BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 14 2007, pages 1768–1774
doi:10.1093/bioinformatics/btm232

Gene expression

Assessment of survival prediction models based on microarray data

Martin Schumacher[1,*], Harald Binder[2] and Thomas Gerds[2]

[1]Department of Medical Biometry and Statistics, Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg and [2]Freiburg Center of Data Analysis and Model Building, University Freiburg, Germany

Received on December 21, 2006; revised on April 24, 2007; accepted on April 26, 2007

Advance Access publication May 7, 2007

Associate Editor: Chris Stoeckert

ABSTRACT

Motivation: In the process of developing risk prediction models, various steps of model building and model selection are involved. If this process is not adequately controlled, overfitting may result in serious overoptimism leading to potentially erroneous conclusions.

Methods: For right censored time-to-event data, we estimate the prediction error for assessing the performance of a risk prediction model (Gerds and Schumacher, 2006; Graf *et al.*, 1999). Furthermore, resampling methods are used to detect overfitting and resulting overoptimism and to adjust the estimates of prediction error (Gerds and Schumacher, 2007).

Results: We show how and to what extent the methodology can be used in situations characterized by a large number of potential predictor variables where overfitting may be expected to be overwhelming. This is illustrated by estimating the prediction error of some recently proposed techniques for fitting a multivariate Cox regression model applied to the data of a prognostic study in patients with diffuse large-B-cell lymphoma (DLBCL).

Availability: Resampling-based estimation of prediction error curves is implemented in an R package called `pec` available from the authors.

Contact: sec@imbi.uni-freiburg.de

1 INTRODUCTION

In cancer and other chronic diseases, mostly only moderate predictive accuracy can be achieved with clinical data or single biochemical or molecular markers (Schumacher *et al.*, 2006). A risk prediction model in survival analysis assigns survival probabilities to every new patient based on the information that is available at the time origin, e.g. just before start of therapy. It is hoped that high-dimensional genomic and proteomic information, for example obtained via gene expression measurement or protein mass spectrometry, could considerably improve the predictive ability of such models. The principle for developing such models with a sample cohort is to classify the patients such that the variability of the response is small within classes and high between classes. However, censoring disables application of standard discrimination methods and due to the high dimensionality also standard survival methods, for example multivariate Cox regression, cannot be directly applied. Currently used statistical strategies are either *ad hoc* such as gene signatures based on univariate Cox regression analyses, or not yet fully developed. We consider here principle components partial Cox regression (Li and Gui, 2004) and a Cox path algorithm which combines shrinkage and variable selection (Park and Hastie, 2006). It has to be anticipated that in situations where the number of explanatory variables exceeds by far the number of patients in the sample cohort, the overfitting of naively applied statistical strategies and resulting overoptimism of the prediction error may be overwhelming.

The standard practice is to use only part of the data of a study as a training set for the development of the risk prediction model. The complementary data is then used for estimating the prediction error (Molinaro *et al.*, 2005; Segal, 2006) . We show via resampling that this data splitting approach leads to estimates of the prediction error that are characterized by high finite sample variation. Furthermore, the approach is generally inefficient as it is wasting useful and highly valuable information. Studies in this field are usually rather small with respect to the number of patients or individuals (a few hundred) but large in the number of potential predictors (some thousand). Data splitting would further reduce the size of the already small training set that is used for the development of the risk prediction model thus increasing problems of instability and overfitting. However, there are alternatives available. The bootstrap cross-validation and the 0.632+ estimator are based on resampling (Efron, 1983; Efron and Tibshirani, 1997) and designed to improve on k-fold cross-validation.

There is, however, only limited experience so far whether the bootstrap resampling approach will also work in high-dimensional settings. Molinaro *et al.* (2005) and Braga-Neto and Dougherty (2004) report on problems in situations with very small sample sizes ($n = 40$), but restrict their investigation to the classification problem. In this article, we use modifications of the bootstrap resampling approach for censored time-to-event response variable data (Gerds and Schumacher, 2006, 2007; Graf *et al.*, 1999) and explore the practical performance in the data of the study by Rosenwald *et al.*,

*To whom correspondence should be addressed.

(2002), i.e. in a situation which is characterized by a large number of predictor variables ($d = 7399$) and a moderate sample size ($n = 240$). The aim is to improve the accuracy of prognosis of patients with diffuse large-B-cell lymphoma (DLBCL). The response in the retrospective sample cohort is either the time from chemotherapy to death of the patient or the patient is alive at the end of the individual follow-up period (right censored).

2 METHODS

2.1 Developing risk prediction models

Suppose given is a sample cohort of size n where for each patient observed is the survival status $Y_i(t)$ between the time origin and an individual right censoring time C_i, and a d-dimensional vector of covariates Z_i which is measured at baseline (t= 0). If the patient is alive at time t then $Y_i(t) = 1$ and $Y_i(t) = 0$ otherwise. At baseline, the aim is to predict future survival status for all patients in the population. The predictions are quantified in terms of survival probabilities. Predicted survival probabilities are interpreted by the patient (Kattan, 2002) and can be derived from a hazard regression model, e.g. by the product limit method (Andersen *et al.*, 1993).

Classical survival techniques are based on maximization of the multivariate partial likelihood (Cox, 1972). They suffer from the limitation that for obvious reasons the dimension of the covariate vector Z_i,d, must be smaller than the number of individuals, n. In the high-dimensional setting d is usually much larger than n, thus prohibiting direct maximization of the multivariate partial likelihood. To overcome these difficulties, several techniques have been proposed; we consider three of them in the sequel.

The first approach is an *ad hoc* approach based on univariate Cox regression analyses considering one feature (i.e. the expression level of a gene) at a time. The features corresponding to the twenty smallest P-values so obtained are considered a gene signature. In a second step, a multivariate Cox regression model is fitted with all features in the gene signature; this leads to individual prognoses as described in Section 2.2.

The second approach is a generalization of the partial least squares method (e.g. Garthwaite, 1994; Wold, 1966) for time-to-event data proposed by Li and Gui (2004) as partial Cox regression analysis. In order to further reduce the dimension, the authors combined that approach with prior principal component analysis on the matrix of covariate vectors, leading to a principal components partial Cox regression (PC-PCR) method. So the principal components (instead of the original covariates) are used to derive partial components. For each of the latter a Wald test P-value is calculated from univariate Cox models. The first k components with P-value smaller than a cutoff are included in a multivariate Cox analysis. We use a cutoff of 0.05 which results in 4 components for the full data and 4–5 components for most of the bootstrap samples used. From a final fit, one may recover estimated regression coefficients for all the d initial covariates by reversing the transformations.

The third approach, CoxPath, is based on penalized maximization of a multivariate partial likelihood. In classical linear regression with L_2 norm penalty this method is known as 'ridge regression' (Hoerl and Kennard, 1970). However, we think that penalties based on the L_1 norm are advantageous for the high-dimensional setting, because they result in sparser fits; this requires special estimation techniques (Tibshirani, 1996). Often the aim is to compute whole coefficient paths, corresponding to a range of different values for the penalty parameter. For example, this facilitates choice of the penalty parameter. Path algorithms (Efron *et al.*, 2004; Park and Hastie, 2006) provide these paths in a computationally efficient way. Although this method needs quite some computer power, it is feasible in the high-dimensional setting. At a given value for the penalty parameter the result is similar to that of a multivariate Cox regression analysis, however with shrinked regression coefficients. We found that there are problems with automated choosing of the penalty parameter, since standard selection criteria based on degrees of freedom, such as AIC or BIC, fail spectacularly in our example. Therefore, we use 5-fold cross-validation on the full data to select the model complexity. On the full data this resulted in a model with only 37 out of 7399 features. In the bootstrap samples, we use the same model complexity, resulting in a similar number of features. We are aware that ideally cross-validation should be performed in every bootstrap sample. But, due to the considerable additional computational burden we leave this step out of the present evaluation of the CoxPath procedure. Furthermore, the properties of model complexity selection in high-dimensional bootstrap samples are still unclear.

2.2 Measures of prediction error

A prediction rule r is a statistical method or strategy that uses the information of training data $\mathbf{X} = \{X_i : i \in \mathcal{T}\}$ to develop a risk prediction model $r(\mathbf{X})$. Here X_i represents the information on survival and covariates which is available for patient i, see Section 2.1, and $\mathcal{T}$ refers to a subset of the sample cohort, i.e. $\mathcal{T} \subseteq \{1, \ldots, n\}$. All three approaches introduced in the previous section provide a multivariate Cox regression-like result for the conditional hazard function based on $\mathbf{X}$:

$$\lambda(t \mid Z_i) = \widehat{\lambda}_0(t) \exp(\widehat{\beta}^T Z_i). \tag{1}$$

Here, $\widehat{\beta}$ is a d-dimensional vector of estimated (shrinked) regression coefficients and $\widehat{\lambda}_0$ an estimate of the baseline hazard function. From the fitted model there are several ways to construct a risk prediction model. A popular approach first builds classes, e.g. based on quantiles of the linear predictors $\widehat{\beta}^T Z_i$, and then predicts survival probabilities based on stratified Kaplan–Meier analyses. A patient of the population is first classified according to her or his value for the linear predictor and then assigned the respective survival probabilities of this class, which are her or his predictions. However, the transformation from individual values for the linear predictor to a finite often small number of classes can produce substantial bias (Gerds and Schumacher, 2001). Therefore, we consider the values of (1) as a continuous classifier and derive survival probabilities directly via the usual product limit method. The predicted survival probability for patient i when obtained with the risk prediction model $r(\mathbf{X})$ is denoted $r_{\mathbf{X}}(t \mid Z_i)$ in the following. So, e.g. if r is a simple Cox model, then $r_{\mathbf{X}}(t \mid Z_i)$ is obtained by estimating the regression coefficients and the baseline hazard and deriving from that the predicted survival probability.

The prediction error is defined via Brier's score (Brier, 1950) as a function of time as follows. If the patient is alive at time t the predicted survival probability should be close to 1, and otherwise close to 0; in summary, the following expression should be minimized:

$$\overline{err}(t; r_{\mathbf{X}}) = \sum_i \{Y_i(t) - r_{\mathbf{X}}(t \mid Z_i)\}^2 W(t, \widehat{G}, X_i) \tag{2}$$

where summation is over a validation set and will be standardized. The weights remove a large sample censoring bias and are given by

$$W(t, \widehat{G}, X_i) = \frac{1\{T_i \leq t, T_i \leq C_i\}}{\widehat{G}(T_i-)} + \frac{1\{T_i > t\}}{\widehat{G}(t)}$$

where, T_i is the time of death for the uncensored patients in the validation set and $\widehat{G}(t)$ denotes an estimate of the conditional probability of being uncensored at time t given the history (Gerds and Schumacher, 2006; Van der Laan and Robins, 2003). In the high-dimensional setting considered here the history includes the genomic information. Thus, to achieve feasibility of the inverse of probability of censoring weighted estimate we tentatively assume that the censoring mechanisms is independent of the survival and the history. Then the Kaplan–Meier estimate of the censoring survival function substituted for $\widehat{G}$ makes (2) consistent for the prediction error (Gerds and Schumacher, 2006; Graf *et al.*, 1999). The prediction error as we use it is a special case of a more general loss function approach (Korn and Simon, 1991). In the special case of a classification problem with predicted class labels instead of predicted class probabilities, it reduces to the well-known misclassification rate.

2.3 Bootstrap cross-validation and the 0.632+ estimator

Estimation of prediction error is generally difficult because there usually is no data available for this task. The usual way is to take some data away from the modeling process, yielding the split-sample estimate of the prediction error. Some problems with this approach have been discussed in Section 1; they will be explored in Section 3.

The idea of the 0.632+ estimator is to (linearly) combine the *apparent error* that underestimates the prediction error with a *bootstrap cross-validation* estimate that overestimates the prediction error. The *apparent error* is obtained when the summation in (2) is over the training set **X**. *Overoptimism* refers to the potential negative bias which occurs when a risk prediction model is validated with the same data in which it was developed. For example, in the context of the Cox path algorithm, if the penalty is chosen such that the number of covariates in the model exceeds the number of patients, then the overfitting can be so overwhelming that the apparent error equals zero at all times. To obtain the *bootstrap cross-validation* estimate, we first draw with replacement B bootstrap samples $\mathbf{X}_b^*$ each of size n from the full sample population $\{\mathbf{X}_i\text{: } i \in 1,\ldots,n\}$. Then the risk prediction model r_b^* developed on $\mathbf{X}_b^*$ is validated in the set of all individuals not included in the b-th bootstrap sample, i.e. in $\mathbf{X}_b^0 = \{X_i \notin \mathbf{X}_b^*\}$. The *bootstrap cross-validation* estimate of the prediction error at time t is the average:

$$\widehat{Err}_{B0}(t,r) = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{b_0}\sum_{i\in\mathbf{X}_b^0}\left\{Y_i(t) - r_b^*(t|Z_i)\right\}^2 W(t,\widehat{G},X_i)$$

where b_0 is the cardinality of $\mathbf{X}_b^0$. The estimate $\widehat{Err}_{B0}(\,)$ tends to a positive bias for the true prediction error of r because the bootstrap samples contain not the full information of the sample cohort which would ideally be used for developing the rule. Efron and Tibshirani (1997) propose a linear combination of the form

$$\widehat{Err}_{\omega}(t,r) = \{1-\omega(t)\}\overline{err}(t,r) + \omega(t)\widehat{Err}_{B0}(t,r)$$

where the special choice $\omega(t) = 0.632$ gives the famous 0.632 estimator (Efron, 1983) and where the time dependent extension is proposed in Gerds and Schumacher (2007). Now, our setting is characterized by a potentially over-determined system of covariates and patients. The idea is to let $\omega(t)$ reflect the amount of overfitting of a given prediction rule. Therefore, consider as a worst case scenario the situation that event status and covariates are independent. The so-called *no-information error* (Efron and Tibshirani, 1997) assesses the performance of the prediction rule in this scenario:

$$NoInf(t,r) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\{Y_i(t) - r(t|Z_j)\}^2 W(t,\widehat{G},X_i).$$

Based on this quantity define the *relative overfit* as

$$\widehat{R}(t) = \frac{\widehat{Err}_{B0}(t,r) - \overline{err}(t,r)}{NoInf(t,r) - \overline{err}(t,r)}.$$

This yields the time-dependent version of the 0.632+ estimator with weights $\widehat{\omega}(t) = 0.632/(1-0.368\widehat{R}(t))$ (Gerds and Schumacher, 2007). If $\widehat{R}(t) \sim 0$, then $\widehat{Err}_{0.632+} \sim \widehat{Err}_{0.632}$, and if $\widehat{R}(t)$ is close to 1, then $\widehat{Err}_{0.632+}$ is close to $\widehat{Err}_{B0}$.

2.4 Benchmark values

Various benchmark values are available for the prediction error at time t of a risk prediction model. For example, the values 0.25 and 0.33 correspond to constant predicted survival probability of 50% and to a random number between 0 and 100%, respectively. For judging the prediction error curves, we will however rely on the benchmark prediction error which is obtained with the overall Kaplan–Meier estimator for the survival function. This simple ‘risk prediction model’ corresponds to a classifier which assigns the same class to all patients in the population. It ignores the available covariate information completely and thus provides a suitable benchmark value similar as is obtained with the null model in linear regression.

3 APPLICATION TO DLBCL STUDY

The DLBCL study is a prognostic study (Rosenwald *et al.*, 2002) with retrospective collection of tumor-biopsy specimens and clinical data for 240 patients with untreated DLBCL. After a median follow-up period of 2.8 years, 138 deaths have been observed; the 5-year overall survival rate is estimated as 48%. Lymphochip cDNA microarray technology with $P = 7399$ genes was applied; a detailed description can be found in the original publication (Rosenwald *et al.*, 2002). For development of a survival prediction model, the data set was split into training set ($n = 160$) and test set ($n = 80$) by these authors. Up to now, various attempts to create predictive models of survival (from time of chemotherapy) in DLBCL patients developed in these data have been published; a comprehensive review is provided by Segal (Segal, 2006).

To impute missing gene expression levels, we basically follow Li and Gui (2004) who prepared the same data. In particular, we impute for each missing feature value the mean expression level of the nearest eight features, based on Euclidian distance, observed for this patient. In the rare cases where all these eight values are missing, we impute the mean of the block consisting of the nearest features over all patients.

For calculating the prediction error curves of the three methods, PC-PCR, CoxPath and *ad hoc* Cox, in each of $B = 100$ bootstrap samples the whole process of developing the risk prediction models is carried out separately. Throughout, the inverse of probability of censoring weighted scheme $W(t,\widehat{G},X_i)$ is estimated by the Kaplan–Meier estimate of the censoring survival function in the whole data set.

Figures 2–4 show the prediction error curves for those methods. for each method, the apparent error (broken lines) is contrasted with the bootstrap cross-validation estimator (dotted lines) and the 0.632+ estimator (solid, black lines) of the prediction error curve. The prediction error curves obtained in the bootstrap samples ($B = 100$) are shown as shadow plots. The bootstrap cross-validation estimate is the mean of these curves.

We first look at the apparent error (broken lines). Compared to the Kaplan–Meier benchmark value (gray, solid lines), there seems to be a substantial reduction of prediction error by all three methods. PC-PCR seems to perform best with an estimated prediction error close to zero; when CoxPath is evaluated at the step chosen by 5-fold cross-validation the apparent error is slightly below the one obtained for the *ad hoc* method which considers the best 20 features of univariate

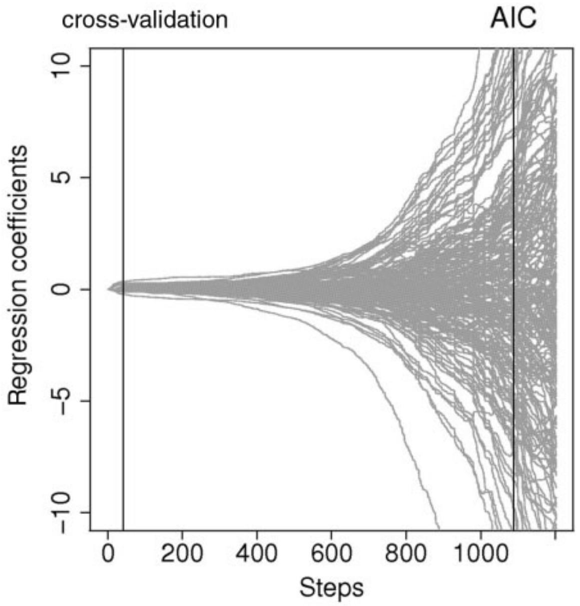

Fig. 1. Rosenwald DLBCL study—results of CoxPath in full data. Each gray line represents the shrinked regression coefficient of one feature.

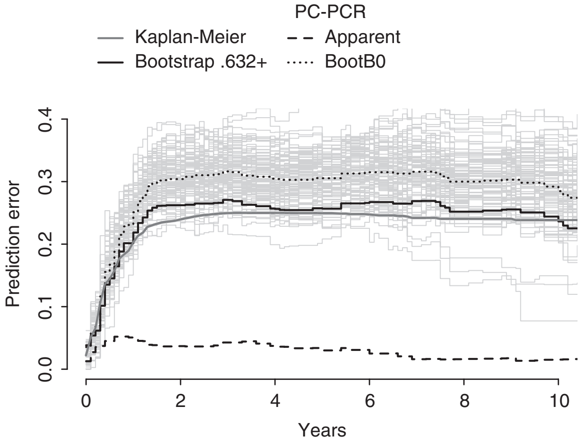

Fig. 2. Rosenwald DLBCL study—prediction error: PC-PCR.

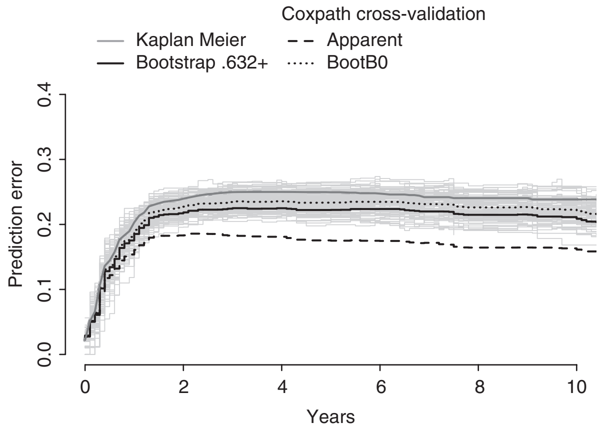

Fig. 3. Rosenwald DLBCL study—prediction error: CoxPath cross-validation.

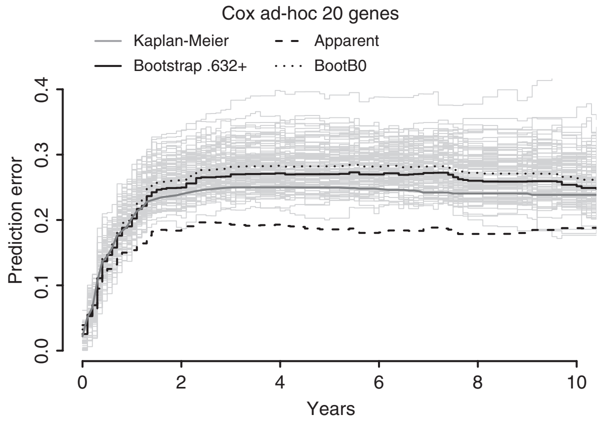

Fig. 4. Rosenwald DLBCL study—prediction error: Cox: *ad hoc* best 20 genes.

analyses in a multivariate Cox model. However, if we evaluate the model at step 1088 which is suggested by the partial likelihood version of AIC (see Fig. 1), then the resulting model includes 205 features and the corresponding apparent error (data not shown) is similar to the one of PC-PCR.

For all three methods, the high variability of the individual bootstrap prediction error curves is clearly visible as shown by the width of the shadow plots. As similar variability is observed for the steps of 2-fold cross-validation, showing high finite sample variance of the simple data splitting approach.

The 0.632+ estimates of the true prediction error indicate that, with the exception of CoxPath, the developed prediction models are no better than those based on the Kaplan–Meier prediction ignoring all available covariate information. The results of CoxPath are promising in that the 0.632+ estimate is at least smaller than the Kaplan–Meier benchmark value. However, recall that the choice of the penalty parameter is crucial and that we have used a workaround. As an alternative to using cross-validation in every bootstrap sample, a consequent step towards automated penalty parameter selection would be to use *bootstrap cross-validation* and a criterion that is based on a summary of the prediction error curves. However, the current implementation of CoxPath is still very computer intensive and a feasible solution is part of our future research.

The relative overfit (defined at the end of Section 2.3) of the three methods is displayed in Figure 5 with values between 60 and 80% throughout the follow-up period of about 10 years. This indicates that the 0.632+ estimator is close to the bootstrap cross-validation estimator for all methods. However, it can also be seen that the latter is somewhat too pessimistic.

Figure 6 summarizes the results for the three methods in terms of prediction error curves obtained by the 0.632+ estimator (Fig. 6) and obtained by splitting the data set into

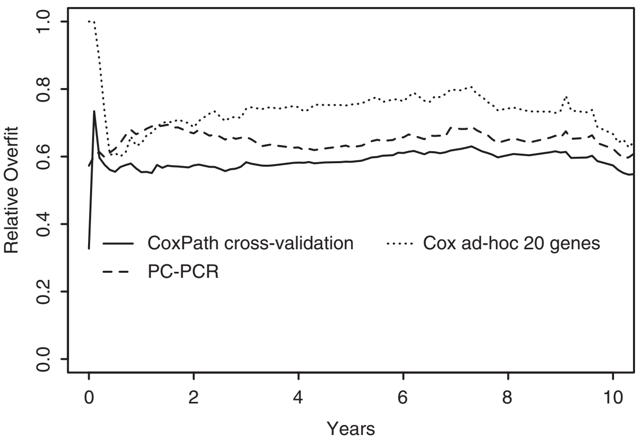

Fig. 5. Rosenwald DLBCL study—apparent error—relative overfit.

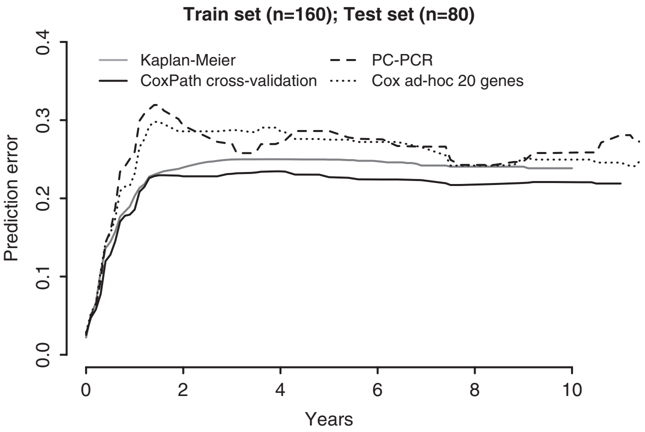

Fig. 7. Rosenwald DLBCL study—trainingset ($n = 160$); testset ($n = 80$).

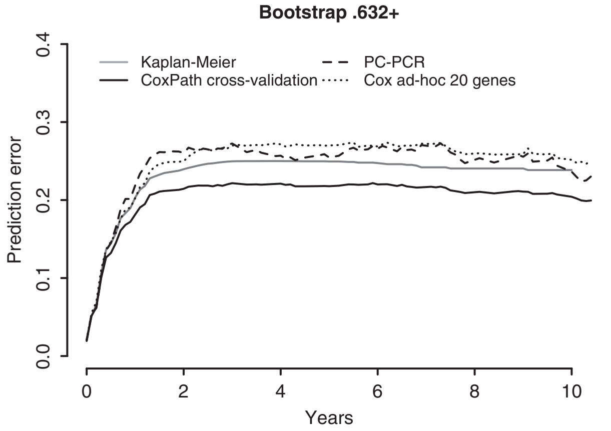

Fig. 6. Rosenwald DLBCL study—Bootstrap 0.632+.

fixed training set ($n = 160$) and test set ($n = 80$) (Fig. 7), as it has been done in earlier publications (Rosenwald *et al.*, 2002; Segal, 2006). It can be seen that the resulting prediction error curves are in nice agreement, however, they are associated with much larger variability.

4 DISCUSSION AND CONCLUSION

Recent years have seen a rapid increase in availability of high-dimensional genomic information in biomedical research that could potentially be used for risk prediction, diagnosis, prognosis and targeted therapeutic interventions (Simon, 2005). The situation is almost always characterized by a relatively small number of individuals and a large number of potential predictors, thus prohibiting the direct use of classical approaches of statistical modeling and data analysis. A variety of proposals have been made to overcome the problem intrinsic to such a high-dimensional setting, most of them are focused on classification problems. Some suggestions have been made with regard to time-to-event data in order to also cover the prognostic and risk prediction situation where time and potential censoring has to be taken into account. Most of these suggestions can be regarded as extensions of the well-known Cox proportional hazards regression model (Cox, 1972) or, better, as techniques for fitting a Cox regression model to high-dimensional data. We used two of these proposals (Li and Gui, 2004; Park and Hastie, 2006) in an exemplary manner and contrasted them with an *ad hoc* approach; other proposals are available but are beyond the scope of this article.

The high-dimensional setting can be seen as a unique opportunity to create better risk prediction and prognostic models compared to those that are already available and are solely based on clinical data and/or single markers, but considerable overfitting has to be anticipated. In order to come to a valid assessment of predictive accuracy, data splitting is usually employed. We argue that especially in a high-dimensional setting model building should be based on all available data and not be restricted to a smaller subset. More efficient than simple data splitting are resampling methods, in which the process of model building on the full data is imitated. Especially, we advocate the use of the 0.632+ estimator of prediction error (Efron and Tibshirani, 1997) that we have adapted for application to time-to-event data (Gerds and Schumacher, 2007). The resulting prediction error curves show the relative merits of a risk prediction model over time when compared to the Kaplan–Meier benchmark value that ignores all (high-dimensional) covariate information (Gerds and Schumacher, 2006).

We also investigated the usefulness of the 0.632+ estimator and other resampling plans for assessing and comparing predictive power of various modeling approaches with varying degree of flexibility using the data of the GBSG-2 study, a clinical trial in breast cancer patients, that has already been used by ourselves and other authors beforehand (Schumacher *et al.*, 2006). An important conclusion from this empirical investigation (Gerds and Schumacher, 2007) is that in a standard regression setting (e.g. a pre-specified Cox model with a limited number of covariates but moderate number of events) the apparent error is almost identical to the prediction error obtained with bootstrap-based approaches, indicating that the former tracks the true prediction error quite well. This is in agreement with the large sample results of Gerds and Schumacher, (2006). For the two flexible modeling strategies studied (regression trees and artificial neural networks) it was shown that the apparent error rates are seriously biased, but this was markedly reduced when penalty terms or restrictions were used. Thus, we got similar experiences as we obtained in the high-dimensional setting investigated in this article.

In the application to the Rosenwald DLBCL data (Rosenwald *et al.*, 2002) we were able to show that the 0.632+ estimator also 'works' in a high-dimensional setting, i.e. it enables to track the true prediction error without splitting the data into training and test set, even for a model where the apparent error is close to zero. Our findings can be summarized as follows: all prediction models developed show prognostic potential in terms of apparent error; PC-PCR (Li and Gui, 2004) seems to be best with a prediction error curve close to the zero line. However, overfitting turned out to be substantial, leading to far too optimistic values of prediction error. The 0.632+ estimates of prediction error indicate that, with the exception of CoxPath (Park and Hastie, 2006), the developed prediction models are no better than the Kaplan–Meier benchmark value ignoring all covariate information. It is worth noting that the 0.632+ estimates are in nice agreement to those obtained with usual data splitting, a procedure that was also employed by Segal (Segal, 2006) in his investigation on the various published risk prediction models that have been created with the Rosenwald DLBCL data. Although he used a different measure to assess their predictive performance, namely time-dependent ROC curves (Heagerty and Zheng, 2005; Heagerty *et al.*, 2000), he came to the very similar conclusion that there is only little if any prognostic information in the prediction models developed in these high-dimensional genomic data so far.

Although our investigation was not focused on a thorough investigation or on an improvement of existing proposals for the analysis of survival data with high-dimensional covariates, we found that the Cox path algorithm involving L_1 regularization appeared to have the most potential. This should be further evaluated and compared with other recent proposals especially based on different methods for regularization (Gui and Li, 2005; Van Houwelingen *et al.*, 2006).

There have been a few investigations on prediction error estimation in a high-dimensional setting. Within a classification framework, Molinaro *et al.* (2005) compared by simulation various approaches, including the 0.632+ and leave-one-out cross-validation estimates. They found large variability of the latter, especially when flexible prediction models were used. This is a known drawback of leave-one-out cross-validation (Wehberg and Schumacher, 2004) that originally has lead Efron and Tibshirani to propose improvements on cross-validation (Efron, 1983; Efron and Tibshirani, 1997). On the other hand, (Molinaro *et al.*, 2005) report on problems with the 0.632+ estimate, especially in very small samples, ($n = 40$) which might be attributed to difficulties with 'ties', i.e. individuals that are included more than once in a bootstrap sample. These difficulties diminished with increasing sample sizes ($n = 80$ and $n = 120$). Although we were not confronted with that problem in a serious way when using the 0.632+ estimate of prediction error in the Rosenwald DLBCL data, this issue deserves further attention. Recently, a proposal of combining bootstrap resampling and leave-one-out cross-validation has been published (Fu *et al.*, 2005), but has not been investigated in a high-dimensional setting yet. Finally, we would like to emphasize that the 0.632+ estimate will only 'work', i.e. will track the true prediction error, if all steps of model building are repeatedly performed within each bootstrap sample. This implies, e.g. that preliminary selection of genes may not be ignored. The consequences of doing so have already been impressively demonstrated (Simon *et al.*, 2003).

ACKNOWLEDGEMENTS

This investigation has been supported by the Deutsche Forschungsgemeinschaft

Conflict of Interest: none declared.

REFERENCES

Andersen,P.K. *et al.* (1993) *Statistical Models Based on Counting Processes*. Springer Series in Statistics, Springer, New York.

Braga-Neto,U.M. and Dougherty,E.R. (2004) Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, **20**, 374–380.

Brier,G.W. (1950) Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.*, **78**, 1–3.

Cox,D.R. (1972) Regression models and life tables. *J. R. Stat. Soc. B.*, **34**, 187–220.

Efron,B. (1983) Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. Stat. Assoc.*, **78**, 316–331.

Efron,B. and Tibshirani,R. (1997) Improvements on cross-validation: the 0.632+ bootstrap method. *J. Am. Stat. Assoc.*, **92**, 548–560.

Efron,B. *et al.* (2004) Least angle regression. *Ann. Stat.*, **32**, 407–499.

Fu,W.J. *et al.* (2005) Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics*, **21**, 1979–1986.

Garthwaite,P. (1994) An interpretation of partial least squares. *J. Am. Stat. Assoc.*, **89**, 122–427.

Gerds,T.A. and Schumacher,M. (2001) On functional misspecification of covariates in the cox regression model. *Biometrika*, **88**, 572–580.

Gerds,T.A. and Schumacher,M. (2006) Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biom. J.*, **48**, 1029–1040.

Gerds,T.A. and Schumacher,M. (2007) Efron-type measures of prediction error for survival analysis. *Biometrics*, doi: 10.1111/j.1541-0420.2007.00832.

Graf,E. *et al.* (1999) Assessment and comparison of prognostic classification schemes for survival data. *Stat. Med.*, **18**, 2529–2545.

Gui,J. and Li,H. (2005) Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, **21**, 3001–3008.

Heagerty,P. and Zheng,Y. (2005) Survival model predictive accuracy and roc curves. *Biometrics*, **61**, 92–105.

Heagerty,P. *et al.* (2000) Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, **56**, 337–344.

Hoerl,A. and Kennard,R. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Kattan,M. (2002) Statistical prediction models, artificial neural networks, and the sophism 'I am a patient, not a statistic'. *J. Clin. Oncol.*, **20**, 885–887.

Korn,E.L. and Simon,R. (1991) Explained residual variation, explained risk, and goodness of fit. *Am. Stat.*, **45**, 201–206.

Li,H. and Gui,J. (2004) Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, **20** (Suppl. 1), I208–I215.

Molinaro,A.M. *et al.* (2005) Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, **21**, 3301–3307.

Park,M. and Hastie,T. (2006) l_1 regularization path algorithm for generalized linear models. *Technical report*, Department of Statistics, Stanford University.

Rosenwald,A. *et al*. (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *N. Engl. J. Med*., **346**, 1937–1947.

Schumacher,M. *et al*. (2006) Prognostic factor studies. In Crowley,J. and Pauler Ankerst,D. (eds) *Handbook of Statistics in Clinical Oncology*. 2nd edn. Chapman & Hall, pp. 289–333.

Segal,M.R. (2006) Microarray gene expression data with linked survival phenotypes: diffuse large-B-cell lymphoma revisited. *Biostatistics*, **7**, 268–285.

Simon,R. (2005) Roadmap for developing and validating therapeutically relevant genomic classifiers. *J. Clin. Oncol*., **23**, 7332–7341.

Simon,R. *et al*. (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl Cancer Inst*., **95**, 14–18.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*., **58**, 267–288.

Van der Laan,M.J. and Robins,J.M. (2003) *Unified Methods for Censored Longitudinal Data and Causality*. Springer, New York.

Van Houwelingen,H.C. *et al*. (2006) Cross-validated cox regression on microarray gene expression data. *Stat. Med*., **25**, 3201–3216.

Wehberg,S. and Schumacher,M. (2004) A comparison of nonparametric error rate estimation methods in classification problems. *Biome. J*., **46**, 35–47.

Wold,H. (1966) Estimation of principal components and related models by iterative least squares. In Krishaiaah,P. (ed.) *Multivariate Analysis*. New Academic Press, New York, pp. 391–420.