Contemporary Clinical Trials 36 (2013) 636–641

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Overfitting in prediction models – Is it a problem only in high dimensions?

Jyothi Subramanian [a], Richard Simon [b,*]

[a] Emmes Corporation, USA
[b] Biometric Research Branch, National Cancer Institute, USA

ARTICLE INFO

Article history:
Received 25 February 2013
Revised 29 May 2013
Accepted 17 June 2013
Available online 27 June 2013

Keywords:
Classifiers
Prediction accuracy
Overfitting
Clinical trials
Patient selection

ABSTRACT

The growing recognition that human diseases are molecularly heterogeneous has stimulated interest in the development of prognostic and predictive classifiers for patient selection and stratification. In the process of classifier development, it has been repeatedly emphasized that in situations where the number of candidate predictor variables is much larger than the number of observations, the apparent (training set, resubstitution) accuracy of the classifiers can be highly optimistically biased and hence, classification accuracy should be reported based on evaluation of the classifier on a separate test set or using complete cross-validation. Such evaluation methods have however not been the norm in the case of low-dimensional, $p < n$ data that arise, for example, in clinical trials when a classifier is developed on a combination of clinico-pathological variables and a small number of genetic biomarkers selected from an understanding of the biology of the disease. We undertook simulation studies to investigate the existence and extent of the problem of overfitting with low-dimensional data. The results indicate that overfitting can be a serious problem even for low-dimensional data, especially if the relationship of outcome to the set of predictor variables is not strong. We hence encourage the adoption of either a separate test set or complete cross-validation to evaluate classifier accuracy, even when the number of candidate predictor variables is substantially smaller than the number of cases.

Published by Elsevier Inc.

1. Introduction

In many disease areas, and especially in oncology, recognition of the molecular heterogeneity of the disease has fueled the search for prognostic and predictive classifiers that identify patients who require new treatment regimens and who are likely to benefit from specific new regimens. Such classifiers can be used for selection and stratification of patients in clinical trials and for structuring the analysis plan of clinical trials. Advances in genomic technologies has moreover made it possible to measure gene expression levels for tens of thousands of genes, and these have been used in combination with traditional clinico-pathological variables to develop composite pharmacogenomic classifiers that could potentially be useful in the design and analysis of clinical trials [1]. The number of cases available for classifier development, however, remains much less, usually of the order of hundreds or less. This is commonly referred to as the high-dimensional, low sample size (HDLSS) (i.e., $p \gg n$) setting.

Overfitting, which is characterized by high accuracy for a classifier when evaluated on the training set but low accuracy when evaluated on a separate test set, has been recognized as a problem in $p \gg n$ settings [2]. In HDLSS settings, it has been repeatedly emphasized that the apparent (training set, resubstitution) accuracy of a classifier is highly optimistically biased and hence should never be reported and that accuracy should be estimated based on the evaluation of the classifier on separate test sets or through complete resampling in which the model is redeveloped for each resampling [2,3]. The use of resampling techniques or independent test sets for

* Corresponding author. Tel.: +1 2402766028.
E-mail address: rsimon@mail.nih.gov (R. Simon).

1551-7144/$ – see front matter. Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.cct.2013.06.011

the evaluation of prediction accuracy are however not widespread in the traditional $p < n$ situations, even though overfitting is likely to be a problem in these settings also [4,5]. In the context of clinical trials, prediction problems with $p < n$ can arise, for example, when a classifier is developed on a combination of clinico-pathological and a small number of candidate genetic biomarker variables selected based on an understanding of the biology of the disease. When p is less than n, there exist rules of thumb, for example, specifying that the *effective*[1] sample size for training should be at least 10 times the number of candidate predictors [6,7]. However, these rules of thumb appear to have been developed for ensuring stability of regression coefficients [8,9] and it is not clear whether adoption of these rules also avoid overfitting.

We conducted simulation studies to investigate the existence and extent of the problem of overfitting under traditional low-dimensional settings. As p increases and starts exceeding n, traditional classification techniques like logistic regression or Fisher's linear discriminant analysis cannot be directly applied and some form of variable selection and/or shrinkage estimation becomes mandatory [4,5,10]. Shrinkage based approaches in fact are reported to be preferable in comparison to p-value based variable selection methods [5]. In our simulations, we study overfitting as a function of the ratio of p to the effective sample size, with and without feature selection. The results of these simulations and the significance of the results are reported in this paper.

2. Material and methods

2.1. Binary class prediction

In the binary class prediction problem we have a training set $\{X_i, Y_i\}$ of n observations where $Y_i \in \{0, 1\}$ is the outcome class label and $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is a p-dimensional vector of predictor variables (features). The goal is to build a rule utilizing the information in X in order to predict Y. The rule is often known as a *classifier*. By developing the classifier on the training set of data, future unobserved outcomes can be predicted based on their corresponding measured predictor variables. Many methods exist for developing classifiers, including linear and quadratic discriminant analysis, logistic regression, decision trees, support vector machines, and others [11]. Additionally, variable selection may also be used in order to reduce the number of predictors in the classifier.

2.2. Simulations

For all our simulations, the number of candidate predictors, p was fixed at 10. Of the 10 predictors, 5 predictors were informative and the remaining 5 were non-informative. The number of samples in the training set, n, was varied from 20 to 1000. Half of the samples (i.e. $n/2$) were randomly assigned to class 0 ($Y = 0$) and the other half to class 1 ($Y = 1$). The effective sample size in our simulations was thus $n/2$. The informative predictors were generated from $\mathrm{N}(0, \mathrm{I}_5)$ for class 0 and $\mathrm{N}(\mu, \mathrm{I}_5)$ for class 1. The non-informative predictors were generated from $\mathrm{N}(0, \mathrm{I}_5)$ for both class 0 and class 1. Separate simulations were carried out for the values of μ in 0, 0.25 and 0.5 to represent the null signal and signals of increasing strength from moderate to high. Additionally, a simulation was conducted with two informative predictors with $\mu = 0.25$ and three informative predictors with $\mu = 0.5$.

To study the sensitivity of results to correlation among predictors, additional simulations were carried out with block diagonal correlation structures, where the informative and non-informative predictors were assumed to be correlated with pairwise common correlation coefficient r. Values of $r = 0.25$ and 0.75 were used.

Diagonal linear discriminant analysis (DLDA) was used as the classification method [10]. DLDA corresponds to Fisher's linear discriminant analysis where the class specific densities are assumed to have the same diagonal covariance matrix. In DLDA, a new sample with feature vector $X^* = (x^*_1, x^*_2, \ldots, x^*_p)$ is assigned to class 0 if

$$\sum_{j=1}^{p}\frac{\left(x_j^*-\overline{x}_j^{(0)}\right)^2}{s_j^2}\le\sum_{j=1}^{p}\frac{\left(x_j^*-\overline{x}_j^{(1)}\right)^2}{s_j^2}$$

and otherwise assigned to class 1. DLDA has the advantage that for p predictors, only p variances need to be estimated. In contrast to this, Fisher's linear discriminant analysis requires the estimation of $p(p + 1)/2$ elements of the covariance matrix. DLDA is commonly used in $p > n$ settings as it is more robust to overfitting compared to Fisher's LDA and often results in greater predictive accuracy even when the features are correlated [11]. For $p > n$ problems, ordinary logistic regression too cannot be used because the design matrix is singular. Stepwise logistic regression tends to provide substantially overfit models in that setting and so penalized version of logistic regression are often used to shrink the regression coefficients.

DLDA was used in our simulations because of its stability in $p < n$ problems and its resistance to overfitting compared to Fisher's LDA and stepwise logistic regression in $p > n$ problems. When the class specific covariance matrices are equal and diagonal, DLDA is equivalent to logistic regression.

Since, typically, some form of variable selection is incorporated even in the low-dimensional case; simulations were conducted with and without variable selection to study the impact of variable selection on overfitting. The variable selection methods studied were:

(i) selecting variables with the largest k absolute value univariate t-test statistics with $k = 3$ or 5.
(ii) using cross-validation to select the optimal number of variables in the model.

[1] For the linear regression problem, the effective sample size is the actual sample size. In the case of proportional hazards regression, the effective sample size is the number of events, and in case of binary class prediction, the effective sample size is the number of observations in the smaller of the two classes [4].

(iii) using a nearest shrunken centroid classifier [12] which applies soft thresholding shrinkage.

Variable selection using a p-value cutoff is often used to limit the number of predictors in the model. However, using a fixed number of predictors in the model enables us to more clearly evaluate how overfitting depends on the number of predictors in the model than would be possible using a p-value cutoff. In the case of (ii), the number of variables giving the best cross-validated prediction accuracy was identified by computing the cross-validated prediction accuracy after including variables with the largest k absolute value univariate t-test statistics for all values of k from 1 to 10 and selecting the number of variables to be the value of k giving the maximum cross-validated prediction accuracy. This approach is similar to optimizing the significance level α for variable selection in a cross-validation loop [10], but again instead of optimizing α, we optimize the number of variables in the model directly. The nearest centroid classification method can be thought of as an extension of nearest centroid classification based on the Euclidean distance metric but with shrunken class centroids used in the place of actual class centroids. The shrinkage applied plays the role of reducing the number of variables in the classifier. The amount of shrinkage was chosen through cross-validation [12].

The ratio of the effective sample size to the number of candidate predictor variables was denoted by ρ and the extent of overfitting was evaluated for a large range of values of ρ. The classifier was developed from the training data either using all the variables or after pre-selection of variables. In every case, the true prediction accuracy of the classifier was evaluated on a large independent test set of 1000 observations (500 observations each in class 0 and class 1) that followed the same distribution as the training set. Generating such a large test is possible when the model generating the data is known and a large test set can in turn provide a precise and reliable estimate of the true prediction accuracy.

2.3. Evaluation of prediction accuracy

Prediction models should be evaluated based on their ability to predict the outcome class accurately for new observations having similar distribution as the training data. Prediction accuracy, not goodness of fit or statistical significance of model fit, is the objective. Overall prediction accuracy is defined as the percentage (or proportion) of samples correctly classified and can be decomposed into sensitivity and specificity. Though sensitivity and specificity are also important evaluation measures of a classification model, we focus on overall prediction accuracy to illustrate our points. Prediction accuracy, computed on a large completely independent test dataset gives an unbiased measure of the performance of the classifier. The difference between the apparent (resubstitution) prediction accuracy and prediction accuracy as measured on an independent test set is a measure of the *degree of overfit* to the training sample.

The simulation was conducted by repeating each combination of conditions 100 times. The *degree of overfit* was measured as the average difference in the prediction accuracy of the classifier between the training set (apparent accuracy) and the test set over the 100 replications.

Simulations were conducted using code written in R (version 2.15.1) [13]. The dDA function from the sfsmisc library was used for DLDA [14]. The pamr.train and pamr.cv functions from the pamr library were used for nearest shrunken centroid method [15].

3. Results

3.1. Simulation results: the null case

The results for the simulations under the null for all uncorrelated predictors are illustrated in Fig. 1. It can be seen that in this setting, the true average estimated prediction accuracy (over the simulation replications) of the prediction model as measured on the test set is 50%. This is as it should be, since in the null situation no variables are informative for the class label, Y, and as per our design, the prevalence of the $Y = 1$ observations is 50%. When prediction accuracy is measured on the training sets, however, the apparent accuracy is greater than 50% in all cases. When the ratio of number of cases to number of candidate variables ρ was 10, the apparent error rate was over-estimated (on an average) by up to 8 percentage points. The average degree of overfit is 6% when ρ increases to 20. The situation is similar irrespective of whether feature selection was or was not part of the modeling process. Even for values of ρ as high as 50, the average apparent accuracy does not overlap with the true accuracy, though the difference becomes small (an average of around 4% overfit). It is thus

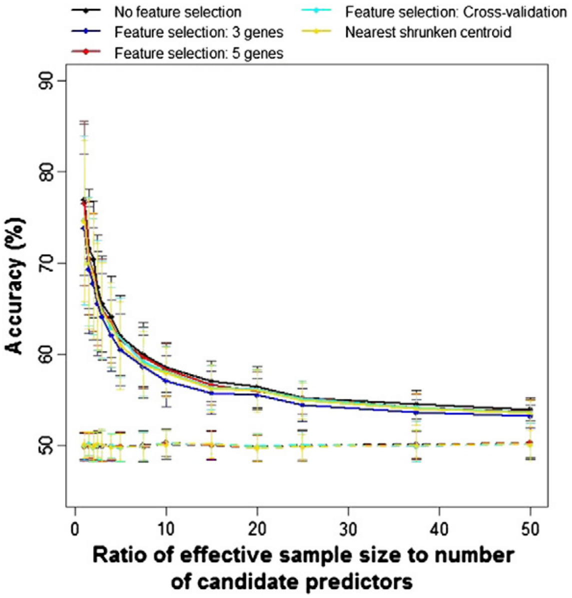

Fig. 1. Illustration of overfit in prediction models under the null. The dashed line represents the accuracy of classification as evaluated in the test set and the solid lines represent the accuracy of classification in the training set. The error bars represent ±1 standard deviation.

clear that, under the null, overfitting is a problem even for low-dimensional data.

3.2. Simulation results: the alternative case

For independent predictors with moderate signals, the average degree of overfit is around 5% when ρ equals 10, but decreases considerably for values of $\rho > 20$ or more (Fig. 2a). Overfitting does not seem to be a serious problem in those $p < n$ situations with strong signal and $\rho \geq 10$. With an effective sample size of 100 for 10 candidate predictors, the degree of overfit is around 2%, and continues to decrease slowly with further increase in ρ (Figs. 2b and A2 in Appendix). In the case where we have a mix of predictors, some with moderate and some with strong signals, the degree of overfit for ρ of 10 is around 3 – 4% which is in between the degree of overfit for the all moderate and the all strong signal predictor situations (Fig. 2(c)). The degree of overfit seems to depend on the extent of signal strength in the predictors and decreases with increase in predictor signal. Again, as in the case of the null, the degree of overfit does not differ whether feature selection was or was not part of the modeling process. The results also do not seem to show a dependency on the actual feature selection method employed.

Simulations with correlated predictors also showed that the degree of overfit has no dependence on the actual feature selection method employed. Hence, results for the case of correlated predictors are presented only for the feature selection with nearest shrunken centroid (Figs. A1 and A2, Appendix). It can be observed that the degree of overfit decreases, but not substantially, when the pairwise correlation between predictors is high.

4. Discussion

Statisticians have long recognized overfitting as an important problem in classifier development [4,5]. In particular, the problem of overfitting has been recognized as a major concern in HDLSS settings [2]. Through simulations, we have demonstrated overfitting to be a problem not just in the HDLSS case, but also in the more traditional settings where the number of candidate variables is much less than the number of observations. We also studied the dependencies between the degree of overfit to ρ (ratio of sample size to the number of candidate predictor variables), predictor signal strength and feature selection. We simulated data under a variety of conditions but all simulations have limitations and cannot be fully comprehensive. For the purposes of this paper, however, it was necessary to know the true prediction accuracy in order to evaluate degree of overfitting of the models examined. Simulations enable the generation of a large test set following the same distribution as the training data to evaluate the expected true prediction accuracy.

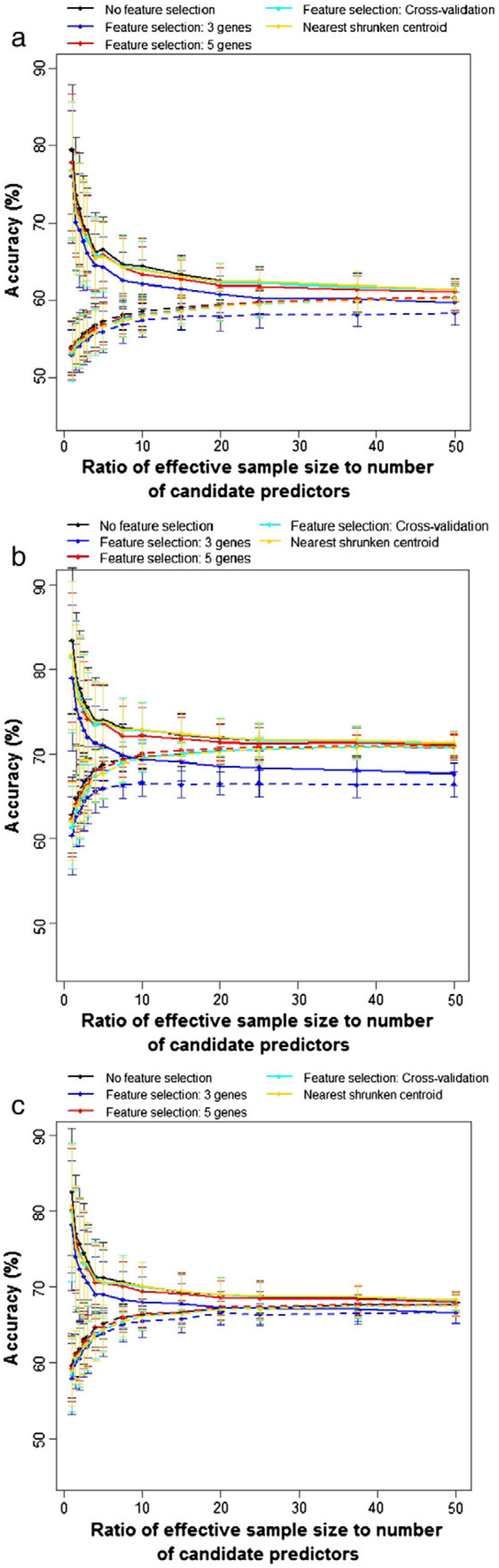

Fig. 2. Illustration of overfit in prediction models under **(a)** moderate signal **(b)** high signal and **(c)** combination of moderate and high signal predictors. The dashed line represents the accuracy of classification as evaluated in the test set and the solid lines represent the accuracy of classification in the training set. The error bars represent ±1 standard deviation.

Though evaluation of our findings on real data sets would be ideal, a reliable estimate of the expected true prediction accuracy would be difficult to obtain because of the absence of a large test set. We have tried to use the results from simulations to understand some of the important features that affect overfitting rather than just cataloging the simulation results.

In the simulations, feature selection did not have a strong influence the degree of overfit. This is to some extent a result of our selection of classifiers like DLDA and shrunken centroids that are more resistant to overfitting than those which explicitly or implicitly utilize correlation between feature information [11]. The relevant p in assessing whether overfitting is likely to be a problem is the number of candidate variables, not the number of variables in the model after variable selection. In fact, variable selection can result in increasing model overfitting by overestimating the magnitude of the selected variables unless it is accompanied by shrinkage of model coefficients as it is with the shrunken centroid method.

We did notice a dependency between the degree of overfit and the signal strength in the predictors – the degree of overfit increased with decreasing predictor signal strength. The common rules of thumb suggested for ensuring the stability of regression models do not necessarily avoid overfitting, at least not under null and weak signals and the degree of overfit becomes negligible only under strong signal predictors. In the presence of high correlation among the predictors, the degree of overfit decreases, but not substantially. This decrease is probably because the *effective* number of candidate predictors decreases when they are correlated.

The only case when overfitting did not seem to be a problem in low-dimensional situations was when the signal in the predictors was strong and the ratio of the effective sample size to the number of candidate predictors was also greater than 10. Thus the rules of thumb formulated for ensuring stability of regression coefficients seem to also have a role in decreasing the degree of overfit, but only under strong signals. Hence, reporting the apparent prediction accuracy of a classifier should be avoided except in such ideal settings.

In all our simulations, the event rates were fixed at 50%. This was for the sake of simplicity. A higher or lower event rate would decrease the effective sample size and hence, for the same number of candidate predictors, larger samples would be required to arrive at the same value of ρ. We do not however expect the differences in event rates to significantly alter the relationship between ρ and the degree of overfit or to change the principal conclusions of this study.

Our evaluations were based on using DLDA and shrunken centroid classifiers. These are methods that have been found in $p > n$ studies to be relatively resistant to overfitting. Stepwise logistic regression is often applied in $p < n$ classification problems. However, because stepwise logistic regression is known to be much more prone to overfitting than DLDA, we do not believe that our principal conclusion that overfitting is often a problem in $p < n$ studies would be altered for studies based on stepwise logistic regression.

5. Conclusion

Use of the apparent accuracy (i.e. training set estimated accuracy) for models developed on low-dimensional data is common in medical journals and is even more common when attempting to develop predictive classifiers where treatment by covariate interactions as well as main effects are candidates for the model. We however encourage authors to report prediction accuracy based on complete cross-validation or evaluation on an independent test set to avoid over-optimism, for low-dimensional problems except in cases where the effective sample size is at least ten times the number of candidate predictors and the proportion of variability explained by the model is substantial. For prospective development of a predictive classifier to be used in a phase III trial for selecting or stratifying patients, one often has a number of candidate variables, but the number of cases available is often severely restrained by the size of the phase II database. In some cases the classifier development and classifier use for stratifying the population for analysis are being combined into the phase III clinical trial. In these situations, use of complete cross-validation is often essential. The power of such an integrated cross-validation based approach has been recently demonstrated to illustrate the new paradigm of predictive analysis of clinical trials for patient stratification using an example from a prostate cancer trial [1].

Appendix A

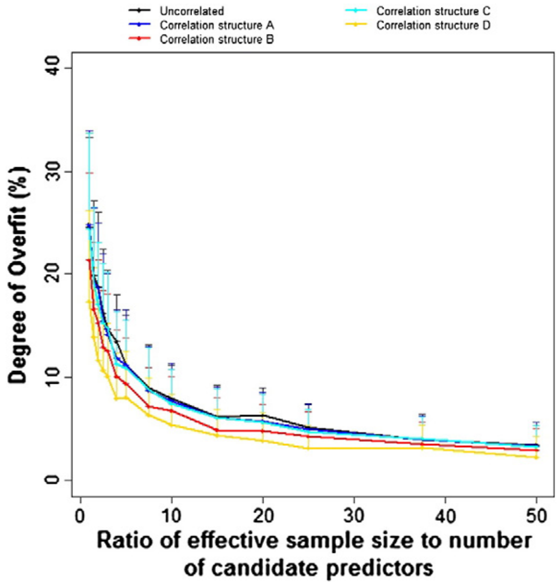

Fig. A1. Effect of different correlation structures among predictors to the degree of overfit for the null data. Correlation structures (A): all informative predictors pairwise correlated with $r = 0.25$, non-informative predictors uncorrelated (B) same as (A), but with r = 0.75 (C) informative and non-informative predictors pairwise correlated with r = 0.25 (D) same as (C), but with $r = 0.75$. The bars represent +1 standard deviation.

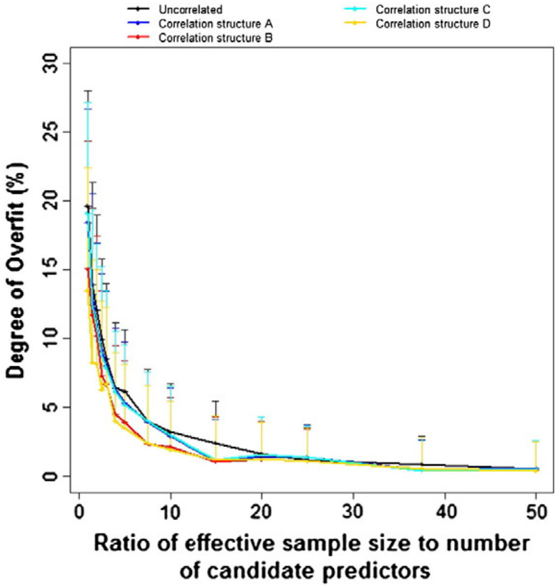

Fig. A2. Effect of different correlation structures among predictors to the degree of overfit for the high signal data. Correlation structures (A): all informative predictors pairwise correlated with $r = 0.25$, non-informative predictors uncorrelated (B) same as (A), but with r = 0.75 (C) informative and non-informative predictors pairwise correlated with r = 0.25 (D) same as (C), but with $r = 0.75$. The bars represent +1 standard deviation.

References

[1] Simon R. Clinical trials for predictive medicine. Stat Med 2012;31:3031–40.

[2] 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.

[3] Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci USA 2002;99:6562–6.

[4] Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–87.

[5] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed.Springer; 2009 .

[6] Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. Ann Intern Med 1993;118:201–10.

[7] Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med 2004;66:411–21.

[8] Concato J, Peduzzi P, Holford TR, Feinstein AR. Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. J Clin Epidemiol 1995;48:1495–501.

[9] Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol 1995;48:1503–10.

[10] Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y. Design and analysis of DNA microarray investigations. Springer; 2004 .

[11] Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene-expression data. J Am Stat Assoc 2002;97:77–87.

[12] 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.

[13] Core Team R. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Austria: Vienna3-900051-07-0; 2012 [URL http://www.R-project.org/].

[14] Maechler M. sfsmisc: utilities from Seminar fuer Statistik ETH Zurich. R package version 1.0–23. http://CRAN.R-project.org/package=sfsmisc; 2012.

[15] Hastie T, Tibshirani R, Narasimhan B, Chu G. pamr: Pam: prediction analysis for microarrays. R package version 1.54. http://CRAN.R-project.org/package=pamr; 2011.