The Performance of Risk Prediction Models

Thomas A. Gerds[*,1], **Tianxi Cai**[2], and **Martin Schumacher**[3]

[1] Department of Biostatistics, Øster Farimagsgade 5 opg. B, Postboks 2099, 1014 København
[2] Department of Biostatistics, 655 Huntington Avenue, Building I Room 411,Boston, MA 02115 2
[3] Institute of Medical Biometry and Medical Informatics, University of Freiburg, Stefan-Meier-Strasse 26, D-79104 Freiburg, Germany

Received 16 May 2008, revised 28 May 2008, accepted 7 June 2008

Summary

For medical decision making and patient information, predictions of future status variables play an important role. Risk prediction models can be derived with many different statistical approaches. To compare them, measures of predictive performance are derived from ROC methodology and from probability forecasting theory. These tools can be applied to assess single markers, multivariable regression models and complex model selection algorithms. This article provides a systematic review of the modern way of assessing risk prediction models. Particular attention is put on proper benchmarks and resampling techniques that are important for the interpretation of measured performance. All methods are illustrated with data from a clinical study in head and neck cancer patients.

Key words: AUC; Brier score; 0.632+ Bootstrap; Crossvalidation; Machine learning; Model selection; R^2; Risk prediction; ROC curve.

1 Introduction

An important medistatistical task is to translate research findings into diagnoses and prognoses validated to be of clinical utility (Simon, 2005b). Indeed, in many medical applications the final aim of all statistical modelling is to assist medical decision making. Examples where systematic data analysis led to prognostic tools that are used in practice are the international prognostic index (IPI, 1993), the prostate cancer risk calculator (Thompson et al., 2006), and the Kattan nomogram (Kattan et al., 2000). A strategy for developing a prognostic tool that finally provides evidence based information for a new patient requires careful planning and evaluation. It requires on the statistical side a tool that distinguishes candidate prognostic factors from noise and also a measure for the performance of candidate models. Eventually the model search has to be assessed as a whole (Simon, 2005a).

The central aim of this article is to review and illustrate the modern usage of the classical statistical tools for prognostic and diagnostic studies. The main message is that the performance of a model has to be estimated, and that there are potential biases incurred by overfitting and by censored data.

For illustration, we use data from a clinical trial on epo treatment of anemic cancer patients (Henke et al., 2006). Considered are performance measures derived from receiver operating characteristic (ROC) methodology and from probability forecasting theory. The tools have been described and compared by many (e.g. Hand, 1997; Shapiro, 1999; Hilden, 2000; Hand, 2001; Pepe, 2003; Gail and Pfeiffer, 2005; Pencina et al., 2008). The main conclusions are typically quite robust to how precisely the performance is measured. For example, the area under the ROC curve and the Brier score often provide similar answers (e.g. Redelmeier et al., 1991; Bartfay et al., 2006). However, the predictive performance of a risk prediction model has various aspects and components, and different measures of

[*] Corresponding author: e-mail: tag@biostat.ku.dk, Phone: +45 353 27914

predictive performance give different emphasize to these components. They include accuracy, calibration, dispersion, discrimination, separability, precision, regret and utility; see Hand (1997) for a comprehensive discussion of the terminology.

Special topics to be discussed and illustrated in more detail are benchmark values and resampling strategies for assessing generalization performance. Further topics are the handling of censored data in survival analysis, time-dependent performance measures, a general R^2 measure, and the added predictive performance of a new marker.

1.1 The focus of model building

Which modelling approach is appropriate depends first of all on the designated (medical) decision, and on *Who is asking the question?*. A patient is generally interested in a personal diagnosis or prognosis but does in principle not care about the insides of the model, i.e. what the specific prognostic factors are or what the assigned risk group is (Kattan, 2002). A researcher on the other hand will often be interested in the insides of the model for example when the aim is to understand a biologic system or to assess the value of a new biomarker. A community is interested in the utility of a model for the population, and so on. The point is that the process of developing a model can and should have a focus that corresponds closely to the target parameter (Claeskens and Hjort, 2003). For example, the study group of the Prostate Cancer Prevention Trial (Thompson et al., 2006) used logistic regression modelling and selected the model which maximized the crossvalidated area under the ROC curve. Thus, they focused on the ability of the model to discriminate the prostate cancer risk of the men in the considered population. It might well be that the finally selected model would not have been suggested by standard goodness-of-fit procedures.

When a new patient seeks advice, then the prognostic model should extract and communicate the information inherent in data collected on similar patients. On a high level a model performs well if it provides information that leads to successful medical decision. However, in real life the patient and the physician will incorporate other practical aspects into the decision making. Thus it is often sufficient to assess and validate the ability of the model to predict accurately the probability for the patient's future status. Given this information a decision can be made and risk categorization is not needed. If, however, risk is categorized then the risk classes should be connected to treatment classes (Pencina et al., 2008). In this article we consider solely predicted risk on a continuous scale.

1.2 The epo study

For illustration we re-analyze data from a clinical trial on erythropoietin therapy for anemic cancer patients (Henke et al., 2006). Anaemia is a deficiency of red blood cells and/or hemoglobin. This randomized placebo controlled trial investigated if treatment with epoetin beta (300 U/kg) enhances hemoglobin concentration and how this translates into an improvement of the survival chances of head and neck cancer patients. The study population considered here consists of 154 patients with a tumor located in the oropharynx (35%), the oral cavity (40%), the larynx (15%) or in the hypopharynx (24%). The patients were randomized within the following strata: radiotherapy following complete resection (Stratum 1: 47%), radiotherapy following incomplete resection (Stratum 2: 19%) and radiotherapy without surgery (34%). Blood hemoglobin levels were measured weekly during the administration of radiotherapy (7 weeks); follow-up visits were scheduled every 3 month. Treatment with epoetin beta was defined successful when the hemoglobin level increased to 14 g/dL in women and 15 g/dL in men, or when the weekly hemoglobin change was higher or equal to 2 g/dL. While the target hemoglobin levels were significantly more often reached in the treatment group, unfortunately the survival chances were significantly better in the placebo group. This latter finding is restricted to Strata 2 and 3 and could be explained by a stimulating effect of epoetin beta on the remaining tumor tissue (Henke et al., 2003). Using Cox regression analysis, Henke et al. (2006) identified the c20

Table 1 Information on the considered predictive factors for two patients of the epo study.

PatNr	Age	Gender	HbBase	Resection	Treat	epoRec
134	51	male	12.6	Compl	Epo	positive
151	50	male	14.3	Incompl	Placebo	negative

expression (erythropoietin receptor status) as a new biomarker for the prognosis of locoregional progression-free survival.

For the purpose of illustration we consider two different status response variables. The first is the epo treatment success defined if the target hemoglobin concentration was reached during the radiation therapy. The treatment was successful in $n = 66$ (44)% and failed in $n = 83$ (56%) cases. The second response is the local disease free survival status. E.g., after three years this status is 'alive' for 41 patients, 'dead' for 101 patients and unknown (right censored) for 7 patients. Different modeling strategies are considered for predicting these status variables based on the predictors: age, gender, treatment, baseline hemoglobin level and erythropoietin receptor status; see Table 2 and Henke et al. (2006) for more details.

In subsequent sections these data are used to illustrate various aspects of modelling and model assessment. In particular, we will follow the predictions obtained with different models for the two sample patients that are characterized in Table 1.

For patient no. 134 the epo treatment was successful and he died 52.8 month after enrollment without prior local relapse. Patient no. 151 received placebo and did not reach the required hemoglobin concentration level. He died after 13.9 month without prior local relapse.

2 Risk Prediction

2.1 The formal framework

Consider a medical data set $\mathcal{L}_n = \{X_i, Y_i(t)\}_{i=1,\ldots,n}$ that contains samples from n patients and consist of an input matrix $X_i = (X_{ik})_{k=1,\ldots,K}$ of K marker variables (the predictors), that are available at time 0, and an outcome vector $Y_i(t)$ that describes the patients' individual status at time $t \geq 0$. Usually X_i provides information on treatment, conventional factors, such as age and gender, biomarkers, genotypes, phenotypes, family history, and so on. A single element of X_i is called a prognostic or predictive factor (in some fields the term 'predictive factor' implies that the patient is treated). The status could be the survival or disease status, or describe the response to treatment, or a similar condition. We assume that $Y_i(t) = 1$ means that the event of interest has occurred before time t and $Y_i(t) = 0$ that it has not. Note that a diagnostic study deals with the patient's current status $Y_i(0)$.

In what follows, risk prediction modelling strategies for the response at time t are denoted r_t. The data $\mathcal{L}_n$ are used to estimate model internal parameters. This yields a trained prediction model $r_t(\mathcal{L}_n) = \hat{r}_{n,t}$ which assigns to patient i a probability for the status at time t:

$$r_t(\mathcal{L}_n)(X_i) = \hat{r}_{n,t}(X_i)\,.$$

This probability is interpreted as a prediction for the unknown status. For example, in a diagnostic study $\hat{r}_{n,0}(X_i) = 0.3$ would be interpreted as a 30% probability that patient i is diseased. In a survival study $\hat{r}_{n,2}(X_i) = 0.8$ would be interpreted as a 80% chance that the patient dies within the first 2 years. In practice different modelling strategies may result in similar or very different results for single patients even if they are build on the data from the same study.

2.2 Simple threshold models

The most simple form of a risk prediction model is a threshold model obtained by splitting the population based on the values of a single marker and an estimated threshold:

$$\hat{r}_{n,t}(X_i) = \begin{cases} 1 & X_i^1 \geq \hat{\xi} \\ 0 & X_i^1 < \hat{\xi}. \end{cases} \tag{1}$$

Such a model predicts either 0% or 100%. The data $\mathcal{L}_n$ are used to estimate the threshold $\hat{\xi}$. To find the threshold with maximal performance, it is possible to utilize one of the performance measures that are discussed below, e.g. the area under the ROC curve as in Uno et al. (2007). Alternatively one finds $\hat{\xi}$ that maximizes the test statistic of an appropriate test for comparing two groups. Although the p-value of such a maximally selected statistic can be easily calculated (Miller and Siegmund, 1982; Lausen and Schumacher, 1992) there is the danger of serious overestimation of effects (Holländer and Schumacher, 2006; Schumacher et al., 2006) that may lead to considerable overfitting of the resulting prediction model.

When there are multiple sources of information available to assist in prediction, a prognostic model may be obtained by combining two or more threshold models. Combining markers is especially useful when they detect differently diseased patients. It is straightforward to combine threshold models for different marker by 'AND' or 'OR' conjunctions. However, the results are usually much better when predictors are combined in a systematic multivariable model building procedure.

As a first example we build a threshold model for the the epo treatment success based on the baseline hemoglobin level. The 75 treated patients in the epo study had 41 different baseline hemoglobin levels. Thus there are 41 possible thresholds. Not a single one will be optimal regarding all criteria. For example, the overall misclassification rate is minimized at $\hat{\xi} = 10.9$ but the Gini split criterion (equivalently the area under the ROC curve) suggests $\hat{\xi} = 11.3$. A correspondingly constructed prediction rule combines the treatment assignment and the baseline hemoglobin level with the 'AND' conjunction: if treatment='Yes' and HbBase$\geq$ 11.3 g/dl then predict the epo treatment success.

2.3 Traditional modelling

Logistic and Cox regression belong to the data modelling culture (Breiman, 2001b). Inference is based on the assumption that the data are generated from a specific form of the conditional distribution of the status at time t given the predictors. The usual aim is to estimate regression coefficients. But once the model is fitted, that means all model specific parameters are estimated, predictions can be obtained for new patients. In most applications the precise form of the model is not pre-specified, and 'logistic regression' and 'Cox regression' are complex strategies that involve the selection of variables and link function.

Let $\mathcal{K} \subseteq \{1, \ldots, K\}$ be a subset of the available predictors, $g = g_1, \ldots, g_k$ a vector of link functions, and $\beta = \beta_1, \ldots, \beta_k$ a vector of regression coefficients. The corresponding logistic regression model defines a risk prediction model via the well-known inverse of the logit transformation:

$$\hat{r}_{n,t}(X_i) = \left[1 + \exp\left\{\hat{\beta}_0 - \sum_{k \in \hat{\mathcal{K}}} \hat{\beta}_k \hat{g}(X_{ik})\right\}\right]^{-1}. \tag{2}$$

Here β_0 is an intercept which often has an interpretation as the baseline risk. The data are used to estimate the regression coefficients β_k. In many cases, the data are also used to search for $\mathcal{K}$ and g such that the resulting model satisfies some optimality criterion, for example predictive performance. In a systematic model selection procedure one can for example use regression splines to estimate g. In the commonly encountered ad-hoc procedures one would for example compare the fit of $\hat{g}(\text{age}) = \text{age}$ to the fit of $\hat{g}(\text{age}) = \text{age} + \text{age}^2$. Whatever strategy is adopted, it is important that for assessing the performance of the finally selected model, all the data driven steps in model selection are considered as part of the model, and repeated for example in crossvalidation.

Table 2 Estimated odds ratios $\exp(\hat{\beta})$ obtained with different logistic regression strategies. The outcome variable is the status of the response to epo treatment. Model $LRM^{(2)}$ omits the marker variable 'epoRec' (erythropoietin receptor status), model $LRM^{(3)}$ is the result of automated backward elimination and in model $LRM^{(4)}$ the patients age is dichotomized at the "optimal" threshold age $>65^*$. The value of the Akaike information criterion (AIC) is given in the last row only for the strategies that did not search for the model.

Variable	$LRM^{(0)}$	$LRM^{(1)}$	$LRM^{(2)}$	$LRM^{(3)}$	$LRM^{(4)}$
Intercept	0.80	0.00	0.00	0.00	0.00
Age	–	0.97	0.98	–	0.19^*
Gender (female : male)	–	4.72	3.98	6.16	6.12
HbBase	–	3.26	2.96	3.22	3.77
Treat(Epo : Placebo)	–	90.49	55.36	60.31	113.25
Resection(Incompl : no)	–	1.75	1.89	–	1.87
Resection(Compl : no)	–	4.13	3.38	–	5.23
epoRec (pos : neg)	–	5.81	–	4.49	5.19
AIC	206.61	103.47	109.8	–	–

Table 2 shows the estimated regression coefficients of five different logistic regression models for the epo treatment success in the epo study: $LRM^{(0)}$ is the null model that ignores all predictive factors and thus predicts the prevalence to each patient. $LRM^{(1)}$ includes all predictors, $LRM^{(2)}$ excludes the erythropoietin receptor status and $LRM^{(3)}$ is obtained from automated backward elimination using the R function 'fastbw' of the Design library (see Harrell, 2001). Model $LRM^{(4)}$ is an ad-hoc strategy which searches through all possible dichotomizations of the variable age to find the one that provides the minimal p-value in a logistic regression model. Otherwise $LRM^{(4)}$ is like $LRM^{(1)}$. Comparing the first two columns of Table 2 shows the typical attenuation of effect estimates in multiplicative models where an important factor is omitted (Struthers and Kalbfleisch, 1986). However, the fact that the effect of the factor erythropoietin receptor status ('epoRec') is significant ($p = 0.0076$) does not necessarily imply that the factor is very important for the predictive ability of the model (Kattan, 2003; Pencina et al., 2008; Uno et al., 2007; Ware, 2006).

Predicted probabilities for epo treatment success are obtained by inserting the estimated coefficients of Table 2 into formula (2). For the sample patients of Section 1.2 the predictions of the different models are given in Table 3.

The Cox regression model is prominently used in survival analysis where the status of some patients is right censored. A Cox model predicts for each (new) patient the survival probability as a function of time. In practice, the predictions are defined until the maximal follow-up time in the data that were used to fit the model. The model building can be complex in the same way as just discussed for the logistic regression model. Using the notation from above, the Cox regression model predicts the following survival probability for a patient with predictors X_i at time t:

$$\hat{r}_{n,t}(X_i) = \exp\left(-\hat{\Lambda}_0(t)\exp\left[\sum_{k\in\hat{\mathcal{K}}}\hat{\beta}_k\ \hat{g}(X_{ik})\right]\right). \tag{3}$$

The data are used to estimate the baseline hazard function $\Lambda_0(t)$ and the regression coefficients β. The data may also be used to specify $\hat{\mathcal{K}}$ and $\hat{g}$, i.e. to search for the functional form.

For illustration, two Cox regression models are considered for predicting the local disease free survival status. $CRM^{(1)}$ adjusts for the tumor resection and the interaction between the epo treatment

Table 3 Predicted epo treatment success probabilities in % for the two sample patients of Table 1. Here CART refers to the classification tree and RF to the random forest model described in Section 2.4; $\text{LRM}^{(0)}$ is the null model that ignores all predictive factors and predicts the prevalence to each patient. For all but the null model $\text{LRM}^{(0)}$ the data of the two patients are concordant since the patient 134 had a positive status and his predicted probabilities are higher ranked than those of patient 151 who had a negative status.

	PatNr	$\text{LRM}^{(0)}$	$\text{LRM}^{(1)}$	$\text{LRM}^{(2)}$	$\text{LRM}^{(3)}$	$\text{LRM}^{(4)}$	CART	RF
Prediction	134	44.3	97.39	93.13	92.15	98.31	92.86	97.6
	151	44.3	18.73	46.83	24.08	25.3	10.81	10.8
Brier Score	134	31.03	0.07	0.47	0.62	0.03	0.51	0.06
	151	19.62	3.51	21.93	5.8	6.4	1.17	1.17
Pair concordant		no	yes	yes	yes	yes	yes	yes

group and the epo receptor status. This model was used in Henke et al. (2006) to explain the lowered survival chances in the epo treatment group. $\text{CRM}^{(2)}$ is like $\text{CRM}^{(1)}$ but adjusts further for the factors age, gender and baseline hemoglobin level.

2.4 Machine learning

Classification trees and random forests are alternatives to the traditional modelling culture models of the previous section that belong to a broader class of machine learning models (Breiman, 2001b). A classification tree combines multiple threshold models in logical if-then combinations (Breiman et al., 1984). In these tree structures, nodes represent classifications and branches represent binary rules for single markers. The model building follows an algorithm of the following kind. In the first tree growing step one uses an appropriate criterion, such as the Gini index, to assess all possible binary rules. The best binary split is implemented yielding to two nodes as in Eq. (1). In the second step one compares the remaining binary rules separately for the two nodes found in the first step. In most cases this procedure is pursued until an appropriate stopping criterion is met. A commonly used stopping criterion is a bound on the minimum number of patients in terminal nodes. Eventually a tree model assigns to a new patient a terminal node $\mathcal{T}(X_i)$. A predicted probability for the status of this patient is obtained as the fraction of patients with positive response status in the corresponding terminal node:

$$\hat{r}_{n,t}(X_i) = \frac{\sum_{\mathcal{T}(X_j)=\mathcal{T}(X_i)} \{Y_j(t) = 1\}}{\sum_{\mathcal{T}(X_j)=\mathcal{T}(X_i)}} . \tag{4}$$

Both sums range over all patients in the training data $\mathcal{L}_n$ that share the terminal node with the new patient. The data are extensively used in the process of model building and it is typically not known in advance how many parameters are to be estimated.

Figure 1 shows a classification tree model (CART) obtained in the data of the epo study for the outcome variable epo treatment success by using the same predictor variables that have entered into the logistic regression model $\text{LRM}^{(1)}$.

The model depicted in Figure 1 classifies the sample patient 134 of Table 1 in the rightmost terminal node; here the probability of treatment response is predicted as 92.9%.

An advantage of classification trees is that they are simple to understand and interpret. A drawback of the method is that it tends to instability and overfit, especially if it is applied in a naive way, see e.g. (Schumacher et al., 2006).

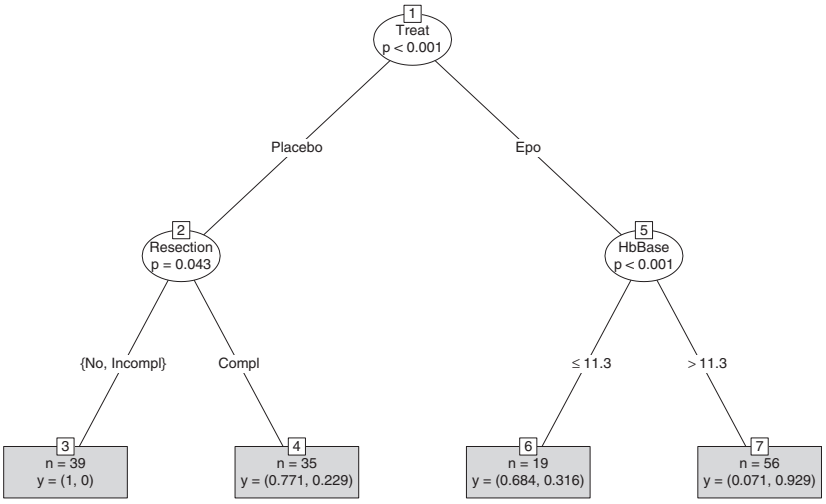

Figure 1 Classification tree model for the outcome variable epo treatment success.

A random forest is an ensemble method that consists of many, usually between 100 and 10 000, classification trees (Breiman, 2001a). It usually provides more stable and reliable results than a single tree model. However, it does not have a nice graphical representation and is often considered a black box – which may not be a problem for the patient as long as the predictive performance is high, compare Section 1.1. To build the model, the single trees are grown in independent bootstrap samples without stopping until all terminal nodes are pure. Predictions of the random forest model correspond to a majority vote from all single trees. A random forest approach uses the data in a much more exhaustive way than a single tree model.

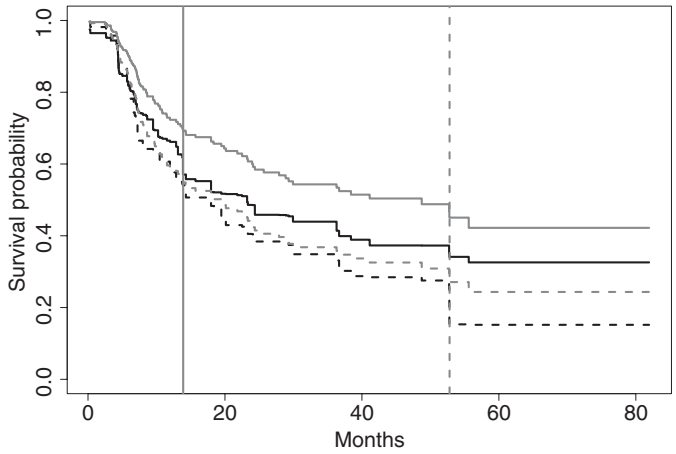

Figure 2 Predicted survival probabilities for the two sample patients of Section 2 (patient 134 solid lines; patient 151 dashed lines) obtained with models $\text{CRM}^{(1)}$ (gray lines) and RSF (black lines). The vertical lines represent the observed death times of these patients.

Using 500 bootstrap samples a random forest model (RF) is trained in the data of the epo study for the outcome variable epo treatment success and the same predictors that have entered the logistic regression model LRM$^{(1)}$. The results for the sample patients of Section 1.2 obtained with the random forest model and the CART model are given in Table 3.

In a similar way a survival model can be build based on the random forest idea. Using again 500 bootstrap samples a random survival forest model (RSF) is trained in the data of the epo study for the local disease free survival status based on the treatment groups, age, gender and baseline hemoglobin level. Figure 2 displays the predictions obtained with the Cox model CRM$^{(1)}$ and the random survival forest model (RSF) for the two sample patients of Section 2.

3 Performance Measures

3.1 The formal framework

In this section we formulate and analyze certain estimation problems where the predictive performance of a model is defined as a parameter of the population which has to be estimated. In several instances performance can be formulated as an expectation taken with respect to the joint distribution of the data of one or two patients. This formulation enables the 'plug-in' principle to estimation and thereby also simplifies the implementation of resampling techniques.

Let $F(t \mid x)$ denote the conditional probability of the event $Y_i(t) = 1$ given $X_i = x$ and $H(x)$ the marginal probability law of the vector X_i. Then the joint probability law of the data of a single patient is $F \times H$. Define also the conditional survival function $S(t \mid x) = 1 - F(t \mid x)$ and the prevalence $\varrho(t) = F(t \mid x)\, H(\mathrm{d}x)$ at time t.

Let $\mathcal{D}$ be a decision space. A performance measure Q assigns to a given risk prediction model a value in the decision space: $Q(\hat{r}_{n,t}) = Q(\hat{r}_{n,t}; F, H) \in \mathcal{D}$. Here 'given model' means that given the data $\mathcal{L}_n$ the model has to be completely specified. In this section we consider solely estimates of predictive performance that use the data of an external validation sample $\mathcal{V}_m = \{X_i^v, Y_i^v(t)\}_{i=1,\ldots,M}$ which is drawn from the same distribution as $\mathcal{L}_n$. These estimates are obtained by the 'plug-in' principle, that is by substituting the theoretical expectation by its empirical counterpart. and denoted $\hat{Q}_m(\hat{r}_{n,t}) = Q(\hat{r}_{n,t}; F_m^v, H_m^v)$. Here $F_m^v \times H_m^v$ is the empirical distribution of the data $\mathcal{V}_m$. The data $\mathcal{L}_n$ are only used to fit the prediction model. This situation resembles the split-sample approach to *internal validation* which is further discussed in Section 5.

3.2 ROC methodology

Traditionally the ROC curve and derived measures are used for evaluating the discriminative ability of a single continuous marker. In this article the aim is to evaluate the predicted probabilities that are obtained with multivariable risk prediction models as described in Section 2. Thus, in all considerations that follow, the predicted probabilities are considered as a new 'single continuous marker'. This has to be kept in mind for the interpretation of performance measured with ROC methodology. Another key feature of the ROC curve is that it is invariant to monotone transformations of the marker values and hence works with the ranks of the predicted probabilities.

3.2.1 The ROC curve

The decision space for the ROC curve is the set of functions that map from the false positive rate, FPR $(\hat{r}_{n,t}, \xi) = P(\hat{r}_{n,t}(X_i) \geq \xi \mid Y_i(t) = 0)$, to the true positive rate, TPR $(\hat{r}_{n,t}, \xi) = P(\hat{r}_{n,t}(X_i) \geq \xi \mid Y_i(t) = 1)$. Thus, formally $\mathcal{D} = \{f : [0,1] \mapsto [0,1]\}$. To assess the predictive performance of a model the aim is to estimate the graph ROC $(\hat{r}_{n,t}, \cdot) = \{[\text{FPR}\ (\hat{r}_{n,t}, \xi), \text{TPR}\ (\hat{r}_{n,t}, \xi)] : \xi \in [0,1]\}$. The restriction $\xi \in [0,1]$ refers to that the marker values considered here are predicted probabilities.

The true positive rate can be expressed as a functional of (F, H) as follows

$$\text{TPR}\,(\hat{r}_{n,t}, \xi; F, H) = \frac{P\{\hat{r}_{n,t}(X_i) \geq \xi, Y_i(t) = 1\}}{P(Y_i(t) = 1)} = \frac{\int \mathcal{I}\{\hat{r}_{n,t}(x) \geq \xi\}\, F(t \mid x)\, H(\mathrm{d}x)}{\varrho(t)}\,. \tag{5}$$

Analogously the false positive rate can be written as

$$\text{FPR}\,(\hat{r}_{n,t}, \xi; F, H) = \frac{P\{\hat{r}_{n,t}(X_i) \geq \xi, Y_i(t) = 0\}}{P(Y_i(t) = 0)} = \frac{\int \mathcal{I}\{\hat{r}_{n,t}(x) \geq \xi\}\, S(t \mid x)\, H(\mathrm{d}x)}{1 - \varrho(t)}\,. \tag{6}$$

The plug-in estimates substitute $F_m^v \times H_m^v$ for $F \times H$; they are respectively given by

$$\widehat{\text{TPR}}_m(\hat{r}_{n,t}, \xi) = \text{TPR}\,(\hat{r}_{n,t}, \xi; F_m^v, H_m^v) = \frac{\sum_{i=1}^m \mathcal{I}\{\hat{r}_{n,t}(X_i^v) \geq \xi, Y_i^v(t) = 1\}}{\sum_{i=1}^m \mathcal{I}\{Y_i^v(t) = 1\}} \tag{7}$$

and

$$\widehat{\text{FPR}}_m(\hat{r}_{n,t}, \xi) = \text{FPR}\,(\hat{r}_{n,t}, \xi; F_m^v, H_m^v) = \frac{\sum_{i=1}^m \mathcal{I}\{\hat{r}_{n,t}(X_i^v) \geq \xi, Y_i^v(t) = 0\}}{\sum_{i=1}^m \mathcal{I}\{Y_i^v(t) = 0\}}\,. \tag{8}$$

Unfortunately, the decision space for the ROC curve is not ordered, as ROC curves may cross, and it may be impossible, by solely considering the ROC curves, to distinguish the performance of several risk prediction models. See Hilden (1991, 2005) thorough discussion and examples.

3.2.2 Optimal points

Several "optimal" points have been considered as criteria for choosing between two competing ROC curves. It is well known that there is no unique way of combining in a single index all the properties of a diagnostic test (Greenhouse et al., 1950). However, one of the prominently considered criteria is the Youden Index (Youden, 1950), the point on the ROC curve which has the maximum vertical distance to the positive diagonal:

$$J(\hat{r}_{n,t}; F, H) = \max_{\xi \in [0,1]} \left\{\text{TPR}\,(\hat{r}_{n,t}, \xi; F, H) - \text{FPR}\,(\hat{r}_{n,t}, \xi; F, H)\right\}. \tag{9}$$

Note that the plug-in estimate for J, when it is derived directly from (7) and (8), may not be uniquely defined (Hsieh and Turnbull, 1996). Alternatively one may consider the point on the ROC curve which is closest to the optimal point at which $\text{TPR}\,(\hat{r}_{n,t}, \xi) = 1$ and $\text{FPR}\,(\hat{r}_{n,t}, \xi) = 0$. Perkins and Schisterman (2006) point out that this "optimal" point can be consistent and inconsistent with Youden's index. Other points on the ROC curve that can be used to decide between two models are the symmetry point where $\text{TPR}\,(\hat{r}_{n,t}, \xi) = 1 - \text{FPR}\,(\hat{r}_{n,t}, \xi)$, and the point that corresponds to a preselect value for the false positive rate (Pepe, 2003). For our purposes it is most important that all these criteria can be formulated as functionals of $F \times H$, and hence can estimated with the 'plug-in' principle, e.g.

$$J(\hat{r}_{n,t}; F_m^v, H_m^v) \to J(\hat{r}_{n,t}; F, H) \quad \text{as} \quad m \to \infty\,.$$

3.2.3 The area under the ROC curve

For assessing the performance of predicted risk usually the whole risk spectrum is important and hence all points of the ROC curve. The area under the ROC curve (AUC) is one of the most commonly used summary measures. It is given by

$$\text{AUC}\,(\hat{r}_{n,t}) = P\{\hat{r}_{n,t}(X_1) < \hat{r}_{n,t}(X_2) \mid Y_1(t) = 0, Y_2(t) = 1\} = \int_0^1 \text{ROC}\,(\hat{r}_{n,t}, \xi)\, \mathrm{d}\xi\,.$$

Here $\text{ROC}\,(\xi) = \text{TPR}\,\{\text{FPR}^{-1}(\xi)\}$ and $(Y_i(t), X_i)$, $i = 1, 2$ denote data from two randomly selected patients. The ROC area can be expressed as a functional of the joint distribution $F \times H$ as follows:

$$\text{AUC}\,(\hat{r}_{n,t}; F, H) = \frac{1}{\varrho(t)\{1-\varrho(t)\}} \int\int \mathcal{I}\{\hat{r}_{n,t}(x) < \hat{r}_{n,t}(\tilde{x})\}\, F(t \mid x)\, H(\mathrm{d}x)\, S(t \mid \tilde{x})\, H(\mathrm{d}\tilde{x})\,. \quad (10)$$

The decision space for the AUC is the interval $[0, 1]$ and generally the higher the AUC of a model the better.

One version of the plug-in estimate to Eq. (10) that accounts for ties between the predicted probabilities and uses the external validation set is

$$\text{AUC}\,(\hat{r}_{n,t}; F^v_m, H^v_m) = K_m^{-1} \sum_{i=1}^{m} \sum_{j=1}^{m} (1 - Y^v_i(t))\, Y^v_j(t) \times \left[\frac{1}{2}\, \mathcal{I}\{\hat{r}_{n,t}(X^v_i) = r(X^v_j)\} + \mathcal{I}\{\hat{r}_{n,t}(X^v_i) < \hat{r}_{n,t}(X^v_j)\}\right], \quad (11)$$

where $K_m = m^2 \hat{\varrho}_m(t)\,\{1 - \hat{\varrho}_m(t)\}$ and $\hat{\varrho}_m(t) = n^{-1} \sum_i Y^v_i(t)$ is the empirical prevalence in the validation set at time t.

The statistic (11) is equivalent to the Gini index (Hand and Till, 2001) and to the Wilcoxon test statistic (Hanley and McNeil, 1983). In the situations considered in this article, the AUC can be interpreted as the average ability of the risk prediction model to differentiate the risk of patients over the whole risk spectrum. Being a statistic based on ranks, in a given sample, the AUC can not see a difference when the predicted probability of one patient is increased by a small amount. Table 3 shows this deficit since the data of the two patients are concordant for all models disregarding the considerable differences in the predicted probabilities.

Sometimes it is useful to restrict the area under the ROC curve to a certain region. In our setting one could restrict the attention to the high or the low risk part of the population. The partial area under the curve (McClish, 1989; Pepe, 2003) is defined by

$$p\text{AUC}\,(\hat{r}_{n,t}, \tau) = \int_a^b \text{ROC}\,(\hat{r}_{n,t}, \xi)\, \mathrm{d}\xi\,.$$

3.3 Utility and regret

The measures of predictive performance discussed in Section 3.2 are constructed solely from the true and false positive rates. They do not incorporate the prevalence and hence do not measure utility on a population or patient level. In many applications this is a disadvantage and not a strength of the method (Hilden, 2000; Guggenmoos-Holzmann and van Houwelingen, 2000; Greenland, 2008).

3.3.1 ROC utility measures

To measure the utility of a predictive model one can incorporate the prevalence into the ROC curve and optionally introduce problem-specific weights. See Habbema and Hilden (1981) for a general utility and regret framework for medical decision making. Hilden (1991) shows how the ROC curve and hence its summaries can be modified to incorporate the prevalence and misclassification costs. A further concrete example is the 'generalized Youden Index' considered recently (Schisterman et al., 2008):

$$J^*(\hat{r}_{n,t}, \varkappa_t; F, H) = \max_{\xi \in [0,1]} [\{1 - \varrho(t)\}\,(1 - \text{FPR}\,(\hat{r}_{n,t}, \xi; F, H)) + \varkappa(t)\,\varrho(t)\,\text{TPR}\,(\hat{r}_{n,t}, \xi; F, H)].$$

Here $\varkappa(t)$ is the relative loss of the false negative rate at time t when compared with a false positive rate at that time. The generalized Youden Index corresponds to the expected utility ROC curve defined in Hilden (1991). Further discussion on optimal utility operating points can be found in Cantor and Ganiats (1999) and Obuchowski (2003).

More familiar combinations of TPR, FPR and the prevalence $\varrho(t)$ are the positive predictive value PPV $(\hat{r}_{n,t}, \xi; F, H) = P(Y_i(t) = 1 \mid \hat{r}_{n,t}(X_i) \geq \xi)$ and the negative predictive value, NPV $(\hat{r}_{n,t}, \xi; F, H) = P(Y_i(t) = 0 \mid \hat{r}_{n,t}(X_i) < \xi)$; we refer to Pepe (2003) and Moskowitz and Pepe (2004b, a).

3.3.2 Expected loss

An important class of performance measures scores directly the value of the probability $\hat{r}_{n,t}(X_i)$ as a prediction for the status $Y_i(t)$. A scoring rule is called proper (strictly proper) if it is minimized (uniquely) at the true status probability $F(t \mid X_i)$ (Savage, 1971; Gneiting and Raftery, 2007). An often considered strictly proper scoring rule is the Brier score (Brier, 1950). It is given by the squared distance between the patients observed status and the predicted probability, viz. $\{Y_i(t) - \hat{r}_{n,t}(X_i)\}^2$.

The values of the Brier score can be interpreted as the loss or regret which is incurred when the prediction $\hat{r}_{n,t}(X_i)$ is issued to a patient whose true status is $Y_i(t)$. The expected loss of the risk prediction model in the population is given by

$$\mathrm{BS}\,(\hat{r}_{n,t}; F, H) = E\{Y_i(t) - \hat{r}_{n,t}(X_i)\}^2 = \left[\{1 - \hat{r}_{n,t}(x)\}^2\, F(t \mid x) + \{\hat{r}_{n,t}(x)\}^2\, S(t \mid x)\right] \mathrm{d}H(x)\,. \tag{12}$$

The decision space for the Brier score is the interval $[0, 1]$ and generally the lower the better. The plug-in estimate to (12) based on the external validation set is

$$\mathrm{BS}\,(\hat{r}_{n,t}; F_m^v, H_m^v) = \frac{1}{m} \sum_{i=1}^{m} \{Y_i^v(t) - \hat{r}_{n,t}(X_i^v)\}^2\,. \tag{13}$$

As pointed out in Moskowitz and Pepe (2004b), certain performance measures such as the Brier score do not distinguish between the losses incurred for individuals with positive and negative status. To accommodate such differential losses, one may consider the following modification:

$$Y_i(t)\, \varkappa_1(t)\{1 - \hat{r}_{n,t}(X_i)\}^2 + (1 - Y_i(t))\, \varkappa_0(t)\, \{\hat{r}_{n,t}(X_i)\}^2\,.$$

Here $\varkappa_0(t), \varkappa_1(t)$ are different costs incurred for misspecified individuals with positive and negative status, respectively.

The logarithmic score is closely related to the deviance and therefore a popular tool. It is the strictly proper scoring rule given by $\varkappa_1 \log\{\hat{r}_{n,t}(X_i)\}$ where $\varkappa_1 > 0$ is a cost parameter as above. Unfortunately, the plug-in estimate of the expected logarithmic score is not finite when one of the predicted probabilities is zero. To accommodate this problem Hilden et al. (1978) defined an appropriate ε-modified version of the logarithmic score.

3.4 The added value of a new marker

Today, risk prediction models exists for most of the common diseases. It is thus often demanded to combine markers and to investigate the improvement when a new biomarker is incorporated into an existing risk prediction model (Pencina et al., 2008; Huang et al., 2007). Any of the performance measures described above can be utilized for this purpose: one simply compares the performance of a model that does not use the new marker to the performance of the extended model. However overall performance measures like the AUC and the Brier score can appear quite insensitive if the new marker is effective only for a small part of the population.

Motivated by the insensitivity of the AUC, Pencina et al. (2008) proposed to summarize the incremental predictive value of a new marker based on its potential in reclassification and discrimination. When interest lies in assigning subjects into different risk categories, the incremental value is summarized by a net reclassification improvement, defined as the total gain in reclassifying cases into higher risk categories and controls into lower risk categories. When interest lies in improving discrimination

accuracy, the incremental value is evaluated based on the integrated discrimination improvement which is essentially the difference in the integrated Youden's index. Since the incremental value of a new marker may vary across different subpopulations, Tian et al. (2007) proposed procedures for identifying subgroups that may or may not benefit from the additional marker assessment.

3.5 Explained variation

A general measure of the variation explained by a risk prediction model is obtained by benchmarking the prediction error to the predicted probability of the null model $\hat{\varrho}_{n,t} = n^{-1} \sum_i Y_i(t)$ that ignores all the predictors:

$$R^2(\hat{r}_{n,t}; F, H) = 1 - \frac{\text{BS}\,(\hat{r}_{n,t}; F, H)}{\text{BS}\,(\hat{\varrho}_{n,t}; F, H)}\,.$$

This R^2 measure is not confined to any specific model fitting technique and satisfies also most of the other requirements proposed in the literature, for example by Kvalseth (1985). However, to compare the performance of different models it is enough to consider the Brier score, $\text{BS}\,(\hat{r}_{n,t}; F, H)$, since the denominator in the previous display does not depend on the model.

3.6 Following performance over time

In survival analysis predictions are usually available as functions of time. It is thus natural to also consider the predictive performance as a function of time (Graf and Schumacher, 1995). For example, Segal (2006) uses the time-dependent area under the curve to compare various risk prediction models for the survival status of patients diagnosed with diffuse large-B-cell lymphoma. In Schumacher et al. (2003) the Brier score is followed over time to compare different prognostic breast cancer models. In the framework of this section it is straightforward to define time-dependent versions of other performance measures via the mapping $t \mapsto Q(\hat{r}_{n,t}; F, H)$. A seemingly undiscovered example is the time-dependent generalized Youden Index $t \mapsto J^*(\hat{r}_{n,t}, \varkappa_t)$.

For these extensions it is important to discuss the different concepts of time-dependent true and false positive rates. To incorporate the time domain in evaluating prediction models for event times, various time dependent discrimination measures have been proposed by defining different diseased and diseased-free populations specific to any time t of interest. For example, Heagerty et al. (2000) considered a cumulative time dependent sensitivity and specificity functions with $Y_i(t) = 1$ being the diseased population and $Y_i(t) = 0$ being the disease-free population. Heagerty and Zheng (2005) and Cai et al. (2006) also considered diseased population at time t as those who experience the event at t and disease-free population as those with $Y_i(t) = 0$. Pepe et al. (2008) also extended the positive and negative predictive value functions to be time dependent.

4 Benchmarking

Benchmarking a candidate model is important for the interpretation of the results obtained with any assessment tool. There are different types of benchmarks for the performance of a trained model: theoretical benchmarks, the performance of a null model, the apparent performance of the model in its own data, and the performance of the model in subsamples and permutations of the original data.

4.1 Theoretical benchmarks

It is often possible to structure the decision space by means of theoretical considerations. For example the expected Brier score is a number between 0 and 1 but a useful risk prediction model should not have a value above 25%. This is the value achieved when issuing a predicted probability of 50% to each patient – the worst case scenario for decision making.

Another often used benchmark is the performance of a random number generator. Suppose a uniformly distributed random number is used to predict the probability of all patients, then the expected Brier score is approximately 33%. To see this, consider uniformly distributed random variate $Z \sim U[0,1]$ and use equation (12):

$$\mathrm{BS}\,(Z;F,H) = \int_0^1 \{(1-x)^2\,\varrho(t) + x^2\,\{1-\varrho(t)\}\,\mathrm{d}x\,.$$

Here $\varrho(t) = P(Y_i(t)=1)$ is the true prevalence, and simple calculus leads to a value of 33% for the expected Brier score.

For a random number prediction the ROC curve equals the positive diagonal, the area under the curve is 0.5 and Youden's Index is equal to zero. These benchmark values correspond to any prediction rule whenever the predictor X_i does not provide any information on the status response $Y_i(t)$ i.e. when X_i and $Y_i(t)$ are stochastically independent. In such a situation we have $P(X_i > c \mid Y_i(t)=1) = P(X_i > c \mid Y_i(t)=0) = P(X_i > c)$ yielding $\mathrm{ROC}\,(c) = c$ regardless of what is being predicted.

Now consider a constant prediction rule $\hat{r}_{n,t}(X_i) = \pi$ which assigns the same value π to every patient. Then the ROC-plot consists of the two points, $(0,0)$ and $(1,1)$, independently of the value of π. For a constant prediction rule the expected Brier score can be decomposed into $\varrho(t)\{1-\varrho(t)\} + \{\pi - \varrho(t)\}^2$. Thus, among the constant predictions, the expected Brier score is minimal if π equals the true prevalence. Thus for the Brier score the constant prediction using the estimated prevalence will provide an informative benchmark value while ROC based measures do not differentiate between these various prediction rules.

4.2 The performance of a null model

A useful benchmark is the performance of a null model that ignores all patient specific information and simply predicts the empirical prevalence $\hat{\varrho}_{n,t} = \frac{1}{n}\sum_i Y_i(t)$ to each patient. A useful risk prediction model should perform better than the null model, at least if there is information in the predictors X_i (see Section 3.5). The AUC value equals 50% for the null model as outlined above. For censored data the Kaplan–Meier estimate plays the role of the null model as it ignores all covariate information (see Section 5.2).

4.3 The apparent performance

The predictive performance of a model when estimated in the data that were used to build the model is called apparent or re-substitution performance. It is an *upper bound* for the true predictive performance of the model. The apparent performance is uniformly defined for all the performance measures displayed in (5), (6), (9), (10) and (12) by

$$Q^{\mathrm{App}}(r_t) = Q(\hat{r}_{n,t}; F_n, H_n)\,.$$

Here $F_n \times H_n$ is the empirical distribution of the sample $\mathcal{L}_n$. The true predictive performance of the trained model $\hat{r}_{n,t}$ will in most cases be lower than this benchmark because the predictions for the patients in $\mathcal{L}_n$ depend on the true status response via estimated parameters and other supervised model building steps.

Table 4 shows the apparent AUC and the apparent Brier score for the epo treatment success and the prediction models that are described in Section 2. If we would consider the apparent estimate of the

Table 4 The apparent performance of the prediction models for the epo treatment success.

	$\mathrm{LRM}^{(0)}$	$\mathrm{LRM}^{(1)}$	$\mathrm{LRM}^{(2)}$	$\mathrm{LRM}^{(3)}$	$\mathrm{LRM}^{(4)}$	CART	RF
AUC	50.00	94.66	93.32	93.53	95.43	89.65	99.76
Brier score	24.67	8.69	9.58	8.63	8.28	10.04	3.00

Table 5 The bootstrap crossvalidation performance of the models considered for predicting the epo treatment success.

	$\text{LRM}^{(0)}$	$\text{LRM}^{(1)}$	$\text{LRM}^{(2)}$	$\text{LRM}^{(3)}$	$\text{LRM}^{(4)}$	CART	RF
AUC	50	91.79	91	90.72	90.31	87.58	91.68
Brier score	25.19	11.58	11.85	11.9	13.7	12.52	11.02

Brier score or the AUC as an estimate for the true performance then we could get the wrong impression that the random forest model de-classifies all the other models in terms of predictive performance.

4.4 Bootstrap crossvalidation

Crossvalidation is a central tool for the process of building a predictive model. This is especially so in the machine learning world; it has received less attention among traditional statisticians (Breiman, 2001b). Crossvalidation works by fitting the same model in k different subsets of the original data. The crossvalidation performance is the average predictive performance measured in the data of the patients that were left out in the k-th run. An effective version of this procedure is bootstrap-crossvalidation where the model is fitted in B different bootstrap subsamples $\mathcal{L}^*_{n,1}, \ldots, \mathcal{L}^*_{n,B}$ drawn from the data $\mathcal{L}_n$ either with or without replacement. The samples not in the b-th bootstrap sample are used to validate the bth fit. The bootstrap crossvalidation performance provides a *lower bound* for the true predictive performance because the data used for training the model generally includes less information than the full data. Let $\hat{r}^*_{b,t}$ denote the fit of the model r_t in the sample $\mathcal{L}^*_{n,b}$ and let F^*_b, H^*_b represent the samples not in the set $\mathcal{L}^*_{n,b}$. In the framework of Section 3 the boostrap-crossvalidation benchmark for the model trained in the full data $\hat{r}_{n,t}$ is given by

$$Q^{\text{Boot CV}}(r_t) = \frac{1}{B}\sum_{b=1}^{B} Q(\hat{r}^*_{b,t}; F^*_b, H^*_b)\,.$$

This definition applies uniformly to the performance measures in (5), (6), (9), (10) and (12).

Table 5 shows the bootstrap crossvalidated benchmark values of ROC area and the Brier score for the epo treatment success and the prediction models that are described in Section 2.

4.5 The no-information value

The idea behind the no-information benchmark is to measure a model's potential to overfit. This is done by training the model in artificially modified data where the response variable is independent of the predictors. This independence can be achieved by systematic or random permutation of the the status response values, at the same time holding fixed the predictor matrix. Efron and Tibshirani (1997) motivated their estimate of the noinformation benchmark for the misclassification rate. Wehberg and Schumacher (2004); Gerds and Schumacher (2007) studied a generalization of this estimate for the Brier score. Recently, Adler and Lausen (2007) defined the noinformation sensitivity and specificity.

To define a uniformly applicable estimate, let $\tilde{\mathcal{L}}_{n,1}, \ldots, \tilde{\mathcal{L}}_{n,B}$ arise from $\mathcal{L}_n$ by permutation of the status response, that is in each set $\tilde{\mathcal{L}}_{n,b}$ the predictor information of patient i is reallocated to the status response value of a different patient $j \neq i$. Let $\tilde{F}_b, \tilde{H}_b$ represent the sample $\tilde{\mathcal{L}}_{n,b}$. The no-information benchmark for the model $\hat{r}_{n,t}$ is given by

$$Q^{\text{NoInf}}(r_t) = \frac{1}{B}\sum_{b=1}^{B} Q(\hat{r}_{n,t}; \tilde{F}_b, \tilde{H}_b)\,.$$

Table 6 The no-information performance of the prediction models for the epo treatment success obtained in 1000 versions of the original data where the response was randomly reallocated and thus the predictor variables are stochastically independent of the response.

	$LRM^{(0)}$	$LRM^{(1)}$	$LRM^{(2)}$	$LRM^{(3)}$	$LRM^{(4)}$	CART	RF
AUC	50	49.96	50.02	49.8	50.11	49.96	49.92
Brier score	24.67	40.32	39.45	39.97	40.74	39.32	41.3

Table 6 shows the average values of the AUC and the Brier score over 1000 of such no-information data sets. The prediction models are described in Section 2. The results show that the AUC always takes a value close to 0.5 as explained in Section 4.2.

The Brier score is able to differentiate between the various prediction rules. Here, the no-information value can be considered as some kind of 'worst-case' scenario in which predictors do not provide any information on status response but they are used for prediction. Table 6 shows that the models $LRM^{(4)}$ and RF have the largest noinformation error but are closely followed by CART and the other regression-based prediction rules. Clearly, the null model attaching the estimated prevalence (0.443) as prediction to every patient is not affected.

5 Estimation

There are two major sources of bias when estimating the true performance, also called the generalization error, of a status prediction model. Bias is introduced by censored data (see Section 5.2) and when the predictions depend on the response status (see Section 5.1).

5.1 Internal validation

For example all the predicted probabilities for the future status of two sample patients reported in Table 3 depend on the actually observed value of these patients because their data was part of the data used for building the models. Boostrap-crossvalidation efficiently removes this dependency. However, the bootstrap-crossvalidation benchmark suffers from a different bias, as noted already in the previous section. It is pessimistic since the model is build on less information than the full data provide. The 0.632+ formula as proposed in Efron and Tibshirani (1997) for the misclassification error has been extended for the Brier score and the AUC, e.g. in Steyerberg et al. (2001, 2003); Wehberg and Schumacher (2004); Gerds and Schumacher (2007); Jiang and Simon (2007).

The idea of the 0.632+ method is that the true performance of a model is trapped between its apparent performance (upper bound) and its bootstrap-crossvalidated performance (lower bound). The Efron and Tibshirani (1997) proposed a linear combination of these two benchmarks yielding the 0.632+ estimate:

$$Q^{0.632+}(r_t) = (1 - \omega(r_t))\, Q^{\text{App}}(r_t) + \omega(r_t)\, Q^{\text{BootCV}}(r_t)\,. \tag{14}$$

The weights are defined by $\omega(r_t) = 0.632/1 - 0.368\hat{R}(r_t)$ where the relative overfitting rate (Efron and Tibshirani, 1997) is given by

$$\hat{R}(r_t) = \frac{Q^{\text{BootCV}}(r_t) - Q^{\text{App}}(r_t)}{Q^{\text{NoInf}}(r_t) - Q^{\text{App}}(r_t)}\,.$$

It is not straightforward to define a 0.632+ version of the ROC curve. One possibility is to consider fixed values for the false positive rate. One can use a grid e.g. of 100 or 1000 equidistant values between 0 and 1 and compute, for each of these false positive rates, a point wise 0.632+ estimate of

Table 7 The 0.632+ performance of the prediction models for the epo treatment success. The values are linear combinations based on the respective values in Tables 4, 5 and 6 and by using the formula (14).

	$LRM^{(0)}$	$LRM^{(1)}$	$LRM^{(2)}$	$LRM^{(3)}$	$LRM^{(4)}$	CART	RF
AUC	50	92.67	91.61	91.48	91.95	87.2	94.11
Brier score	25.19	10.58	11.06	10.78	11.93	11.66	8.49

the corresponding true positive rate. In the same way one could start by fixing true positive rates and estimate the corresponding 0.632+ false positive rates. These averages of ROC curves have been called vertical and horizontal, respectively, in Fawcett (2004). They may be criticized in situations where neither the false nor the true positive rate can be influenced. The vertical 0.632+ ROC curve is given by:

$$\mathrm{ROC}^{0.632+}(\hat{r}_{n,t};\cdot) = [\{\mathrm{TPR}^{0.632+}(\hat{r}_{n,t};\xi),\xi\};\xi \in \{0,0.01,\ldots,0.99,1\}]\,.$$

Figure 3 shows the 0.632+ ROC curve for $LRM^{(4)}$ obtained by vertical averaging of 1000 bootstrap-crossvalidation runs and the corresponding benchmark curves. The apparent ROC curve is clearly too high in some regions when compared to the more honest 0.632+ estimate of the ROC curve.

5.2 Censored data

In the typical applications of survival analysis the future status of some patients is right censored. Estimated performance, like any other parameter estimate in survival analysis, should not depend on censoring. In particular, all the plug-in estimates discussed so far in this article cannot be applied to censored data. To deal with censoring Korn and Simon (1990) proposed a model-based approach which essentially relies on a correctly specified model. However, when the purpose is to compare the performance of different, potentially misspecified, models it is preferable to use an approach that does

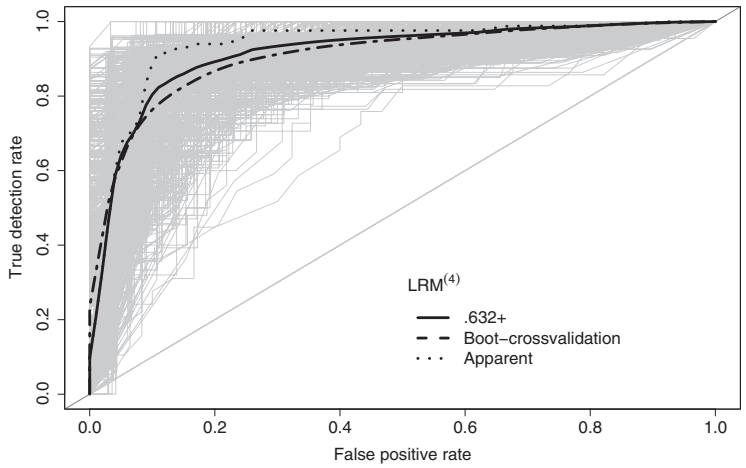

Figure 3 Vertical 0.632+ estimates of the ROC curve for $LRM^{(4)}$ (solid line) is a point wise linear combination of the apparent ROC curve (dotted line) and the boostrap-crossvalidation ROC curve (dashed line). The 1000 ROC curves obtained in the process of bootstrap-crossvalidation are given as a grey shadow.

not depend on the models which are compared. Such 'model-free' methods have been described in Heagerty et al. (2000); Uno et al. (2007) for ROC analysis and in Graf et al. (1999); Gerds and Schumacher (2006) for the Brier score. We refer to these articles for details of computation.

From a high-level viewpoint it is possible to explain the principle of how most of these methods deal with censoring. For this we use that the performance measures in (5), (6), (9), (10) and (12) are defined as functions of F and H. With censored event status one derives an estimate $\hat{F}_n(t \mid X_i)$ of the conditional distribution $F(t \mid X_i)$ from an estimate of the conditional survival function, and then substitutes this together with the empirical distribution H_n of the predictors. Then, if $\hat{F}_n \to F$ *as* $n \to \infty$, it is often possible to show that the plug-in estimate $Q(\hat{r}_{n,t}; \hat{F}_n, H_n)$ is asymptotically independent of the censoring mechanism. However, it is not clear how to estimate the conditional distribution F in practice and there is a tradeoff between bias and the curse-of-dimensionality (Van der Laan and Robins, 2003). The problem can be described as follows: If a parametric or semi-parametric model is used to estimate $F(t \mid X_i)$ then the plug-in estimate $Q(\hat{r}_{n,t}; \hat{F}_n, H_n)$ for the performance measure $Q(\hat{r}_{n,t}; F, H)$ will be biased as soon as the model is misspecified. On the other hand, using a nonparametric estimate for the conditional survival function F will remove the bias, but it requires that the predictor space is low dimensional. A class of nonparametric esitmates for the conditional survival function is studied in Dabrowska (1987), see also Akritas (1994). A typical semi-parametric estimate for $F(t \mid X_i)$ can be derived from a Cox regression model. In order to reduce the potential bias of the latter Heagerty and Zheng (2005) proposed to use a regression spline for the functional form of covariate link. However, the large sample performance of such a flexible approach is not clear. Heagerty et al. (2000) used a nonparametric estimate to estimate the time-dependent ROC curve and they used a nonparametric estimate of the survival function motivated by results of Akritas (1994). Here the dimensionality problem is avoided by restricting the attention to a one-dimensional marker.

Instead of estimating the conditional survival function $S = 1 - F$ directly one can also use an inverse of the probability of censoring weighted (IPCW) estimate. This was studied in Graf et al. (1999); Uno et al. (2007); Gerds and Schumacher (2006) for the performance measures discussed in this article. The IPCW approach requires an estimate of the conditional censoring survival function. For individual right censoring times C_i denote $G(t \mid X_i) = P(C_i > t \mid X_i)$ for the conditional censoring survival function. The process $Y_i(t)$ is only observed in the interval $[0, C_i]$ and we may define $T_i = \min\{t : Y_i(t) = 1\}$ as the event time and $\tilde{T}_i = \min(T_i, C_i)$ as the minimum of censoring time and event time. Under the assumption that the censoring times are conditionally independent of the event times given the predictors X_i, the IPCW estimate is motivated by the relation

$$S(t \mid X_i) = \frac{G(t \mid X_i)}{G(t \mid X_i)} S(t \mid X_i) = \frac{P(\tilde{T}_i > t \mid X_i)}{G(t \mid X_i)} . \tag{15}$$

Substituting an estimate $\hat{G}_n(t \mid X_i)$ for $G(t \mid X_i)$ in (15) yields an estimate for $S(t \mid X_i)$ (see Gerds and Schumacher (2006) for details) and via $F = 1 - S$ also for F which then can be used to estimate any of the performance measures (5), (6), (9), (10), (12). From a theoretical viewpoint estimation of a conditional censoring distribution is as hard as estimation of a conditional survival distribution. Thus the IPCW approach suffers from the same tradeoff between bias and curse-of-dimensionality as the direct method discussed above. There is however an important special case: If it is reasonable to assume that the censoring distribution is independent of the predictors, then one can use the Kaplan–Meier estimator for the censoring distribution in (15). This approach was proposed in Graf et al. (1999). However, as pointed out in Van der Laan and Robins (2003) this approach is inefficient. The estimate will only be fully efficient if a nonparametric estimate is used for $G(t \mid X_i)$ which again is restricted to low dimensional predictor spaces. Gerds and Schumacher (2006) compares the marginal Kaplan–Meier weights to weights based on semi-parametric and nonparametric models for the conditional censoring survival function $G(t \mid X_i)$.

In Figure 4 we have used a Cox regression model to obtain the IPCW weights. The figure compares different prediction models by means of their 0.632+ estimate of the expected Brier score

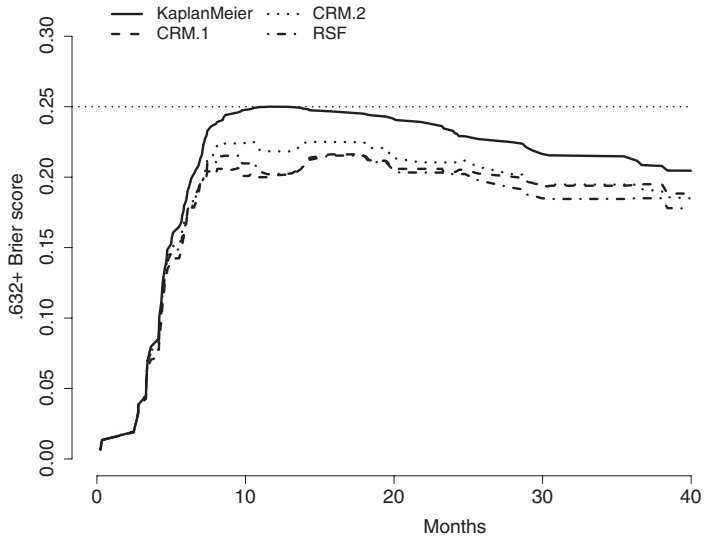

Figure 4 0.632+ estimates of the expected Brier score. An IPCW weighting scheme is obtained for all models with a Cox regression model for the censoring times as outlined in Gerds and Schumacher (2006, 2007).

(Gerds and Schumacher, 2007). A similar result is obtained when the time-dependent area under the curve is considered for the four different prediction models.

5.3 Confidence

As noted in Section 4.3 severe biases are associated with the re-substitution or apparent performance estimators, especially when the prediction model of interest has high complexity. When constructing confidence intervals, it is important to correct for such biases to ensure the proper coverage level. Recently, Tian et al. (2007) and Uno et al. (2007) showed that for certain prediction models, the crossvalidated estimators of certain performance measures, such as the ROC curve, are asymptotically equivalent to the apparent performance estimators, at the order of root-n. Therefore, to minimize computational burden, one may construct confidence intervals by centering at the crossvalidated estimators with width determined based on the variability of the apparent performance estimators. Furthermore, for standard regression models, the bootstrap method could potentially be used to assess the variability of the estimated performance measures. However, when the prediction models are constructed based on flexible procedures such as the random forest, it is unclear whether such confidence limits remain to be valid.

6 Summary and Discussion

New biomarkers are important for the development of medical treatment options. They are also highly relevant for improving the information given to patients. However, rarely a single marker alone provides enough information for reliable decision making. Multiple markers have to be integrated into existing models. In many applications both the conventional model and the new model provide predicted probabilities that can be compared to assess the ability of a new marker. The ROC curve and the Brier score are both tools that can be used for this purpose. Both methods are developed to deal with censored data, and the area under the ROC curve and the Brier score can be followed longitudinally in time for applications in survival analysis.

Often one aims at comparing non-nested prediction models with different complexity. A desirable performance measure should not rely on a specific form of the likelihood function and it should be applicable to all statistical cultures. For example it can be useful to compare a prognostic model that has been derived from logistic regression with another one that has been derived with the random forest (Breiman, 2001b). For example, in a recent study Bartfay et al. (2006) used the Brier score and the AUC to compare neural network models to logistic regression models. In a Bayesian framework, one could think of competing scientists that choose different modelling strategies and the aim is to elicit the best forecast. Indeed, proper scoring rules are a fundamental concept of Bayesian inference (Bernardo and Smith, 2000). The use of a strictly proper score function guarantees that optimal predictions are elicited (Savage, 1971; Matheson and Winkler, 1976; Gneiting and Raftery, 2007). In this article we have exclusively considered to evaluate directly the predicted risk and not used risk classes. This is also in agreement with Bayesian theory where it is natural to consider probabilistic forecasts of future events. However, in some applications it is useful to classify risk into categories. This can be in order, for example, when risk classes directly correspond to treatment options, or when they are otherwise clinically meaningful. Whatsoever, it is important to keep in mind that categorization of risk generally implies loss of information.

Crossvalidation and bootstrap are important tools that can be used to make the performance of different models comparable. They are the heart of effective internal validation procedures. For the step into routine clinical use, however, often external validation is required where a given prediction model is tested in data from a similar study. For example Greene et al. (2004) validated the nomogram developed by Kattan et al. (2000) in the data from their study.

When developing a risk model with a limited amount of data, the 0.632+ approach allows an unbiased assessment of its predictive performance with the advantage that the information available can efficiently be used and unnecessary data splitting can be avoided. We like to emphasize that the estimated prediction error will only track the true prediction error if all steps of the model building process are repeated within each bootstrap or cross validation sample. In the epo study the problem of not doing so is nicely illustrated by the performance of model $\text{LRM}^{(4)}$ where an 'optimal' cut-point for the variable age is determined that gives the minimal p-value in the respective logistic regression model. This data-driven procedure leads to the best prediction performance among the four regression models in terms of apparent error. However, the prediction error estimated by bootstrap crossvalidation or by the 0.632+ approach is highest among the four models showing that the model should not be used for new patients.

The adequate and efficient use of information in developing risk prediction models as well as an unbiased evaluation of their prediction performance is especially important in high-dimensional ($K \gg n$) data settings, e.g. in micro-array data; e.g. Ruschhaupt et al. (2004); Markowetz and Spang (2005); Simon (2005b). We have recently shown that the 0.632+ performance evaluation can also be successfully used in such a setting (Schumacher et al., 2007), however, some care has to be taken when in a high-dimensional setting resampling is used both for data-driven determination of model complexity and for the assessment of performance (Binder and Schumacher, 2008).

In summing up, we have developed a unifying framework for evaluating the performance of risk prediction models that are of growing importance in biomedical and other application. We have exemplified and illustrated the some of the most popular and currently available approaches by means of the data of the epo study (Henke et al., 2006) where we have studied epo treatment success and local disease free survival as two different status response variables. For the latter it would also be possible to include time-dependent information in the risk prediction model; for a recent concept of quantifying the prediction performance in such a situation we refer to Schoop et al. (2008). We have shown that various performance measures will fit into this general framework the most prominent ones being the Brier score and ROC-based quantities. At this stage we cannot give a general recommendation of which one to prefer since this would depend on context, e.g. type of application, aim of study etc and on their theoretical properties. We hope, however, that our review article will help to clarify issues and highlight differences in order to define the most appropriate strategy for evaluating the performance of risk prediction models in a specific situation.

Acknowledgements *We owe thanks to Prof. Michael Henke for introducing us to the data of the Epo study.*

Conflict of Interests Statement
The authors have declared no conflict of interest.

References

Adler, W. and Lausen, B. (2007). Bootstrap estimated sensitivity, specificity and ROC curve. *Tech. rep.*, Department of Biometry and Epidemiology, University Erlangen-Nuremberg.

Akritas, M. G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *The Annals of Statistics* **22**, 1299–1327.

Bartfay, E., Mackillop, W. J., and Pater, J. L. (2006). Comparing the predictive value of neural network models to logistic regression models on the risk of death for small-cell lung cancer patients. *European Journal of Cancer Care* **15**, 115–124.

Bernardo, J. M. and Smith, A. F. (2000). *Bayesian theory.* Wiley Series in Probability and Statistics. Chichester: Wiley.

Binder, H. and Schumacher, M. (2008). Adapting prediction error estimates for biased complexity selection in high-dimensional bootstrap samples. *Statistical Applications in Genetics and Molecular Biology* **7**, Article 12.

Breiman, L. (2001a). Random forests. *Machine Learning* **45**, 5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures. (With comments and a rejoinder). *Statistical Sciences* **16**, 199–231.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees.*. The Wadsworth Statistics/Probability Series. Belmont, California.

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

Cai, T., Pepe, M. S., Zheng, Y., Lumley, T., and Jenny, N. S. (2006). The sensitivity and specificity of markers for event times. *Biostatistics* **7**, 182–197.

Cantor, S. B. and Ganiats, T. G. (1999). Incremental cost-effectiveness analysis: the optimal strategy depends on the strategy set. *Journal of Clinical Epidemiology* **52**, 517–522.

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. (With discussions and a rejoinder by the authors.) *Journal of the American Statistical Association* **98**, 900–945.

Dabrowska, D. M. (1987). Non-parametric regression with censored survival time data. *Scandinavian Journal of Statistics* **14**, 181–197.

Efron, B. and Tibshirani, R. (1997). Improvement on cross-validation: The 0.632+ bootstrap method. *Journal of the American Statistical Association* **92**, 548–560.

Fawcett, T. (2004). Roc graphs: Notes and practical considerations for researchers. *Tech. rep.*, HP Laboratories, Palo Alto, USA.

Gail, M. H. and Pfeiffer, R. M. (2005). On criteria for evaluating models of absolute risk. *Biostatistics* **6**, 227–239.

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

Gerds, T. A. and Schumacher, M. (2007). On Efron type measures of prediction error for survival analysis. *Biometrics* **63**, 1283–1287.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–378.

Graf, E., Schmoor, C., Sauerbrei, W. F., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* **18**, 2529–2545.

Graf, E. and Schumacher, M. (1995). An investigation on measures of explained variation in survival analysis. *The Statistician* **44**, 497–507.

Greene, K. L., Meng, M. V., Elkin, E. P., Cooperberg, M. R., Pasta, D. J., Kattan, M. W., Wallace, K., and Carroll, P. R. (2004). Validation of the Kattan preoperative nomogram for prostate cancer recurrence using a community based cohort: results from cancer of the prostate strategic urological research endeavor (capsure). *Journal of Urology* **171**, 2255–2259.

Greenhouse, S. W., Cornfield, J., and Homburger, F. (1950). The Youden index: letters to the editor. *Cancer* **3**, 1097–1101.

Greenland, S. (2008). The need for reorientation toward cost-effective prediction: Comments on 'evaluating the added predictive ability of a new marker: From area under the roc curve to reclassification and beyond' by m.j. pencina et al., statistics in medicine. *Statistics in Medicine* **27**, 199–206.

Guggenmoos-Holzmann, I. and van Houwelingen, H. C. (2000). The (in)validity of sensitivity and specificity. *Statistics in Medicine* **19**, 1783–1792.

Habbema, J. D. and Hilden, J. (1981). The measurement of performance in probabilistic diagnosis. IV. Utility considerations in therapeutics and prognostics. *Methods for Information in Medicine* **20**, 80–96.

Hand, D. (2001). Measuring diagnostic accuracy of statistical prediction rules. *Statistica Neerlandica* **55**, 3–16.

Hand, D. J. (1997). *Construction and Assessment of Classification Rules*. John Wiley, Chichester.

Hand, D J. and Till, R. J. (2001). A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning* **45**, 171–186.

Hanley, J. A. and McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **148**, 839–843.

Harrell, F. E. (2001). *Regression modeling strategies. With applications to linear models, logistic regression and survival analysis.*. Springer Series in Statistics. New York, NY: Springer.

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337–344.

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

Henke, M., Laszig, R., Rube, C., Schafer, U., Haase, K. D., Schilcher, B., Mose, S., Beer, K. T., Burger, U., Dougherty, C., and Frommhold, H. (2003). Erythropoietin to treat head and neck cancer patients with anaemia undergoing radiotherapy: randomised, double-blind, placebo-controlled trial. *Lancet* **362**, 1255–1260.

Henke, M., Mattern, D., Pepe, M., Bëzay, C., Weissenberger, C., Werner, M., and Pajonk, F. (2006). Do erythropoietin receptors on cancer cells explain unexpected clinical findings? *Journal of Clinical Oncology* **24**, 4708–4713.

Hilden, J. (1991). The area under the ROC curve and its competitors. *Medical Decision Making* **11**, 95–101.

Hilden, J. (2000). Prevalence-free utility-respecting summary indices of diagnostic power do not exist. *Statistics in Medicine* **19**, 431–440.

Hilden, J. (2005). What properties should an overall measure of test performance possess? *Clinical Chemistry* **51**, 471; author reply 471–472.

Hilden, J., Habbema, J. D. F., and Bjerregaard, B. (1978). The measurement of performance in probabilistic diagnosis — III. Methods based on continuous functions of the diagnostic probabilities. *Methods of Information in Medicine* **17**, 238–246.

Holländer, N. and Schumacher, M. (2006). Estimating the functional form of a continuous covariate's effect on survival time. *Computational Statistics and Data Analysis* **50**, 1131–1151.

Hsieh, F. and Turnbull, B. W. (1996). Nonparametric methods for evaluating diagnostic tests. *Statistica Sinica* **6**, 47–62.

Huang, Y., Pepe, M. S., and Feng, Z. (2007). Evaluating the predictiveness of a continuous marker. *Biometrics* **63**, 1181–1188.

IPI (1993). A predictive model for aggressive non-Hodgkin's lymphoma. The International Non-Hodgkin's Lymphoma Prognostic Factors Project. *New England Journal of Medicine* **329**, 987–994.

Jiang, W. and Simon, R. (2007). A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. *Statistics in Medicine* **26**, 5320–5334.

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

Kattan, M. W. (2003). Judging new markers by their ability to improve predictive accuracy. *Journal of the National Cancer Institute* **95**, 634–635.

Kattan, M. W., Zelefsky, M. J., Kupelian, P. A., Scardino, P. T., Fuks, Z., and Leibel, S. A. (2000). Pretreatment nomogram for predicting the outcome of three-dimensional conformal radiotherapy in prostate cancer. *Journal of Clinical Oncology* **18**, 3352–3359.

Korn, E. L. and Simon, R. (1990). Measures of explained variation for survival data. *Statistics in Medicine* **9**, 487–503.

Kvalseth, T. O. (1985). Cautionary note about R^2. *The American Statistician* **39**, 279–285.

Lausen, B. and Schumacher, M. (1992). Maximally selected rank statistics. *Biometrics* **48**, 73–85.

Markowetz, F. and Spang, R. (2005). Molecular diagnosis. Classification, model selection and performance evaluation. *Methods for Information in Medicine* **44**, 438–443.

Matheson, J. and Winkler (1976). Scoring rules for continuous probability distributions. *Management Science* **22**, 1087–1096.

McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making* **9**, 190–195.

Miller, R. and Siegmund, D. (1982). Maximally selected chi square statistics. *Biometrics* **38**, 1011–1016.

Moskowitz, C. S. and Pepe, M. S. (2004a). Quantifying and comparing the accuracy of binary biomarkers when predicting a failure time outcome. *Statistics in Medicine* **23**, 1555–1570.

Moskowitz, C. S. and Pepe, M. S. (2004b). Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. *Biostatistics* **5**, 113–127.

Obuchowski, N. A. (2003). Receiver operating characteristic curves and their use in radiology. *Radiology* **229**, 3–8.

Pencina, M. J., Agostino, R. B. S. D., Agostino, R. B. J. D., and Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine* **27**, 157–172.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press.

Pepe, M. S., Zheng, Y., and Jin, Y. (2008). Evaluating the roc performance of markers for future events. *Lifetime Data Analysis* **14**, 86–113.

Perkins, N. J. and Schisterman, E. F. (2006). The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology* **163**, 670–675.

Redelmeier, D., Bloch, D., and Hickam, D. (1991). Assessing predictive accuracy: how to compare Brier scores. *Journal of Clinical Epidemiology* **44**, 1141–1146.

Ruschhaupt, M., Huber, W., Poustka, A., and Mansmann, U. (2004). A compendium to ensure computational reproducibility in high-dimensional classification tasks. *Statistical Applications in Genetics and Molecular Biology* **3**, Article 37.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* **66**, 783–801.

Schisterman, E. F., Faraggi, D., Reiser, B., and Hu, J. (2008). Youden Index and the optimal threshold for markers with mass at zero. *Statistics in Medicine* **27**, 297–315.

Schoop, R., Graf, E., and Schumacher, M. (2008). Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates. *Biometrics* **64**, 603–610.

Schumacher, M., Binder, H., and Gerds, T. (2007). Assessment of survival prediction models based on microarray data. *Bioinformatics* **23**, 1768–1774.

Schumacher, M., Graf, E., and Gerds, T.A. (2003). How to assess prognostic models for survival data: a case study in oncology. *Methods for Information in Medicine* **42**, 564–571.

Schumacher, M., Holländer, N., Schwarzer, G., and Sauerbrei, W. (2006). Prognostic factor studies. In: J. Crowley and D. Pauler Ankerst (eds.), *Handbook of Statistics in Clinical Oncology.* Second Edition, Chapman & Hall, 289–333.

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

Shapiro, D. E. (1999). The interpretation of diagnostic tests. *Statistical Methods in Medical Research* **8**, 113–134.

Simon, R. (2005a). Development and validation of therapeutically relevant multi-gene biomarker classifiers. *Journal of the National Cancer Institute* **97**, 866–867.

Simon, R. (2005b). Roadmap for developing and validating therapeutically relevant genomic classifiers. *Journal of Clinical Oncology* **23**, 7332–7341.

Steyerberg, E. W., Bleeker, S. E., Moll, H. A., Grobbee, D. E., and Moons, K. G. M. (2003). Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *Journal of Clinical Epidemiology* **56**, 441–447.

Steyerberg, E. W., Harrell, F. E. J., Borsboom, G. J., Eijkemans, M. J., Vergouwe, Y., and Habbema, J. D. (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology* **54**, 774–781.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika* **73**, 363–369.

Thompson, I. M., Ankerst, D. P., Chi, C., Goodman, P. J., Tangen, C. M., Lucia, M. S., Feng, Z., Parnes, H. L., and Coltman, C. A. J. (2006). Assessing prostate cancer risk: results from the Prostate Cancer Prevention Trial. *Journal of the National Cancer Institute* **98**, 529–534.

Tian, L., Cai, T., and Wei, L. J. (2007). Identifying patients who need additional biomarkers for better prediction of health outcome or diagnosis of clinical phenotype. *Tech. rep.*, Harvard University Biostatistics Working Paper Series.

Uno, H., Cai, T., Tian, L., and Wei, L. (2007). Evaluating prediction rules for *t*-year survivors with censored regression models. *Journal of the American Statistical Association* **102**, 527–537.

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

Ware, J. H. (2006). The limitations of risk factors as prognostic tools. *New England Journal of Medicine* **355**, 2615–2617.

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

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer* **3**, 32–35.

