normal distribution with mean μ_1 and variance 1 for half of the patients and from a normal distribution with mean μ_2 and variance 1 for the other half. The expression measures for these 100 genes were averaged for each patient to produce scores $s_1, s_2, \ldots, s_n$. These scores were then used as the mean parameter (on the log-scale) of the log normal distribution with variance 1, from which survival times were

DOI: 10.1002/sim

simulated. Paired values of μ_1 and μ_2 were chosen so that on average half of the subjects would have a survival time longer than 10 years. The larger the difference between μ_1 and μ_2, the stronger is the association between gene-expression and survival.

To simulate the classifier building procedure, patients were divided into poor and good prognosis groups on the basis of their observed survival times. Subjects with observed survival time shorter than 10 years were assigned to poor prognosis class; others were assigned to the good prognosis class. Survival times greater than 20 years were censored at 20 years. These assigned outcome classes represented the 'true' prognostic classes. On average, specimens were equally distributed among the two classes due to the choice of parameters of the distributions used to generate survival times. Univariate two-sample pooled variance t-statistics comparing the good *versus* poor prognosis groups were computed for each gene. For the null case simulations, either the 10 or 100 genes with largest absolute t-statistics were selected as the 'informative features' to be used in building the classifier. For the alternative case simulations, the top 100 most informative genes (by t-test) were selected. Fisher's diagonal linear discriminant analysis was used to build the multivariate prediction rule using the informative genes to classify the specimens into the two prognosis groups. This entire classifier building process was embedded in a cross-validation loop. For each training set, informative genes were re-selected, the classifier was re-calculated, and the classifier was used to make predictions on the test set. For the alternative case simulations, note that the 100-gene sets selected as the informative features in the training data sets were not guaranteed to be the genes that were truly generated from two different distributions for the good and poor prognosis groups. At the end of each simulated cross-validation, there were true and cross-validated predicted prognosis classes assigned to each patient.

The simulation study was conducted using code written in the R language (http://www.r-project.org/). All simulations were repeated 10 000 times. For the null cases, parameters were varied as follows: 30, 50, 100 or 500 subjects; 10 or 100 most informative genes; LOO, 10-fold or 5-fold cross-validation. For the alternative cases, the number of subjects was always 100, 100 most informative features were selected, and only LOO cross-validation was considered. While we would not usually advocate building classifiers using sample sizes as small as 30 or 50, we include them in our simulations because they cover the range of sample sizes that have been used in published studies that developed microarray-based classifiers.

Under the null case, gene-expression profiles are generated from the same distribution for all patients, and class membership is defined independently from gene-expression profiles; therefore, there should be no association between true and CV-class membership (log odds ratio $= 0$). Our simulation studies examine for potential bias in the estimates and problems with the level of tests of hypotheses of no association.

Under the alternative case, we expect there will be some association between the CV-class and true class and survival (a non-zero log odds ratio). Due to the complexity of the classifier derivation, the true values of the log odds ratios and of the misclassification error rate had to be empirically determined. The true quantities were obtained through an 'inner' simulation loop, where at each 'outer loop' of the primary simulation we simulated 100 data sets of 100 subjects each from the same population from which the original sample, on which the classifier was developed, was drawn. The classifier derived on the full original data set was applied to each new ('inner loop') data set. For each of the 100 'inner loop' data sets, using classifier developed on full original sample, predicted class memberships were obtained and log odds ratios and misclassification error rates were estimated. These estimates were then averaged over the 100 inner loop data sets to empirically determine the true log odds ratio and misclassification error rates. These true quantities

DOI: 10.1002/sim

were compared to the cross-validated estimates obtained in the outer loop in order to estimate bias and, for log odds ratio, confidence interval coverage. Confidence interval coverage was broken into components, recording how often the true value falls completely below the lower confidence bound (overestimation) and how often the true value falls completely above the upper confidence bound (underestimation).

3. RESULTS

3.1. Null case

Table II presents simulation results for the log odds ratio estimates calculated under a null situation using various cross-validation methods.

The most striking findings are that all of these cross-validation methods yielded log odds ratio estimates with SDs substantially larger than the naïve standard error estimate, $\text{s.e.}(\hat{\psi}) = (1/(a+0.5) + 1/(b+0.5) + 1/(c+0.5) + 1/(d+0.5))^{1/2}$, that assumes all observations are independent. Additionally, we note that the use of 10-fold or 5-fold cross-validation resulted in considerably smaller estimated SDs of the log odds ratio estimates. This may be related to the claim in the context of prediction error estimation that LOO cross-validation often results in estimates with large variance [11]. However, it is interesting that the degree to which the true SD is inflated relative to the naïve standard error under LOO cross-validation decreases with decreasing sample size (for example, inflation factors 3.3, 2.4, 1.9, and 1.4 for sample sizes 500, 100, 50, and 30, respectively, using 100 most informative genes in the classifier). For 5-fold and 10-fold cross-validation with 100 most informative genes used in the classifier, the variance inflation is not as strongly related to sample size. When 10 most informative genes are used, the variance inflation seen with 5-fold and 10-fold cross-validation is roughly independent of sample size, while for LOO cross-validation the inflation remains a function of sample size. This suggests that both the complexity of the model and the per cent overlap between LOO training sets influence degree of variance inflation. In part, this may be explained by the fact that for LOO cross-validation the proportion of overlapping observations ($(n-2)/(n-1)$) between any two LOO training sets increases with sample size (n), and this increases the degree of correlation between LOO class predictions.

The impact of the variance inflation is exhibited most dramatically in the level of the tests of significance of the log odds ratio. For example, if one were to use LOO cross-validation for a study of 100 subjects, a nominal 5 per cent two-sided z-test would reject the null hypothesis an estimated 41 per cent of the time (23 per cent lower rejection, 18 per cent upper rejection). The 18 per cent upper rejection rate is probably of greatest concern, as these might represent classifiers likely to be falsely reported as promising, whereas classifiers exhibiting negative association with truth are unlikely to ever be published. The upwardly biased testing level for LOO cross-validation, related to variance inflation, is exacerbated with a larger sample size. For a study of 500 subjects, the estimated rejection rate increased to nearly 55 per cent. Although the performance of the test is not as bad when 5-fold or 10-fold cross-validation is used, the rejection rates for a study with sample size 100 still exceed the nominal values by an unacceptably large margin.

A further interesting observation is that for all methods, the mean estimated log odds ratio approached the correct value of zero as the sample size increased but was strongly biased for small sample sizes. The bias was less when 10 most informative genes were used in the predictor rather than 100, but it was still substantial. The estimated median log odds ratio estimates suggested

DOI: 10.1002/sim

Table II. Null case results for estimated odds ratio relating true and cross-validated predicted outcome class for studies with varied numbers of subjects and using different cross-validation methods.

		Mean log odds ratio estimate*		Median log odds ratio estimate		Estimated SD of log odds ratio estimate		Mean of naïve standard error† of log odds ratio estimate		Estimated rejection rate (in %) for nominal 5% two-sided z-test‡ (nominal 2.5% per tail)				Mean mis-classification rate (in %) (SD)	
Number of selected features (genes)		100	10	100	10	100	10	100	10	100		10		100	10
Cross-validation method	Number of study subjects									Lower rejection rate	Upper rejection rate	Lower rejection rate	Upper rejection rate		
LOO	500	−0.018	−0.003	0.003	0.055	0.594	0.699	0.181	0.182	28.0	26.9	28.4	32.4	50.10 (7.24)	49.96 (8.47)
LOO	100	−0.103	−0.054	−0.079	0.014	1.009	1.136	0.427	0.418	23.2	17.6	24.1	22.8	49.97 (11.46)	50.05 (13.15)
LOO	50	−0.267	−0.116	−0.220	−0.024	1.330	1.450	0.700	0.623	18.6	11.2	21.1	17.9	49.67 (13.56)	49.77 (15.92)
LOO	30	−0.505	−0.259	−0.505	−0.211	1.545	1.730	1.090	0.865	14.6	5.1	18.3	13.6	49.54 (15.57)	50.03 (18.19)
10-fold	500	−0.010	−0.005	−0.011	−0.007	0.274	0.281	0.180	0.179	10.7	9.0	10.8	10.4	50.02 (3.40)	50.02 (3.50)
10-fold	100	−0.099	−0.034	−0.091	−0.035	0.637	0.643	0.419	0.407	12.7	7.4	11.4	9.6	49.96 (7.54)	49.95 (7.84)
10-fold	50	−0.286	−0.121	−0.306	−0.148	0.972	0.944	0.679	0.596	12.2	4.8	11.9	8.1	49.97 (10.49)	50.06 (11.03)
10-fold	30	−0.478	−0.231	−0.505	−0.254	1.273	1.280	1.041	0.822	11.0	2.5	10.6	8.0	50.04 (13.36)	50.05 (14.23)
5-fold	500	−0.009	−0.002	−0.006	−0.005	0.237	0.229	0.179	0.179	7.2	6.2	6.0	6.4	49.99 (2.95)	49.98 (2.85)
5-fold	100	−0.100	−0.039	−0.098	−0.039	0.546	0.519	0.419	0.406	8.5	4.6	6.3	5.7	49.95 (6.54)	49.88 (6.39)
5-fold	50	−0.280	−0.119	−0.278	−0.153	0.841	0.754	0.675	0.591	8.5	2.5	6.6	4.4	49.94 (9.26)	50.05 (9.00)
5-fold	30	−0.462	−0.226	−0.505	−0.254	1.093	1.020	1.016	0.808	7.4	1.2	5.9	4	50.06 (11.90)	50.03 (11.66)

*$\hat{\psi}$ as defined in Section 2.3.

†s.e.($\hat{\psi}$) as defined in Section 2.3.

‡Based on the statistic $z = \hat{\psi}/\text{s.e.}(\hat{\psi})$ as defined in Section 2.3.

DOI: 10.1002/sim

a trend of slightly more departure from the true value of zero for 5-fold and 10-fold cross-validation compared to LOO cross-validation, but there were no statistically significant differences in bias based on the mean log odds ratio estimates. Although LOO cross-validation is known to produce nearly unbiased estimates of prediction accuracy ($(a+d)/n$ in the notation of Table I), the highly non-linear log-odds ratio does not enjoy this same near-unbiasedness. Through a series of additional simulation studies (not shown) we found that this bias could be largely removed by forcing all simulated data sets to have equal numbers of subjects in the good and poor prognosis groups and using leave-two-out cross-validation (one per group). This suggests that the odds ratio may be more sensitive than the prediction accuracy to violations of the equal prior assumption of the simple version of diagonal linear discriminant analysis we used. However, the serious problems with variance inflation persisted regardless of the presence or absence of bias.

3.2. Alternative case

Table III presents results demonstrating properties of the odds ratio estimates, including confidence interval coverage, under alternative cases in which there was a positive association between gene-expression profiles and survival outcomes.

Table III shows that for extremely large odds ratios, the confidence interval coverage is not too far from the intended coverage probability. However, for moderate or smaller odds ratios, estimates can be biased and confidence interval coverage can be quite poor. Note that the SDs of the estimated log odds ratios are very large relative to the magnitude of the log odds ratios for a sample of size 100 which is typical of the size of many gene-expression microarray profiling studies. This implies that even if the confidence intervals had correct coverage probabilities, the large variance of the estimators may result in very wide confidence intervals. These wide confidence intervals indicate that studies with sample size around 100 may not yield sufficiently precise association estimates to distinguish a clinically useful prognostic indicator from a minimally informative indicator. In addition, it is clear that the odds ratio can provide a misleading impression of the performance of the predictor. For example, an odds ratio as large as 15 or 16 (third case presented in Table III) would appear extremely impressive in the context of an epidemiologic study, but we see from Table III that the corresponding classifier misclassification rate was a rather unsatisfying 20 per cent.

4. DISCUSSION

We demonstrated that assessing the strength of a classifier by testing the statistical significance of the odds ratio relating true and cross-validated predicted classes or calculating confidence intervals for that odds ratio is problematic. When a classifier is truly uninformative (null case), application of standard inference procedures to test for significance of the association when cross-validation has been used to determine predicted classes carries with it a very high likelihood of obtaining false positive statistical significance, and reported p-values will be too small. If there is some modest predictive value in the data-derived classifier, confidence intervals for the true association between predicted and true class may be very wide and not have the reported coverage properties. Only when the true association is very large will the confidence interval coverage be approximately correct. The problems arise from the fact that the data pairs (CV-class, true class) are not independent across subjects, and their dependency derives from repeated re-use of the true classes in the cross-validation process. This dependency violates the assumptions of the standard

DOI: 10.1002/sim

Table III. Properties of odds ratio estimates and associated confidence intervals under situations with various degrees of association between true and predicted outcomes for sample size of 100 when leave-one-out cross-validation is used.

Parameters used in data generation (μ_1, μ_2)*	True† log odds ratio (odds ratio)	Mean log odds ratio estimate	Median log odds ratio estimate	SD of log odds ratio estimate	Mean of naïve standard error of log odds ratio estimate	Estimated rates (in %) of nominal 95% CI‡ non-coverage (nominal 2.5% per tail)		Mean rate of mis-classifications (in %) (SD)	True† rate of mis-classifications (in %)
						CI<true (under-estimate)	CI>true (over-estimate)		
(3.58, 1.02)	4.374 (79.394)	4.384	4.259	0.706	0.688	4.0	0.6	10.1 (3.04)	10.1
(3.34, 1.27)	3.453 (31.601)	3.463	3.398	0.574	0.569	3.5	1.2	15.1 (3.58)	15.1
(3.14, 1.46)	2.759 (15.785)	2.761	2.708	0.506	0.504	3.2	1.3	20.1 (4.02)	20.1
(2.98, 1.63)	2.113 (8.269)	2.109	2.157	0.619	0.462	5.2	1.9	25.9 (6.16)	25.9
(2.83, 1.78)	1.094 (2.997)	1.045	1.135	0.960	0.428	15.2	6.6	37.2 (10.84)	37.0
(2.69, 1.62)	0.263 (1.301)	0.166	0.180	1.000	0.423	22.5	17.4	47.05 (11.47)	46.9
(2.30, 2.30)	0.001 (1.001)	−0.113	−0.089	1.010	0.427	23.2	17.4	50.1 (11.50)	50.0

*See Section 2.4 for description of data generation methods and definitions of μ_1 and μ_2.

†True values determined empirically as described in Section 2.4.

‡Calculated as $\hat{\psi} \pm 1.96 \times \text{s.e.}(\hat{\psi})$ as described in Section 2.3.

DOI: 10.1002/sim

statistical procedures for performing tests and constructing confidence intervals for the measures of association and its effects are seen most strikingly when true associations are absent or no more than modest. In addition, our results emphasized a point made by others [16, 17] that measures of association such as an odds ratio are generally poor gauges of classifier performance.

The results summarized above suggest that claims regarding exceptional classifier performance made in some gene-expression microarray papers published in prominent biomedical journals should be interpreted cautiously. For example, in the paper by van't Veer *et al.* [1], the fully cross-validated odds ratio is reported to be 15 with p-value 4.1×10^{-6} and 95 per cent CI 4-56. If the true value is close to the point estimate or larger, then the inference statements are accurate. However, because we do not know the true value and the CI is very wide, we cannot rule out the possibility that the true value is substantially smaller than the point estimate. If the true value of the odds ratio is no more than modest in size, then the results could be subject to some of the problems of poor confidence interval coverage highlighted in this paper. Subsequent independent validation studies, as presented in the second through fourth odds ratios presented in Table II of the paper by van de Vijver *et al.* [2], lend support to the original odds ratio of 15, and therefore, the accuracy of the confidence intervals. Nonetheless, all of the confidence intervals are wide and include modest-size odds ratios such as 3 or 5 as well as some very large odds ratios, so one still cannot reliably answer the question of whether the classifier is a powerful clinical tool or not.

Several authors [3, 18–21] besides van't Veer *et al.* [1] have attempted inference within a cross-validation context. We found examples of studies in which cross-validated predicted classes were used to define groups for log rank tests or for tests of odds ratio, or in which cross-validation was used to define risk scores that were tested for statistical significance as part of a Cox regression model. Prior to initiating a study, one has no information about the likely effect size (odds ratio, regression coefficient, or survival curve separation), and therefore, no indication of whether the problems described in this paper are likely to be encountered if inference in the context of cross-validation such as described in this paper is to be used.

The next question is whether there are satisfactory remedies for these problems. The most important point is to recognize that the prime interest is to evaluate the classifier's predictive accuracy and to determine if the accuracy is better than expected by chance. Cross-validation provides a nearly unbiased estimate of predictive accuracy. Radmacher *et al.* [15] provide a valid permutation test of whether the classifier accuracy is better than expected by chance. Their permutation approach considers many possible permutations of assignments of clinical outcomes to profiles, and calculates for each permuted data set the cross-validated prediction accuracy. The proportion of permutations for which the cross-validated accuracy calculated on the original data set is better (larger) is a valid p-value for testing the null hypothesis that the predictor performance is no better than chance.

If it is desired to assess predictor performance when adjusted for other covariates, the permutation method of Radmacher *et al.* [15] cannot be directly applied. Tibshirani and Efron [22] discuss the idea of 'pre-validation' in logistic regression models in which one of the variables in the model is a predicted class indicator obtained through cross-validation and additional covariates can be incorporated into the regression model. (Without covariates, testing the regression coefficient of the cross-validated predicted class indicator in the logistic model is essentially equivalent to testing the log odds ratio as described in this paper.) They note a problem with the degrees of freedom in the test of the regression coefficient that is related to the problems with type I error rate and confidence interval coverage we observed. We have shown how seriously type I error rates and confidence interval non-coverage rates can be inflated, and we demonstrated the roles that sample

DOI: 10.1002/sim

size, model complexity and method of cross-validation play. The dependency problem we described above is essentially the phenomenon Tibshirani and Efron describe as 'information leak'. They explore a bootstrap method to approximately correct the degrees of freedom for testing regression coefficients. This seems like a promising approach, but would require further investigation to determine how successfully the bootstrap-estimated degrees of freedom can correct for problems in testing levels and confidence interval coverage. Troendle *et al.* [23] demonstrated that bootstrap procedures may not perform well in moderate to small samples of very high dimensional data. In addition, we would be remiss if we did not point out that even if one were able to appropriately correct the problems with the inference procedures, the variances of the measures of association obtained through resampling of typical size gene-expression microarray data sets would still be very large. Also, it would be desirable to base the procedure on a more interpretable alternative to the logistic regression coefficient such as gain in predictive accuracy above predictive accuracy afforded by standard covariates.

In summary, our results provide further evidence that concerns recently expressed [24, 25] about the reproducibility and validity of microarray-based prognostic classifiers are warranted. We applaud the decision of Cardoso *et al.* [26] to conduct a further 'pre-validation' study of the classifier developed and studied in van't Veer *et al.* [1] and van de Vijver *et al.* [2] prior to launching a large trial for definitive validation of that classifier. More and larger independent data sets on which to develop and validate classifiers derived from high-dimensional biologic data are needed if these classifiers are ever to be important clinical tools.

ACKNOWLEDGEMENTS

This study utilized the high-performance computational capabilities of the Biowulf/LoBoS3 cluster at the National Institutes of Health, Bethesda, MD. We gratefully acknowledge the funding provided by Ministero dell'Istruzione, dell'Università e della Ricerca (grant PRIN 2003133820), AIRC (Associazione Italiana per la Ricerca sul Cancro, individual grant to M. A. Pierotti) and an Italy-U.S.A. Fellowship of the Istituto Superiore di Sanità on Oncological Pharmacogenomics—Seroproteomics—to support Dr Lusa's participation in this project.

REFERENCES

1. van't Veer LJ, Dai HY, van de Vijver MJ, He YDD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; **415**(6871):530–536.
2. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AAM, Voskuil DW, Schreiber GJ, Peterse JL, Robert C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R. A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine* 2002; **347**(25):1999–2009.
3. Beer DG, Kardia SLR, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizy ML, Kuick R, Hayasaka S, Taylor JMG, Iannettoni MD, Orringer MB, Hanash S. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine* 2002; **8**(8):816–824.
4. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Staudt LM for the Lymphoma/Leukemia Molecular Profiling Project. The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma. *The New England Journal of Medicine* 2002; **346**(25):1937–1947.
5. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RCT, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine* 2002; **8**(1):68–74.

DOI: 10.1002/sim

6. Tay ST, Leong SH, Yu K, Aggarwal A, Tan SY, Lee CH, Wong K, Visvanathan J, Lim D, Wong WK, Soo KC, Kon OL, Tan P. A combined comparative genomic hybridization and expression microarray analysis of gastric cancer reveals novel molecular subtypes. *Cancer Research* 2003; **63**(12):3309–3316.
7. Bullinger L, Dohner K, Bair E, Frohling S, Schlenk RF, Tibshirani R, Dohner H, Pollack JR. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *The New England Journal of Medicine* 2004; **350**(16):1605–1616.
8. Valk PJM, Verhaak RGW, Beijen MA, Erpelinck CAJ, van Doorn-Khosrovani SBV, Boer JM, Beverloo HB, Moorhouse MJ, van der Spek PJ, Lowenberg B, Delwel R. Prognostically useful gene-expression profiles in acute myeloid leukemia. *The New England Journal of Medicine* 2004; **350**(16):1617–1628.
9. Dave SS, Wright G, Tan B, Rosenwald A, Gascoyne RD, Chan WC, Fisher RI, Braziel RM, Rimsza LM, Grogan TM, Miller TP, LeBlanc M, Greiner TC, Weisenburger DD, Lynch JC, Vose J, Armitage JO, Smeland EB, Kvaloy S, Holte H, Delabie J, Connors JM, Lansdorp PM, Ouyang Q, Lister TA, Davies AJ, Norton AJ, Muller-Hermelink HK, Ott G, Campo E, Montserrat E, Wilson WH, Jaffe ES, Simon R, Yang L, Powell J, Zhao H, Goldschmidt N, Chiorazzi M, Staudt LM. Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *The New England Journal of Medicine* 2004; **351**(21):2159–2169.
10. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute* 2003; **95**(1):14–18.
11. Efron B. Estimating the error rate of a prediction rule: some improvements on cross-validation. *Journal of the American Statistical Association* 1983; **78**(382):316–331.
12. Efron B, Tibshirani R. Improvements on cross-validation: the .632+ bootstrap method. *Journal of the American Statistical Association* 1997; **92**(438):548–560.
13. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene-expression data. *Journal of the American Statistical Association* 2002; **97**(457):77–87.
14. Simon R, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y. *Design and Analysis of DNA Microarray Investigations*. Springer: New York 2004 (Chapter 8).
15. Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene-expression profiles. *Journal of Computational Biology* 2002; **9**(3):505–511.
16. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology* 2004; **159**(9):882–890.
17. Kattan MW. Judging new markers by their ability to improve predictive accuracy. *Journal of the National Cancer Institute* 2003; **95**(9):634–635.
18. Ferguson SE, Olshen AB, Viale A, Barakat RR, Boyd J. Stratification of intermediate-risk endometrial cancer patients into groups at high risk or low risk for recurrence based on tumor gene expression profiles. *Clinical Cancer Research* 2005; **11**(6):2252–2257.
19. McLean LA, Gathmann I, Capdeville B, Polymeropoulos MH, Dressman M. Pharmacogenomic analysis of cytogenetic response in chronic myeloid leukemia patients treated with imatinib. *Clinical Cancer Research* 2004; **10**(7):155–165.
20. Tomida S, Koshikawa K, Yatabe Y, Harano T, Ogura N, Mitsudomi T, Some M, Yanagisawa K, Takahashi T, Osada H, Takahashi T. Gene expression-based, individualized outcome prediction for surgically treated lung cancer patients. *Oncogene* 2004; **23**(31):5360–5370.
21. Onken MD, Worley LA, Ehlers JP, Harbour JW. Gene expression profiling in uveal melanoma reveals two molecular classes and predicts metastatic death. *Cancer Research* 2004; **64**(20):7205–7209.
22. Tibshirani RJ, Efron B. Pre-validation and inference in microarrays. *Statistical Application in Genetics and Molecular Biology* 2002; **1**(1):1–18.
23. Troendle JF, Korn EL, McShane LM. An example of slow convergence of the bootstrap in high dimensions. *The American Statistician* 2004; **58**(1):25–29.
24. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 2005; **365**(9458):488–492.
25. Ioannidis JPA. Microarrays and molecular research: noise discovery? *Lancet* 2005; **365**(9458):454–455.
26. Cardoso F, Piccart MJ, Viale G, Van't Veer L, Saghatchian-d'Assignies M, Glass A, Ellis P, Harris A, Bergh J, Lidereau R, Rutgers E, Delorenzi M, Bogaerts J, Therasse P, Lazar V, Amakrane M, Sotiriou C, Loi S, Buyse M on behalf of the TRANSBIG Consortium. Multi-centre independent validation of a gene prognostic signature for patients with node-negative breast cancer—final results. *Breast Cancer Research and Treatment* 2005; **94**(Suppl 1):S30 (abstract #301).

DOI: 10.1002/sim