Research Article

Statistics in Medicine

Received 20 January 2010, Accepted 13 September 2010 Published online 1 December 2010 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4106

An evaluation of resampling methods for assessment of survival risk prediction in high-dimensional settings[‡,§]

Jyothi Subramanian and Richard Simon[*†]

Resampling techniques are often used to provide an initial assessment of accuracy for prognostic prediction models developed using high-dimensional genomic data with binary outcomes. Risk prediction is most important, however, in medical applications and frequently the outcome measure is a right-censored time-to-event variable such as survival. Although several methods have been developed for survival risk prediction with high-dimensional genomic data, there has been little evaluation of the use of resampling techniques for the assessment of such models. Using real and simulated datasets, we compared several resampling techniques for their ability to estimate the accuracy of risk prediction models. Our study showed that accuracy estimates for popular resampling methods, such as sample splitting and leave-one-out cross validation (Loo CV), have a higher mean square error than for other methods. Moreover, the large variability of the split-sample and Loo CV may make the point estimates of accuracy obtained using these methods unreliable and hence should be interpreted carefully. A k-fold cross-validation with $k=5$ or 10 was seen to provide a good balance between bias and variability for a wide range of data settings and should be more widely adopted in practice. Published in 2010 by John Wiley & Sons, Ltd.

Keywords: microarray data analysis; prognostic signatures; survival analysis; resampling methods; prediction accuracy

1. Introduction

Prediction of prognosis from high-dimensional genomic or proteomic data is a subject of intense research in many disease areas, especially cancer [1–4]. Classifiers that substantially improve the predictive ability of traditional clinico-pathological prognostic factors can help inform improved treatment decisions for individual patients. The outcome variable in prognostic studies is often a time-to-event measure (death or disease recurrence) and the objective is to classify patients into risk groups given the genomic data and other relevant clinico-pathological factors as explanatory variables. However, since most statistical modeling methods were not developed for high-dimensional settings where the number of candidate explanatory variables p is greater than the number of cases $n(p \gg n)$, naïve usage of these methods can lead to serious model overfitting and severely biased performance estimates [5, 6].

Modeling high-dimensional data proceeds through three broad steps: (i) dimension reduction, (ii) model building, and (iii) model assessment. The most commonly used strategy to evaluate the association of a time-to-event response with explanatory variables is Cox's proportional hazards regression [7]. A number of feature selection/dimension reduction/coefficient shrinkage methods have been studied in conjunction with proportional hazards regression to overcome the problem of overfitting with high-dimensional data. These include supervised principal components [8], ridge regression [9], Lasso [10], and least angle regression [11]. Although evaluating a predictive model using fully independent data is always advisable, internal validation is often useful as an intermediate step to obtain an initial assessment of the prediction accuracy. An initial estimate of the predictive accuracy of the model can be obtained using resampling techniques.

Biometric Research Branch, National Cancer Institute, 9000 Rockville Pike, Bethesda, MD 20892-7434, U.S.A.

[*]*Correspondence to: Richard Simon, Biometric Research Branch, National Cancer Institute, 9000 Rockville Pike, Bethesda, MD 20892-7434, U.S.A.*

[†]*E-mail: rsimon@mail.nih.gov*

[§]*Supporting information may be found in the online version of this article.*

Resampling methods are based on a division of the available data into a training set and a test set; while the training set is used for feature selection and model building, the test set is used only for model assessment. Resampling techniques include split-sample, cross validation (CV), and bootstrap-based methods. The study by Molinaro *et al.* [12] provides an extensive comparison of a variety of resampling methods for prediction error estimation in high-dimensional settings with a binary response variable. However, no comparative study on the usefulness of resampling methods for assessing survival prediction models in high-dimensional settings has been reported.

An abundance of prognostic gene expression signatures are being published in many disease areas, and many of them report accuracy estimates obtained through resampling. Hence, a systematic study of resampling methods for assessment of survival models becomes important. In many of the published reports on prognostic signatures, the initial assessment is most often conducted using the split-sample or the leave-one-out (Loo) CV technique (for example, [1, 2, 13, 14]). Hence it is also important to analyze how these methods compare to other simple resampling techniques, such as the k-fold CV.

In this paper, we report the results of a systematic investigation into the behavior of several resampling methods for estimating the accuracy of survival prediction models in high-dimensional settings. These results provide an insight into the behavior of the commonly used techniques and suggest alternatives to be adopted in practice.

2. Methods

2.1. Development of prognosis prediction models

The data consists of a sample of n patients, with follow-up times $t_1, \ldots, t_n$, and event indicators $\delta_1, \ldots, \delta_n$, where $\delta_i = 1$ if the event (recurrence or death) occurred at t_i and $\delta_i = 0$ if the observation was right-censored at t_i. Also available is the $n \times p$ covariate matrix X, representing the gene expression measurements and other clinically important variables. The objective is to predict future survival or disease recurrence using X. In the proportional hazard model [7], the hazard function is modeled as a function of the covariates:

$$h(t|X) = h_0(t)\exp(\beta^{\mathrm{T}} X), \tag{1}$$

where $h_0(t)$ is an unspecified baseline hazard function that is assumed to be the same for all patients and β is the parameter vector. In the traditional setting with $n > p$, the parameter vector and the baseline hazard function are estimated using partial likelihood maximization and the Breslow estimator [15], respectively. However, in situations where $p \gg n$, this can result in a seriously over-fitted model with little predictive value. Three-dimension reduction/coefficient shrinkage strategies that have gained popularity in $p \gg n$ settings are considered in our study:

(i) *Univariate selection (UniCox)*: In this approach, univariate Cox regression analysis is conducted for all covariates considering one covariate at a time. The covariates are then ordered based on the p-value from the score test [16]. Following this, either a pre-fixed number of covariates or covariates whose p-values fall below a threshold are chosen as the reduced feature set. The final step is to fit a multivariate Cox regression model with the selected covariates.

(ii) *Supervised principal components regression (SuperPCR)*: SuperPCR [8] is a modification of conventional principal components regression (PCR), to overcome a possible drawback that the usual principal components may not be associated with the response. In SuperPCR, covariates correlated with survival are first selected through univariate Cox regression. A multivariate Cox regression model is then fit on the first few principal components of this reduced set of covariates.

(iii) *Least absolute shrinkage and selection operator (Lasso)*: Lasso is a penalized estimation approach for regression models that constrains the L1 norm of the regression coefficients [10]. This method is attractive in the $p \gg n$ settings as it simultaneously performs variable selection and shrinkage, by shrinking all regression coefficients towards zero and setting many of them equal to zero, depending on the size of the penalty employed.

2.2. Measures of assessment for prognosis prediction

The most frequently reported measures for assessing the predictive ability of prognostic models are the hazard ratio and the p-value from a log-rank test for the separation of Kaplan–Meier survival curves. Both of these measures have serious limitations. The hazard ratio is a measure of association and not a measure of predictive ability. Strong statistical associations do not necessarily imply that the model can discriminate effectively between patients with good and poor prognosis [17]. The log-rank test requires the specification of a cut-off for defining the high and low risk groups, which

Published in 2010 by John Wiley & Sons, Ltd.
Statist. Med. **2011**, 30 642–653

can be quite arbitrary. Also, this test is invalid for cross-validated measurements [18]. A more appropriate approach for the evaluation of a continuous-valued prognostic factor is the time-dependent ROC curve [19], defined as follows:

Let $r=\beta^{\mathrm{T}}X$ denote the risk score estimated through proportional hazards regression. To construct the time-dependent ROC curve, sensitivity and specificity are considered as time-dependent functions and are defined as

$$\text{sensitivity}\ (c,t)=P\{r>c|T<t\}, \tag{2}$$

$$\text{specificity}\ (c,t)=P\{r\leqslant c|T\geqslant t\}. \tag{3}$$

A plot of sensitivity(c,t) versus 1—specificity(c,t) for all values of the cut-off c leads us to the time-dependent ROC curve, ROC(t) for time t. As outlined in Heagerty *et al.* [19], the conditional probabilities in (2) and (3) can be estimated by the nearest neighbor estimator for the bivariate distribution function F(c,t) as proposed in [20]. The area under the ROC(t) denoted by AUC(t) can then be used as a summary measure for comparing different prognostic models, larger AUC(t)s imply better predictions. For some examples of the use of this approach for comparing prognosis prediction models and for figures illustrating the time-dependent ROC curve, the reader is referred to [21, 22]. The AUC(t) was used as the final quantity of interest in this study for comparing and assessing prognostic models.

2.3. Resampling techniques

Denote the sample of size n consisting of the survival times, event indicators and gene expression measurements for n patients by S_n and let $A(S_n)$ denote the accuracy (as measured by the AUC(t)) of the prognostic model for future samples given the sample S_n. In the absence of an independent large dataset, an estimate for $A(S_n)$ is obtained through resampling from the original sample S_n. We denote the resampling-based estimate of $A(S_n)$ by $\hat{A}^{\mathrm{RS}}(S_n)$.

In the so-called resubstitution estimate/training estimate, $\hat{A}^{\mathrm{resub}}(S_n)$, all points in the sample S_n are used for feature selection, model building as well as estimation of accuracy. Hence, $\hat{A}^{\mathrm{resub}}(S_n)$ is too optimistic, especially with high-dimensional data and complex modeling algorithms.

In the split-sample resampling method, a subset $S_m(m<n)$ is chosen randomly from S_n. The model is trained on points from the complement S_n/S_m and assessed on points in S_m.

In resampling based on the k-fold CV, S_n is partitioned into k subsets, S_{ni}, $i=1,2,\ldots,k$, $(k\leqslant n)$. Each subset is in turn left out during model building. The model is trained on the union of the remaining $k-1$ subsets and predictions are obtained for the left out subset. After the k rounds of training and testing are complete, all the test set predictions are used to estimate the accuracy. There are two ways in which the k-fold CV estimate of AUC(t) can be computed—the pooling strategy and the averaging strategy [23]. In the pooling strategy, all the test set risk-score predictions are first collected and AUC(t) is calculated on this combined set. In the averaging strategy, AUC(t)s are first computed for each test set and are then averaged. In our analysis, the pooling strategy was used to compute the AUC(t).

The Loo CV is a special case of k-fold CV, where $k=n$, i.e. a single observation is left out each time. In this case the AUC(t) can be estimated using only the pooling strategy.

Other resampling methods that have been explored in the case of binary response are Monte Carlo CV and different variations of bootstrap resampling, such as bootstrap CV and the 0.632+ estimator. No substantial improvements over standard CV techniques have been reported for these methods for high-dimensional data in the case of binary response classification problems [12]. Moreover, depending on the number of bootstrap samples taken, the bootstrap methods may be more computationally expensive than even the Loo CV for assessing risk prediction models in $p\gg n$ settings, for small and moderate n. Hence these methods were not evaluated in our study.

2.4. Datasets

(i) *Lung cancer data*: This data is based on a multicenter study on prognostic factors in lung adenocarcinoma (https://caarraydb.nci.nih.gov/caarray/publicExperimentDetailAction.do?expId=1015945236141280) [24]. Only data from the University of Michigan Cancer Center (MI) and the Memorial Sloan-Kettering Cancer (MSK) were considered in our study, leading to 280 samples. Gene normalization and filtering was carried out using BRB-Array tools (developed by Dr Richard Simon and the BRB-ArrayTools Development Team [25], http://linus.nci.nih.gov/BRB-ArrayTools.html), and resulted in 13 538 genes. For our analyses, only the first 5000 genes with the highest overall variance were considered.

(ii) *Diffuse large B-cell lymphoma (DLBCL) data*: This data is based on a prognostic study on DLBCL [1] and contains survival information and expression levels of 7399 genes on 240 patients. The dataset was downloaded from BRB-array tools data archive [26] (http://linus.nci.nih.gov/~brb/DataArchive_New.html). After median normalization, 5000 genes with the largest variance overall were considered for our analyses.

Published in 2010 by John Wiley & Sons, Ltd.
Statist. Med. **2011**, 30 642–653

(iii) *Simulated data, High-signal*: Survival times were generated from a uniform distribution U (2, 200). Each observation was censored with probability 0.2. Expression values for the first 10 genes were then computed as the function of the survival time, t as $(\ln(t)/\beta)+\varepsilon$ with $\beta=1.5$ and $\varepsilon\sim N$ (0, 0.5). These 10 genes form the informative genes. Expression levels for additional 990 genes were generated from a multivariate normal distribution N (μ, Σ), where μ was chosen to be the mean expression of the first 10 genes and sigma was the identity matrix, I_{990}. These 990 genes form the uninformative or noise genes.

(iv) *Simulated data, Null*: Survival times were generated from a uniform distribution U (2, 200). Each observation was censored with probability 0.2. Expression levels for 1000 genes were generated from a multivariate normal distribution $N(0, I_{1000})$. Thus, in this dataset there is no association between the survival times and covariates.

2.5. Analysis

The objective of our study was to investigate the usefulness of various resampling methods for estimating the accuracy of survival prediction models in $p\gg n$ settings on datasets with varying sample sizes and signal-to-noise ratios, while incorporating some form of dimension reduction. The resampling methods evaluated were the split-sample (with two-third of the data for training and one-third for testing), k-fold CV with $k=2$, 5, and 10, and the Loo CV. In high-dimensional settings, it is imperative that feature selection/dimension reduction occurs based only on the training data and is repeated completely within each resampling loop; otherwise substantial bias will be introduced [5]. This is known as complete resampling and was adopted throughout our study. In our simulations, we mimic the situation where a large independent sample is available to estimate the true accuracy of a model. The resampling-based accuracy estimates are then compared to this estimate of true accuracy. All AUC(t)s were evaluated at $t=180$ months. The entire study procedure is outlined in steps 1 to 3. Steps 1–3 were repeated hundred times.

(i) For the real datasets, from the full dataset of size N ($N=280$ for the lung cancer and 240 for the DLBCL dataset), random sample S_n^r of size n was drawn without replacement, where r denotes the rth replication. This was done separately for $n=40$, 80, and 160. Sampling was stratified so that there were equal numbers of event times and censored times in each sample. In the case of simulated datasets, S_n^r was generated as explained in Section 2.4.

(ii) For each resampling strategy, the resampling AUC(t)s, $\hat{A}^{\mathrm{RS}}(S_n^r)$ were computed as outlined in steps (a) to (e):

  (a) Observations from S_n^r were randomly allocated to the training and test set as per the resampling plan.
  (b) For each modeling strategy (UniCox/SuperPCR/Lasso), the coefficient estimates $\hat{\beta}_{\mathrm{train}}$ were obtained using only observations from the training set.
  (c) Risk scores were predicted for observations in the test set as $\hat{\beta}_{\mathrm{train}}^{\mathrm{T}} x_{\mathrm{test}}$.
  (d) In the case of split-sample, the predicted risk scores for the m test set cases was obtained using the above method. In the case of k-fold and Loo CV, steps (b) and (c) were repeated for each partition of S_n^r into the training and test set and the risk scores were predicted for all n cases in S_n^r.
  (e) The ROC(t) and the area under the curve AUC(t) were computed using the predicted risk scores, as outlined in Section 2.2. The R package survival ROC [27] was used to compute the ROC(t) and AUC(t).

(iii) For each modeling strategy, prognostic models were then developed using all the observations in S_n^r. For the simulated datasets, $A(S_n^r)$ was estimated based on generating a large independent test set of 500 observations with the same distribution as the training set. For the real datasets, $A(S_n^r)$ was estimated by the out-of-sample estimate obtained by applying the prognostic model developed using S_n^r to the samples not included in S_n^r. We denote these estimates of $A(S_n^r)$ by $\hat{A}(S_n^r)$.

The expected values of $\hat{A}(S_n)$ and $\hat{A}^{\mathrm{RS}}(S_n)$ were estimated, respectively, as

$$\hat{A}_n=\frac{1}{100}\sum_{r=1}^{100}\hat{A}(S_n^r),\quad \hat{A}_n^{\mathrm{RS}}=\frac{1}{100}\sum_{r=1}^{100}\hat{A}^{\mathrm{RS}}(S_n^r).$$

The bias and MSE were estimated, respectively, as

$$\mathrm{Bias}=\frac{1}{100}\sum_{r=1}^{100}(\hat{A}^{\mathrm{RS}}(S_n^r)-\hat{A}(S_n^r)),\quad \mathrm{MSE}=\frac{1}{100}\sum_{r=1}^{100}(\hat{A}^{\mathrm{RS}}(S_n^r)-\hat{A}(S_n^r))^2.$$

The MSE can be decomposed as a sum of the squared bias and the variance. Hence, for a low MSE, both the bias and variance need to be low. Permutation t-tests were performed to test whether the bias is significantly different from zero. A p-value less than 0.01 was considered significant.

Published in 2010 by John Wiley & Sons, Ltd.
Statist. Med. **2011**, 30 642–653

To fit proportional hazards regression models in the case of UniCox and SuperPCR we used the *coxph* function from the R package *survival* [28]. In the case of UniCox, covariates corresponding to the 10 smallest p-values from univariate Cox regression were chosen for the final multivariate model. For SuperPCR, the first three principal components from 10 covariates with the smallest p-values from univariate Cox regression were selected for the final model. In both cases the p-values were based on the score test. In the case of Lasso, the Lasso implementation for Cox regression available through the R package *penalized* [29] was used. The value of the penalty parameter was chosen so that the resulting model had 10 non-zero coefficients. Ideally, optimized values for these parameters can be obtained using CV within each resampling loop. However, since our study was not designed to be a comparison of the performance of modeling strategies, the parameters were preset and no attempt was made to optimize them.

All computations were carried out using R (version 2.8.0) [30]. The *snow* package [31] was used to parallelize all CV loops.

3. Results

By design, the high-signal and null datasets have the highest and lowest signal-to-noise ratios, respectively. Furthermore, an inspection of the expected AUC(t) estimates, $\hat{A}_n$ (Table I) and a plot of the learning curves for the datasets (Figure S1, Supplementary Information) showed that the four datasets can be arranged in the order of decreasing signal strength as High signal > Lung cancer > DLBCL > Null. Hence, apart from the influence of feature selection methods and sample sizes, we could also study the impact of signal-to-noise ratios on the resampling AUC(t) estimates.

3.1. Distribution of the resampling estimates

The distribution of the AUC(t) estimates is illustrated in Figures 1 and 2 for SuperPCR and in Figures S2 and S3 (Supplementary Information) for UniCox and Lasso, respectively. It can be seen that the distribution of $\hat{A}^{\mathrm{RS}}(S_n)$ is very broad for all the resampling methods as compared to the distribution of $\hat{A}(S_n)$. At $n=40$, the split-sample and Loo CV estimates have the broadest distribution. Although the variance due to resampling decreases as n increases to 160, the variance is still much higher than the variance of $\hat{A}(S_n)$, especially for the split-sample and Loo CV, and especially with the null dataset (Table I).

This variability due to resampling has the undesirable consequence of leading much more frequently to highly pessimistic or highly optimistic resampling AUC(t) estimates than that observed in $\hat{A}(S_n)$ (Figure 3). For example, at very small sample sizes ($n=40$), the split-sample or Loo CV based AUC(t) estimates for the null data can be highly optimistic or pessimistic around 15–20 per cent of the time each. Even when the sample size is increased to 160, pessimistic AUC(t) estimates can be seen to occur around 8–10 per cent of the time for the split-sample and Loo CV methods. Although this effect is greatest when the signal is zero, it can also be observed for real datasets with moderate signals (Figure 4, for the lung cancer dataset).

3.2. Bias due to resampling

The bias in the resampling AUC(t) estimates are compared in Figure 5.

High-signal dataset: The resampling AUC(t) estimates for the high-signal dataset are associated with a significant negative bias. The negative bias reflects the fact that the accuracy of a model based on a dataset of n cases is being estimated using models based on much smaller training sets. The negative bias is greatest for the split-sample and the 2-fold CV, reflecting the larger difference between the actual training set and full dataset for these methods.

Null dataset: The bias due to resampling is not statistically significant for this data, demonstrating that the resampling methods investigated in this study give valid estimates of the true prediction accuracy, i.e. are not systematically optimistic under the null.

Lung cancer and DLBCL datasets: In the case of these real microarray datasets the split-sample and k-fold CV based estimates of accuracy are associated with significant negative bias, especially at the highest sample size when the signal in the training data allows active learning to take place.

In almost all cases where the bias is statistically significant, the magnitude of the bias is greatest for the Lasso. In most cases the estimated expected true AUC(t), $\hat{A}_n$, is higher for Lasso than that for other modeling strategies. However the estimated expected resampling AUC(t), $\hat{A}_n^{\mathrm{RS}}$, is in most cases lower for Lasso. Both these factors contribute to the significant negative bias. It can be seen that for high-signal data, the learning curve is steeper for Lasso than for other modeling methods (Figure S1, Supplementary Information). This shows that while Lasso makes the most effective use of data points, this property also makes Lasso quite sensitive to removal of data points.

Published in 2010 by John Wiley & Sons, Ltd.
Statist. Med. **2011**, 30 642–653

Statistics in Medicine

Table I. Mean and standard deviation of AUC(t) estimates.

n	AUC(t) Estimation method	High signal						Lung cancer						DLBCL						Null					
		Unicox		SuperPCR		Lasso		Unicox		SuperPCR		Lasso		Unicox		SuperPCR		Lasso		Unicox		SuperPCR		Lasso	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
40	$\hat{A}(S_n)$	**0.67**	**0.04**	**0.70**	**0.05**	**0.71**	**0.06**	**0.54**	**0.07**	**0.55**	**0.07**	**0.55**	**0.06**	**0.53**	**0.06**	**0.52**	**0.06**	**0.52**	**0.05**	**0.50**	**0.04**	**0.50**	**0.04**	**0.50**	**0.04**
	Resubstitution	0.81	0.11	0.82	0.08	0.87	0.12	0.87	0.07	0.87	0.07	0.92	0.06	0.87	0.06	0.85	0.07	0.95	0.04	0.90	0.07	0.90	0.07	0.93	0.06
	Loo	0.63	0.13	0.65	0.14	0.62	0.15	0.56	0.14	0.59	0.14	0.57	0.12	0.52	0.14	0.51	0.15	0.52	0.17	0.49	0.18	0.50	0.19	0.49	0.19
	10-fold	0.62	0.12	0.63	0.13	0.60	0.14	0.57	0.11	0.59	0.13	0.51	0.10	0.51	0.12	0.52	0.13	0.52	0.13	0.48	0.15	0.48	0.15	0.47	0.17
	5-fold	0.62	0.11	0.62	0.13	0.57	0.14	0.53	0.09	0.57	0.12	0.52	0.12	0.51	0.09	0.51	0.11	0.51	0.11	0.51	0.15	0.50	0.15	0.50	0.15
	2-fold	0.53	0.13	0.60	0.13	0.55	0.13	0.57	0.16	0.55	0.11	0.53	0.11	0.53	0.11	0.51	0.09	0.52	0.10	0.49	0.12	0.49	0.14	0.47	0.13
	Split-sample	0.57	0.21	0.62	0.20	0.57	0.20	0.55	0.17	0.57	0.17	0.56	0.18	0.46	0.16	0.49	0.18	0.49	0.17	0.52	0.22	0.53	0.22	0.51	0.21
80	$\hat{A}(S_n)$	**0.72**	**0.02**	**0.74**	**0.02**	**0.76**	**0.02**	**0.56**	**0.07**	**0.58**	**0.06**	**0.57**	**0.07**	**0.53**	**0.05**	**0.53**	**0.06**	**0.54**	**0.07**	**0.50**	**0.04**	**0.50**	**0.04**	**0.50**	**0.04**
	Resubstitution	0.77	0.06	0.76	0.06	0.83	0.06	0.84	0.05	0.81	0.07	0.88	0.05	0.82	0.06	0.78	0.05	0.87	0.05	0.89	0.05	0.88	0.05	0.88	0.05
	Loo	0.69	0.07	0.71	0.07	0.70	0.09	0.55	0.13	0.58	0.12	0.59	0.12	0.51	0.11	0.52	0.12	0.51	0.14	0.52	0.16	0.53	0.16	0.51	0.13
	10-fold	0.69	0.07	0.70	0.07	0.67	0.10	0.56	0.10	0.58	0.10	0.51	0.10	0.51	0.09	0.53	0.08	0.52	0.10	0.51	0.12	0.51	0.13	0.51	0.14
	5-fold	0.69	0.06	0.70	0.07	0.65	0.08	0.57	0.10	0.58	0.09	0.52	0.09	0.52	0.08	0.52	0.08	0.52	0.09	0.51	0.12	0.52	0.10	0.51	0.11
	2-fold	0.67	0.08	0.66	0.08	0.63	0.10	0.56	0.09	0.56	0.08	0.53	0.10	0.51	0.08	0.52	0.08	0.53	0.09	0.50	0.09	0.51	0.10	0.50	0.10
	Split-sample	0.60	0.15	0.60	0.14	0.60	0.14	0.57	0.11	0.59	0.11	0.59	0.11	0.50	0.11	0.51	0.11	0.52	0.12	0.50	0.14	0.50	0.13	0.50	0.14
160	$\hat{A}(S_n)$	**0.73**	**0.01**	**0.75**	**0.01**	**0.78**	**0.02**	**0.60**	**0.08**	**0.61**	**0.07**	**0.60**	**0.07**	**0.53**	**0.07**	**0.56**	**0.07**	**0.60**	**0.07**	**0.50**	**0.04**	**0.49**	**0.04**	**0.50**	**0.05**
	Resubstitution	0.75	0.04	0.74	0.04	0.80	0.05	0.80	0.05	0.77	0.05	0.80	0.05	0.77	0.04	0.73	0.04	0.79	0.03	0.85	0.04	0.82	0.05	0.81	0.05
	Loo	0.73	0.04	0.74	0.05	0.75	0.07	0.60	0.11	0.61	0.10	0.63	0.08	0.55	0.09	0.57	0.09	0.59	0.07	0.47	0.13	0.48	0.12	0.48	0.13
	10-fold	0.73	0.04	0.74	0.04	0.71	0.06	0.59	0.08	0.60	0.08	0.56	0.07	0.56	0.07	0.57	0.06	0.58	0.07	0.49	0.09	0.49	0.09	0.49	0.10
	5-fold	0.73	0.04	0.73	0.04	0.71	0.07	0.60	0.07	0.60	0.08	0.53	0.07	0.55	0.06	0.57	0.06	0.57	0.06	0.49	0.07	0.50	0.08	0.48	0.09
	2-fold	0.71	0.05	0.72	0.05	0.69	0.08	0.56	0.08	0.58	0.07	0.54	0.07	0.53	0.06	0.54	0.06	0.54	0.07	0.50	0.07	0.49	0.07	0.50	0.07
	Split-sample	0.69	0.10	0.70	0.09	0.70	0.10	0.60	0.10	0.60	0.09	0.59	0.10	0.54	0.09	0.56	0.09	0.55	0.09	0.50	0.12	0.49	0.11	0.49	0.12

Published in 2010 by John Wiley & Sons, Ltd.

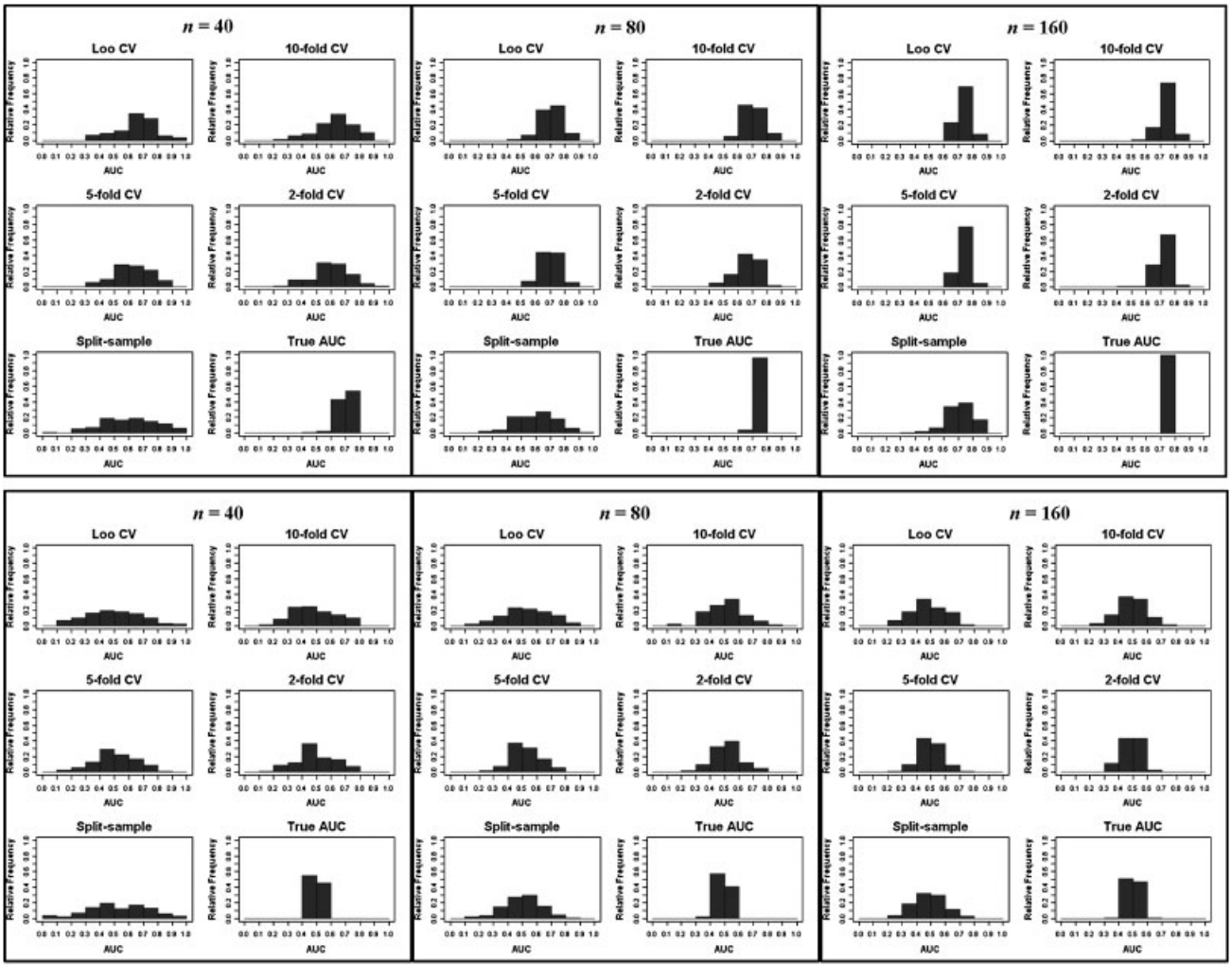

Figure 1. Distribution of the estimated resampling and true AUC(t) for SuperPCR: (top) high-signal data and (bottom) null data.

3.3. MSE due to resampling

The MSEs for the resampling AUC(t) estimates are compared in Figure 6.

High-signal dataset: The split-sample method has the greatest MSE among all resampling methods. The high bias and variance of the split-sample method contributes to this large MSE. At the lowest sample size, 2-fold CV has a higher MSE than Loo, 10-fold, or 5-fold CV, but as the sample size increases, the MSEs for 2-fold CV become equivalent to the other CV methods. Among the modeling methods, in most cases the MSE is largest for Lasso due to the larger bias for Lasso.

Null dataset: As there is no significant bias with the null dataset, the MSE is dominated by the variance. Hence, due to their high variance, the split-sample and the Loo CV have the largest MSE. As the variance of the k-fold CV decreases as k decreases from 10 to 2, the MSE also decreases as k decreases.

Lung cancer and DLBCL datasets: For small sample sizes, the split-sample and the Loo CV have higher MSE than the k-fold CV. As the sample size increases, the MSEs for the various resampling methods become equivalent.

4. Discussion

In a comparative study of resampling methods for prediction error estimation in the case of binary response classification problems, Molinaro *et al*. [12] concluded that for small samples the split-sample and 2-fold CV methods perform poorly and that Loo CV performed well with regard to bias as well as MSE for a wide range of data settings. Although Loo CV has previously been criticized for its high variability [32, 33], its low bias compared with the other methods dominated the MSE comparisons in many of the comparisons made by Molinaro *et al*. [12]. The case of a null dataset was also not

Published in 2010 by John Wiley & Sons, Ltd.
Statist. Med. ***2011***, 30 642–653

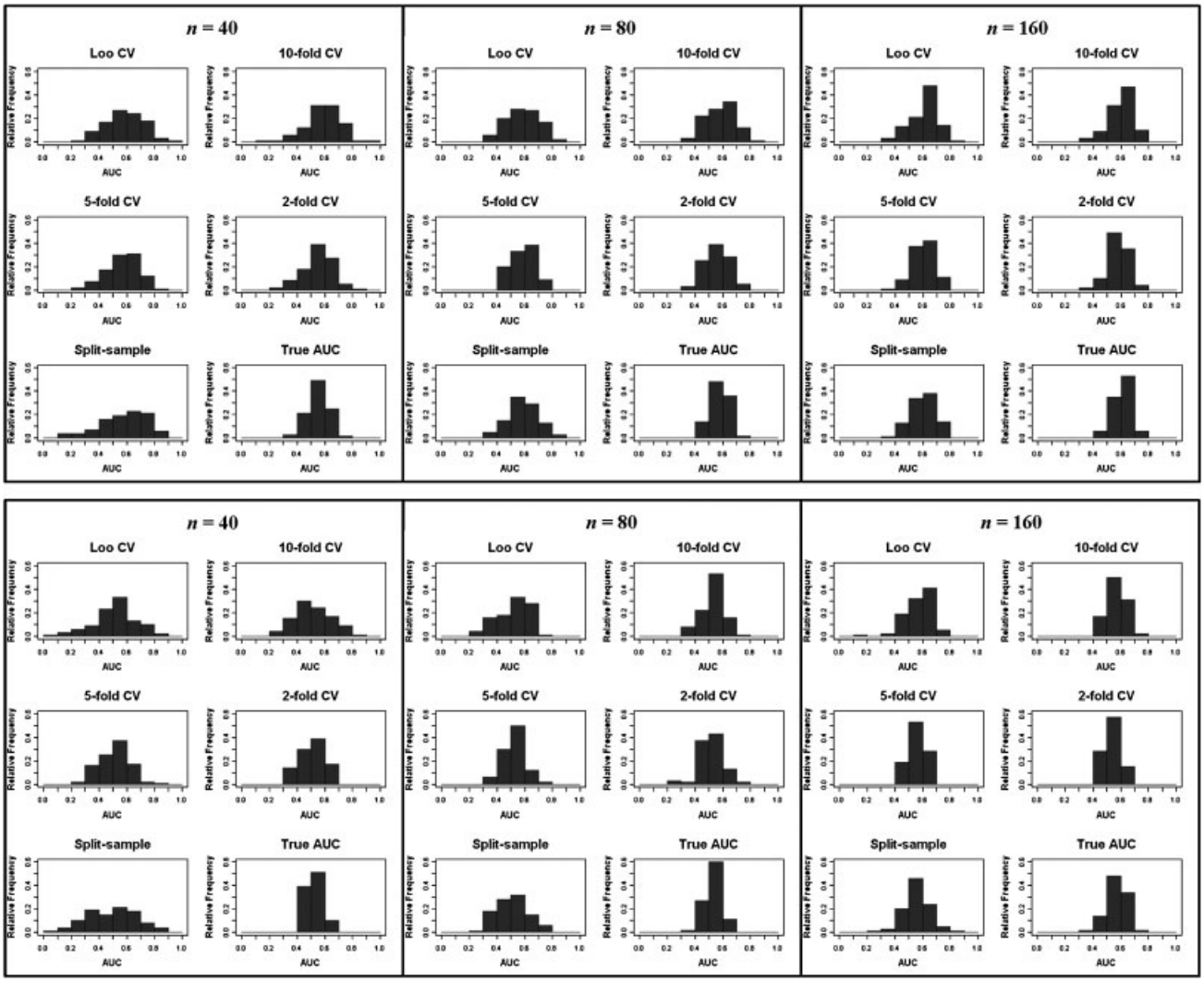

Figure 2. Distribution of the estimated resampling and true AUC(t) for SuperPCR: (top) lung cancer data and (bottom) DLBCL data.

investigated by Molinaro *et al*. In a null setting, none of the methods are biased and the larger variance of Loo CV would have dominated the MSE. All the previous studies cited deal only with the binary response classification problem with the misclassification rate being the error measure. The misclassification rate used for evaluating classification models is an error counting estimate and this is considered to add to the variability of resampling, especially for smaller samples [32]. It is thus necessary to investigate whether the results obtained for high-dimensional classification problems can be directly extended to survival prediction where alternate accuracy measures are used and where there is further loss of information due to censoring.

In this paper, we have compared the performance of several resampling methods for survival risk prediction. The evaluation has been carried out for high-dimensional data for a wide range of sample sizes, signal content, and dimension reduction methods. Although sample sizes used in microarray studies are increasing, there are still many studies that are based on small samples. In a review of 90 microarray studies with cancer outcome data, 65 per cent of the publications had less than 50 patients and only 6 per cent of the publications had more than 100 patients [34]. The general conclusions from our study are summarized below:

1. For null datasets where survival outcome is independent of gene expression, none of the resampling methods investigated are associated with any systematic bias and hence all the resampling methods investigated, including the Loo CV are valid. Thus our simulation studies dispel the widespread (mis)-perception that CV, and especially, the Loo CV can be optimistically biased. Here we emphasize again that for the resampling methods to be valid, feature selection/dimension reduction must be conducted based only on the training data and must be completely repeated within each resampling loop. Failure to do so will result in highly optimistically biased estimates.

Published in 2010 by John Wiley & Sons, Ltd.
Statist. Med. ***2011***, 30 642–653

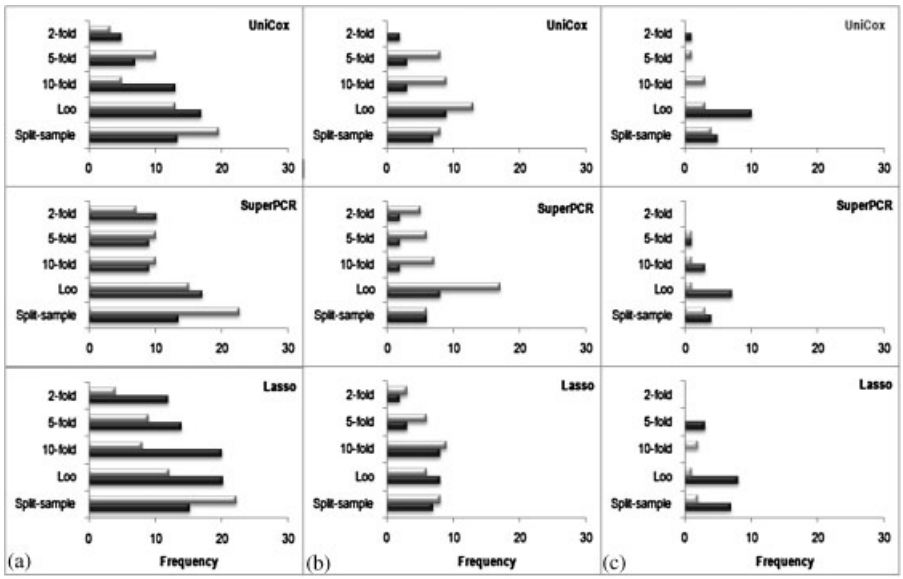

Figure 3. In the case of the null dataset, the AUC(t) is expected to be around 0.5. The variation in the resampling AUC(t) estimates can, however, lead to too pessimistic or too optimistic estimates. The number of times $\hat{A}^{RS}(S_n)$ is too pessimistic (below 0.3, black bars) or too optimistic (above 0.7, light gray bars) in a total of 100 replications for the null dataset is shown. In comparison to this, the number of times $\hat{A}(S_n)$ is below 0.3 or above 0.7 is nil: (a) $n=40$; (b) $n=80$; and (c) $n=160$.

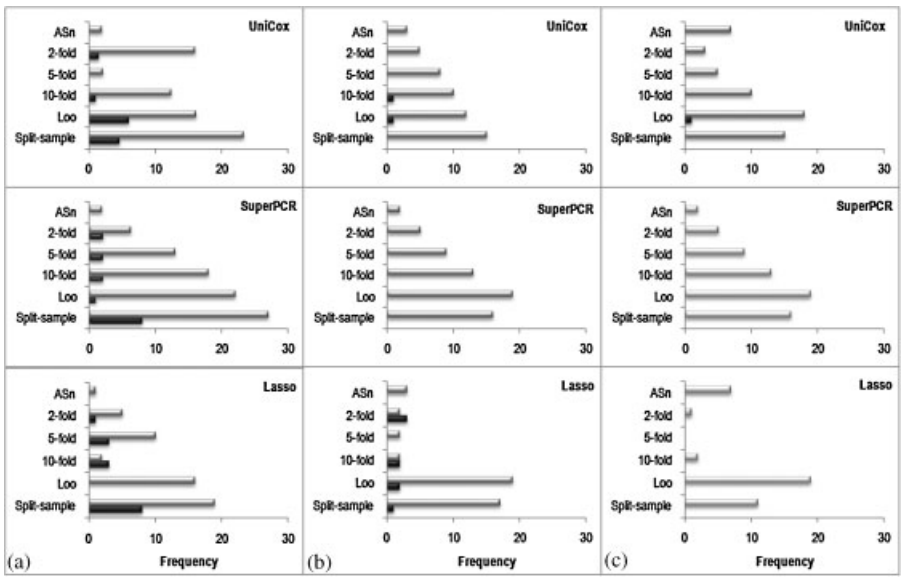

Figure 4. The variation in the resampling AUC(t) estimates more frequently leads to too pessimistic or too optimistic estimates. The number of times $\hat{A}^{RS}(S_n)$ is too pessimistic (below 0.3, black bars) or too optimistic (above 0.7, light gray bars) in a total of 100 replications for the lung cancer dataset is compared with the number of times $\hat{A}(S_n)$ is below 0.3 or above 0.7: (a) $n=40$; (b) $n=80$; and (c) $n=160$.

2. Although we ruled out the presence of optimistic bias for any of the resampling schemes, the variance due to resampling can make the point estimates of accuracy unreliable, especially for the split-sample and the Loo CV and when the sample size is also small. Hence, in addition to point estimates of prediction accuracy it becomes important to also report confidence intervals. As suggested by Heagerty *et al.* [19], confidence bands for AUC(t) can be obtained by bootstrapping.
3. Some of the largest MSEs are seen for the split-sample method with small to moderate samples. This is because split-sample suffers both from a high bias as well as a high variance. Whereas the high bias is due to the reduced dataset available for training, the high variance is because of the small size of the test dataset. Hence split-sample should not be a preferred method for the estimation of prediction accuracy.

Published in 2010 by John Wiley & Sons, Ltd.
Statist. Med. ***2011***, 30 642–653

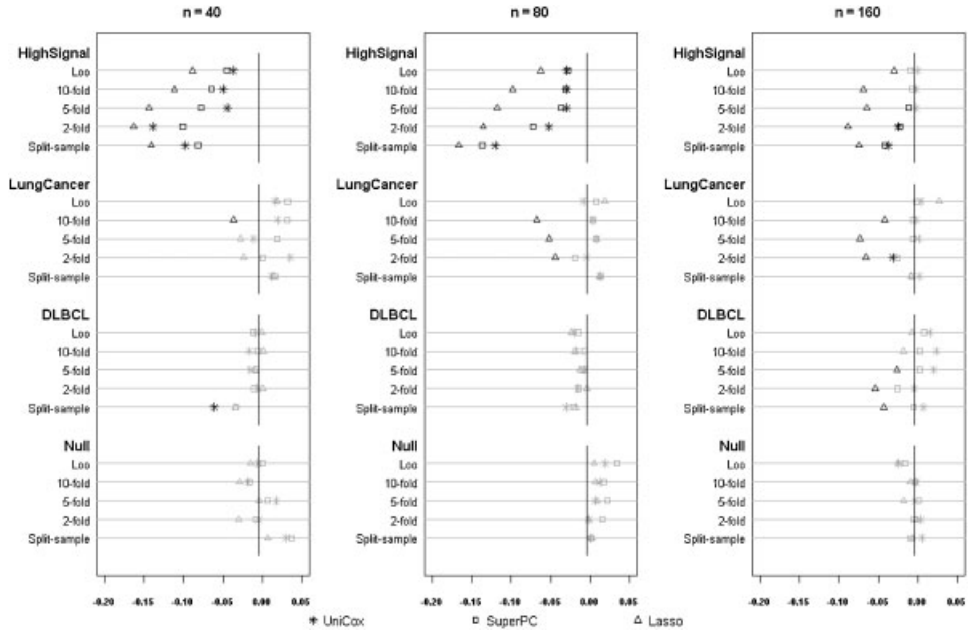

Figure 5. Bias in resampling AUC(t) estimates. Black symbols indicate statistically significant bias ($p<0.01$ from a permutation t-test) and light gray symbols indicate bias not statistically significant.

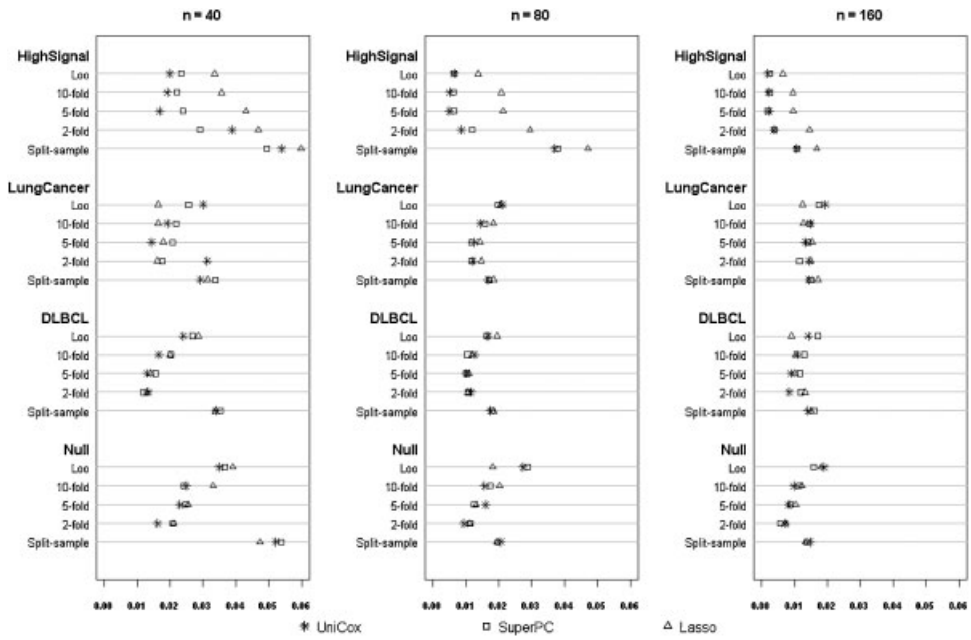

Figure 6. Mean square error in the resampling AUC(t) estimates.

4. A k-fold CV tends to give conservative estimates of accuracy (is negatively biased) especially with moderate to strong signals. With strong signals, this negative bias increases as k decreases from 10 to 2. With moderate to strong signals, the reduction in the number of training samples affects the achievable classification accuracy. This bias would dissipate as the sample size increases and the achievable classification accuracy plateaus. Although Loo CV may have a slight advantage in terms of the bias for moderate to strong signals, this advantage dissipates as signal strength decreases, and in the presence of weak signals coupled with small samples, the large variance of Loo CV leads to a higher MSE.

Published in 2010 by John Wiley & Sons, Ltd.
Statist. Med. ***2011***, 30 642–653

5. For moderate to large samples and moderate to high signal strength, the Loo and k-fold CV have equivalent MSEs. An advantage of the 5-fold and 10-fold CV over Loo CV is computational speed. Also, the potential negative bias of k-fold CV may be preferable to the more frequent optimistic point estimates resulting from the large variance of Loo CV.
6. Lasso-based PH regression provided the largest AUC(t) estimates in the presence of high-signal implying that it makes effective use of the data. However, this property also renders Lasso more sensitive to removal of data due to resampling and hence conservatively biased for small training sets with strong signal datasets. In an earlier study on an evaluation of seven different survival prediction strategies for high-dimensional gene-expression data, penalized methods were found to be among the best performers [35]. In the earlier study, the penalty parameter was optimized through CV, but model assessment was done using only the split-sample technique. It would thus be worthwhile to perform a more in-depth study of penalized methods for survival prediction in high-dimensional settings.

In the case of binary response classification, Loo CV was found to perform poorly when an unstable classifier like CART was used in the presence of a weak signal [12]. This is because the large variance of Loo CV is accentuated when used with an unstable classifier. Although the stability of all the modeling methods investigated in our paper was more or less equivalent, we believe that as in the case of classification, for risk prediction too, the variance of the resampling estimates of accuracy would increase in the presence of an unstable modeling method. Moreover, resampling techniques that are inherently highly variable, such as the split-sample or the Loo CV, are likely to be affected more, especially if the sample size is also small.

In gene expression based prognostic prediction, accuracy estimates using the split-sample and Loo CV methods are frequently used and reported in publications. Based on the results of our analysis we discourage their use with survival data, and especially with small training sets. Through our simulations, we have tried to provide information about the key determinants that should influence choice of a resampling method for evaluating predictive accuracy of survival risk models in high-dimension settings. For datasets with strong signals, bias predominates and a larger k (in case of a k-fold CV) is favored. For datasets with weaker signals, variance predominates and a smaller k may be slightly favored, especially if the sample size is also small. It would be hard to find a universal value of k that is suitable under all sample sizes, datasets, and modeling strategies. However, our study does show that the 5- or 10-fold CV (and likely any k in between) perform reasonably well for datasets of different sample sizes and different signal strengths. It is also not advisable to fine tune the value of k for each dataset under study, as this would again lead to a biased reporting of results. Hence we recommend that for an initial assessment of prognostic factor studies involving time-to-event responses, the k-fold CV with $k=5$ or 10 should be more widely adopted.

More importantly, the point estimates of accuracy obtained through resampling should be supplemented by appropriately computed confidence intervals for a proper interpretation of the results. In the case of high-dimensional classification problems, Jiang *et al.* [36] reported on the inadequacies of many existing confidence interval methods and proposed several confidence interval methods using a bootstrap case cross-validation procedure (BCCV) with and without bias correction. However, there is limited experience on the application of these methods for the estimation of the time-dependent AUC(t) with survival data. We would continue to focus on this important topic in our future investigations.

References

1. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Giltnane JM, Hurt EM, Zhao H, Averett L, Yang L, Wilson WH, Jaffe ES, Simon R, Klausner RD, Powell J, Duffey PL, Longo DL, Greiner TC, Weisenburger DD, Sanger WG, Dave BJ, Lynch JC, Vose J, Armitage JO, Montserrat E, López-Guillermo A, Grogan TM, Miller TP, LeBlanc M, Ott G, Kvaloy S, Delabie J, Holte H, Krajci P, Stokke T, Staudt LM; 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**:1937–1947.
2. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, 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**:530–536.
3. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, Hiller W, Fisher ER, Wickerham DL, Bryant J, Wolmark N. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *The New England Journal of Medicine* 2004; **351**:2817–2826.
4. Vermeulen J, De Preter K, Naranjo A, Vercruysse L, Van Roy N, Hellemans J, Swerts K, Bravo S, Scaruffi P, Tonini GP, De Bernardi B, Noguera R, Piqueras M, Cañete A, Castel V, Janoueix-Lerosey I, Delattre O, Schleiermacher G, Michon J, Combaret V, Fischer M, Oberthuer A, Ambros PF, Beiske K, Bénard J, Marques B, Rubie H, Kohler J, Pötschger U, Ladenstein R, Hogarty MD, McGrady P, London WB, Laureys G, Speleman F, Vandesompele J. Predicting outcomes for children with neuroblastoma using a multigene-expression signature: a retrospective SIOPEN/COG/GPOH study. *The Lancet Oncology* 2009; **10**:663–671.
5. 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**:14–18.

Published in 2010 by John Wiley & Sons, Ltd.

6. Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, Wang Y. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer* 2008; **8**:37–49.
7. Cox DR. Regression models and life tables. *Journal of the Royal Statistical Society*, *Series B* 1972; **34**:187–220.
8. Bair E, Hastie T, Paul D, Tibshirani R. Prediction by supervised principal components. *Journal of the American Statistical Association* 2006; **101**:119–137.
9. Verweij PJM, Van Houwelingen HC. Penalized likelihood in Cox regression. *Statistics in Medicine* 1994; **13**:2427–2436.
10. Tibshirani R. The lasso method for variable selection in the Cox model. *Statistics in Medicine* 1997; **16**:385–395.
11. Gui J, Li H. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* 2005; **21**:3001–3008.
12. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* 2005; **21**:3301–3307.
13. Potti A, Mukherjee S, Petersen R, Dressman HK, Bild A, Koontz J, Kratzke R, Watson MA, Kelley M, Ginsburg GS, West M, Harpole Jr DH, Nevins JR. A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. *The New England Journal of Medicine* 2006; **355**:570–580.
14. Roepman P, Jassem J, Smit EF, Muley T, Niklinski J, van de Velde T, Witteveen AT, Rzyman W, Floore A, Burgers S, Giaccone G, Meister M, Dienemann H, Skrzypski M, Kozlowski M, Mooi WJ, van Zandwijk N. An immune response enriched 72-gene prognostic profile for early-stage non-small-cell lung cancer. *Clinical Cancer Research* 2009; **15**:284–290.
15. Breslow NE. Contribution to the discussion on the paper by D.R. Cox. *Journal of the Royal Statistical Society*, *Series B* 1972; **34**:216–217.
16. Hosmer Jr DW, Lemeshow S. *Applied Survival Analysis.* Wiley: New York, 1999; 93–105.
17. 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. *American Journal of Epidemiology* 2004; **159**:882–890.
18. Lusa L, McShane LM, Radmacher MD, Shih JH, Wright GW, Simon R. Appropriateness of some resampling-based inference procedures for assessing performance of prognostic classifiers derived from microarray data. *Statistics in Medicine* 2007; **26**:1102–1113.
19. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 2000; **56**:337–344.
20. Akritas MG. Nearest neighbor estimation of a bivariate distribution under random censoring. *The Annals of Statistics* 1994; **22**:1299–1327.
21. Segal MR. Microarray gene expression data with linked survival phenotypes: diffuse large-B-cell lymphoma revisited. *Biostatistics* 2006; **7**:268–285.
22. Buyse M, Loi S, van't Veer L, Viale G, Delorenzi M, Glas AM, d'Assignies MS, Bergh J, Lidereau R, Ellis P, Harris A, Bogaerts J, Therasse P, Floore A, Amakrane M, Piette F, Rutgers E, Sotiriou C, Cardoso F, Piccart MJ; TRANSBIG Consortium. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *Journal of the National Cancer Institute* 2006; **6**:1183–1192.
23. Parker BJ, Günter S, Bedo J. Stratification bias in low signal microarray studies. *BMC Bioinformatics* 2007; **8**:326–341.
24. Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma, Shedden K, Taylor JM, Enkemann SA, Tsao MS, Yeatman TJ, Gerald WL, Eschrich S, Jurisica I, Giordano TJ, Misek DE, Chang AC, Zhu CQ, Strumpf D, Hanash S, Shepherd FA, Ding K, Seymour L, Naoki K, Pennell N, Weir B, Verhaak R, Ladd-Acosta C, Golub T, Gruidl M, Sharma A, Szoke J, Zakowski M, Rusch V, Kris M, Viale A, Motoi N, Travis W, Conley B, Seshan VE, Meyerson M, Kuick R, Dobbin KK, Lively T, Jacobson JW, Beer DG. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature Medicine* 2008; **14**:822–827.
25. Simon R, Lam A, Li MC, Ngan M, Menenzes S, Zhao Y. Analysis of gene expression data using BRB-ArrayTools. *Cancer Informatics* 2007; **2**:11–17.
26. Zhao Y, Simon R. BRB Arraytools data archive for human cancer gene expression: a unique and efficient data sharing resource. *Cancer Informatics* 2008; **6**:9–15.
27. Heagerty PJ, Saha P. survivalROC: Time-dependent ROC curve estimation from censored survival data. R package v. 1.0.0, 2006.
28. Therneau T, Lumley T. survival; Survival analysis, including penalised likelihood. R package v. 2.34-1, 2008.
29. Goeman JJ. L(1) penalized estimation in the Cox proportional hazards model. *Biometrical Journal* 2010; **52**:70–84.
30. R Development Core Team. R: a language and environment for statistical computing. *R Foundation for Statistical Computing*, 2008.
31. Tierney L, Rossini AJ. Sevcikova H. snow: simple network of workstations. R package v. 0.3-3, 2004.
32. Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification. *Bioinformatics* 2004; **20**:374–380.
33. Xiao Y, Hua J, Dougherty ER. Quantification of the impact of feature selection on the variance of cross-validation error estimation. *EURASIP Journal on Bioinformatics and Systems Biology* 2007; 16354–16364.
34. Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *Journal of the National Cancer Institute* 2007; **99**:147–157.
35. Bøvelstad HM, Nygård S, Størvold HL, Aldrin M, Borgan Ø, Frigessi A, Lingjaerde OC. Predicting survival from microarray data—a comparative study. *Bioinformatics* 2007; **23**:2080–2087.
36. Jiang W, Varma S, Simon R. Calculating confidence intervals for prediction error in microarray classification using resampling. *Statistical Applications in Genetics and Molecular Biology* 2008; **7**:Article No. 8.