University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2005 *Paper* 174

Survival Ensembles

Torsten Hothorn* Peter Buhlmann† Sandrine Dudoit‡
Annette M. Molinaro** Mark J. van der Laan††

*Institut fur Medizininformatik, Biometrie und Epidemiologie, Friedrich-Alexander-Universitat Erlangen-Nurnberg, Germany, torsten@hothorn.de

†Swiss Federal Institute of Technology (ETH) Zurich, Zurich, Switzerland, buhlmann@stat.math.ethz.ch

‡Division of Biostatistics, School of Public Health, University of California, Berkeley, sandrine@stat.berkeley.edu

**Division of Cancer Epidemiology & Genetics, National Cancer Institute, NIH, Bethesda, MD , annette.molinaro@yale.edu

††Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper174

Survival Ensembles

Torsten Hothorn, Peter Buhlmann, Sandrine Dudoit, Annette M. Molinaro, and Mark J. van der Laan

Abstract

We propose a unified and flexible framework for ensemble learning in the presence of censoring. For right-censored data, we introduce a random forest algorithm and a generic gradient boosting algorithm for the construction of prognostic models. The methodology is utilized for predicting the survival time of patients suffering from acute myeloid leukemia based on clinical and genetic covariates. Furthermore, we compare the diagnostic capabilities of the proposed censored data random forest and boosting methods applied to the recurrence free survival time of node positive breast cancer patients with previously published findings.

1 Introduction

In survival time studies, models regressing the time to event on a set of covariates, i.e., variables expected to be associated with the disease the patient suffers from, are the basis of prognostic and diagnostic modeling. The specification and estimation of such models are complicated by the fact that often only incomplete information about the response variable is available due to censoring. The most widely used representative of regression methods for censored data is the Cox model (Cox, 1972), which addresses the censoring problem by maximizing the partial likelihood while leaving the baseline hazard unspecified under the proportional hazards assumption. In order to motivate the methodology proposed in this paper, it is helpful to classify existing approaches as addressing one of the following problems.

The establishment of a close connection between regression models for uncensored continuous response variables and models designed for censored data was motivated by the problem that the Cox model does not reduce to an ordinary linear regression model in the absence of censoring. Accelerated failure time models (e.g. James, 1998) or the Buckley-James model (Buckley and James, 1979) do have this desirable property.

Many authors proposed flexible alternatives to the Cox model without assuming proportional hazards, such as (partially) nonlinear accelerated failure time models (Stute, 1999; Orbe et al., 2003), spline based extensions (Gray, 1992; Kooperberg et al., 1996; LeBlanc and Crowley, 1999), fractional polynomials (Sauerbrei and Royston, 1999) and neural networks (Ripley et al., 2004).

Current research efforts have focused on data analysis problems with high-dimensional covariate spaces, mainly driven by the requirements of biological applications such as microarray gene expression profiling. In high-dimensional situations, Hastie and Tibshirani (2004) suggest a computationally efficient form of regularization applicable to a wide class of linear models including the Cox model and Huang and Harrington (2005) investigate iterative par-

tial least squares fitting in accelerated failure time models. In contrast, dimension reduction techniques are studied by Li and Li (2004) and Bair and Tibshirani (2004) who advocate the application of low-dimensional compound covariates obtained from an unsupervised clustering of the covariates.

The last but at least equally important research problem is concerned with model selection and evaluation. While classical techniques like residual analysis (e.g. Therneau and Grambsch, 2000) and the detection of influential observations (Bedrick et al., 2002) have been translated into the context of survival analysis, specialized goodness of prediction measures, such as the Brier score for censored data (Graf et al., 1999), are a matter of debate (Henderson, 1995; Altman and Royston, 2000; Schemper, 2003). Although censoring induces non-trivial problems for the comparison of observed and predicted response, such measures are important for cross-validation and other resampling-based model evaluation techniques (Sauerbrei, 1999; Dudoit and van der Laan, 2003; Hothorn et al., 2005).

In this paper, we address the four aforementioned problems simultaneously, by applying the general estimation framework described in van der Laan and Robins (2003) to generalize ensemble learning techniques to censored data problems. The framework allows for the specification of regression models under complete information ('full data world') for arbitrary loss functions. For the estimation of the models under incomplete information ('observed data world') a special weighting scheme ensures that observations likely to be censored are up-weighted compared to the observations of patients likely to suffer an event. As a consequence, in the absence of censoring the models reduce to their counterparts known from the uncensored situation. Most importantly, the goodness of prediction of such models is easily evaluated using cross-validation techniques based on well known loss functions like quadratic, or absolute loss (Keleş et al., 2004). The general estimation framework has recently been applied to problems in longitudinal marginal structural models (Bryan et al., 2004), to the construction of survival trees (Molinaro et al., 2004) and other estimation problems (see Sinisi

and van der Laan, 2004; van der Laan et al., 2004).

Ensemble methods like bagging, random forest and boosting (for a general overview we refer to Bühlmann, 2004a, and references therein) yield flexible predictors for nominal and continuous responses and are known to remain stable in high-dimensional settings. Here, we extend the area of application of ensemble methods to survival analysis. We incorporate weights into random forest like algorithms and extend gradient boosting in order to minimize a weighted form of the empirical risk. The published attempts to use ensemble techniques for modeling censored data are rather limited due to the difficulties induced by censoring. Ridgeway (1999) proposed a boosting algorithm minimizing the partial likelihood and Benner (2002) derived a boosting algorithm from the Brier score for censored data. A special aggregation scheme for bagging survival trees was studied by Hothorn et al. (2004). Breiman (2002) introduced a software implementation of a random forest variant for censored data, however without a formal description of the methodology being available.

Following the road map of van der Laan and Robins (2003), Section 2 defines the regression models and the corresponding risk optimization problems in the full data world and sketches the general estimating framework in the observed data world. In Section 3 we propose both a random forest and a boosting algorithm for censored data. The advantages of our approaches are studied with respect to the stability and flexibility of prognosis and predictions for patients suffering from acute myeloid leukemia, based on high-dimensional covariates from gene expression profiling experiments and clinical data. Moreover, we focus on the diagnostic capabilities of flexible ensemble methods for data from node positive breast cancer patients.

2 Model

The estimation problems to be solved are first defined in the full data world and are then mapped into the observed data world, i.e., in the presence of censoring, following van der Laan and Robins (2003).

2.1 Full Data World

In an ideal world, we are able to observe random variables $\mathbf{Z} = (Y = \log(T), \mathbf{X})$ from some distribution function $\mathcal{F}_{Y,\mathbf{X}}$, where $T \in \mathbb{R}^+$ denotes the survival time and Y its logarithm. The p-dimensional covariate vector $\mathbf{X} = (X_1, \ldots, X_p)$ is taken from a sample space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_p$. We assume that the conditional distribution $\mathcal{F}_{Y|\mathbf{X}} = \mathcal{F}_{Y|f(\mathbf{X})}$ of the response Y given the covariates $\mathbf{X}$ depends on the covariates $\mathbf{X}$ through a real-valued function $f : \mathcal{X} \to \mathbb{R}$. The regression function f, our parameter of interest, is an element of some parameter space $\mathbf{\Psi}$ and has minimal risk

$$\mathbb{E}_{Y,\mathbf{X}} L(Y, f(\mathbf{X})) = \int L(Y, f(\mathbf{X})) d\mathcal{F}_{Y,\mathbf{X}} = \min_{\psi \in \mathbf{\Psi}} \int L(Y, \psi(\mathbf{X})) d\mathcal{F}_{Y,\mathbf{X}}$$

for a suitable full data loss function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$. Our principle aim is to estimate the regression function f. Usually, an estimate $\hat{f}$ of f is computed via constrained minimization of the empirical risk defined by the full data loss function L. However, this minimization problem can only be solved when all quantities are observed. Naturally, this is not the case in the presence of censoring.

2.2 Observed Data World

In realistic set-ups we only observe random variables $\mathbf{O} = (\tilde{Y} = \log(\tilde{T}), \Delta, \mathbf{X})$, with time to event $\tilde{T} = \min(T, C)$ and censoring indicator $\Delta = I(T \leq C)$ from some distribution $\mathcal{F}_{\tilde{Y},\Delta,\mathbf{X}}$. We assume that the conditional censoring distribution $\mathbb{P}(C \leq c|\mathbf{Z})$ only depends on the covariates $\mathbb{P}(C \leq c|\mathbf{Z}) = \mathbb{P}(C \leq c|\mathbf{X})$. This assumption implies a *coarsening at random* (CAR) censoring mechanism (for details we refer to van der Laan and Robins, 2003). Furthermore, for the corresponding conditional censoring survivor function $G(c|\mathbf{X}) = \mathbb{P}(C > c|\mathbf{X})$ we assume that $G(T|\mathbf{X})$ is strictly greater than zero almost everywhere with respect to the full data distribution $\mathcal{F}_{Y,\mathbf{X}}$.

The parameter space $\mathbf{\Psi}$ is the function space of all candidate estimators $\psi : \mathcal{X} \to \mathbb{R}$ for

the regression function f. For an observed learning sample of n independent and identically distributed observations $\mathcal{L} = \{\mathbf{O}_i = (\tilde{Y}_i = \log(\tilde{T}_i), \Delta_i, \mathbf{X}_i); i = 1, \ldots, n\}$, we cannot evaluate the full data loss function $L(Y, \psi(\mathbf{X}))$ for the censored patients. Consequently, we cannot minimize the corresponding empirical risk defined in terms of the full data loss function $L(Y, \psi(\mathbf{X}))$ directly. The methodology presented in van der Laan and Robins (2003) solves this problem by replacing the full data loss function $L(Y, \psi(\mathbf{X}))$ by an observed data loss function $L(\tilde{Y}, \psi(\mathbf{X})|\eta)$ with nuisance parameter η where the risks of both loss functions coincide for all candidate estimators $\psi \in \mathbf{\Psi}$:

$$\mathbb{E}_{Y,\mathbf{X}} L(Y, \psi(\mathbf{X})) = \int L(Y, \psi(\mathbf{X})) d\mathcal{F}_{Y,\mathbf{X}} = \int L(\tilde{Y}, \psi(\mathbf{X})|\eta) d\mathcal{F}_{\tilde{Y},\Delta,\mathbf{X}} = \mathbb{E}_{\tilde{Y},\Delta,\mathbf{X}} L(\tilde{Y}, \psi(\mathbf{X})|\eta).$$

A description of the role of η will be given in Section 2.3. The basic idea is to minimize the empirical counterpart of $\mathbb{E}_{\tilde{Y},\Delta,\mathbf{X}} L(\tilde{Y}, \psi(\mathbf{X})|\eta)$ with respect to the candidate estimators $\psi \in \mathbf{\Psi}$, which is possible even in the imperfect observed data world.

2.3 Inverse Probability of Censoring Weights

One approach for defining the observed data loss function $L(\tilde{Y}, \psi(\mathbf{X})|\eta)$ is the application of inverse probability of censoring weights (IPC weights, van der Laan and Robins, 2003), here the nuisance parameter η is given by the conditional censoring survivor function G:

$$L(\tilde{Y}, \psi(\mathbf{X})|G) = L(\tilde{Y}, \psi(\mathbf{X})) \frac{\Delta}{G(\tilde{T}|\mathbf{X})}.$$

Basically, the full data loss function is weighted by the inverse probability of being censored after time $\tilde{T}$ given the covariates $\mathbf{X}$. The inverse probability $G(\tilde{T}|\mathbf{X})^{-1}$ exists because $G(\tilde{T}|\mathbf{X}) \geq G(T|\mathbf{X}) > 0$ by assumption. The corresponding empirical risk is the weighted average

$$\hat{\mathbb{E}}_{\tilde{Y},\Delta,\mathbf{X}} L(\tilde{Y}, \psi(\mathbf{X})|G) = n^{-1} \sum_{i=1}^{n} L(\tilde{Y}_i, \psi(\mathbf{X}_i)|\hat{G}) = n^{-1} \sum_{i=1}^{n} L(\tilde{Y}_i, \psi(\mathbf{X}_i)) \frac{\Delta_i}{\hat{G}(\tilde{T}_i|\mathbf{X}_i)} \tag{1}$$

and the regression function estimator $\hat{f}$ is derived by (constrained) minimization of (1) with respect to the candidate estimators $\psi \in \mathbf{\Psi}$. Note that the conditional censoring survivor

function G is typically unknown and needs to be replaced by an estimate $\hat{G}$. A Kaplan-Meier estimate $\hat{G}$ is the simplest choice but other procedures, for example a Cox model, are appropriate. For convenience, let $\mathbf{w} = (w_1, \ldots, w_n)$ with $w_i = \Delta_i \hat{G}(\tilde{T}_i|\mathbf{X}_i)^{-1}$ denote the IPC weights. Other choices of the observed data loss function are possible as well, such as that based on doubly robust inverse probability of censoring weights (DR-IPC weights, van der Laan and Robins, 2003).

3 Ensemble Learning

We present two algorithms pursuing some regularized minimization of (1): random forest and gradient boosting for censored data. The random forest approach seeks to minimize the empirical risk indirectly via a stabilization of randomized weak learners fitted on perturbed instances of the learning sample $\mathcal{L}$. In contrast, gradient boosting employs a functional gradient descent algorithm for minimizing the empirical risk (1).

3.1 Random Forest

From the observed learning sample $\mathcal{L} = \{(\tilde{Y}_i, \Delta_i, \mathbf{X}_i); i = 1, \ldots, n\}$, compute the weight vector $\mathbf{w}$. Note that the learning sample can be thought to include the censored observations as well, however with $w_i = 0$ iff $\Delta_i = 0$. The random forest algorithm with weights $\mathbf{w}$ basically works by defining the resampling probability of observation i in terms of the corresponding weight w_i.

Algorithm: Random Forest for Censored Data

Step 1 (Initialization). Set $m = 1$ and fix $M > 1$.

Step 2 (Bootstrap). Draw a random vector of case counts $\mathbf{v}_m = (v_{m1}, \ldots, v_{mn})$ from the multinomial distribution with parameters n and $(\sum_{i=1}^n w_i)^{-1}\mathbf{w}$.

Step 3 (Base Learner). Construct a partition $\pi_m = (R_{m1}, \ldots, R_{mK(m)})$ of the sample space $\mathcal{X}$ into $K(m)$ cells by means of a regression tree. The tree is build using the learning sample $\mathcal{L}$ with case counts $\mathbf{v}_m$, i.e., is based on a perturbation of the learning sample $\mathcal{L}$ with observation i occurring v_{mi} times. Computational details are given below.

Step 4 (Iteration). Increase m by one and repeat steps 2 and 3 until $m = M$.

Prognostic modeling is our main concern, i.e., we are interested in estimating the (log)-survival time $\hat{f}(\mathbf{x})$ for a patient with covariate status $\mathbf{x}$. The predicted status of the response variable is computed based on prediction weights

$$a_i(\mathbf{x}) = \sum_{m=1}^{M} v_{mi} \sum_{k=1}^{K(m)} I(\mathbf{X}_i \in R_{mk} \text{ and } \mathbf{x} \in R_{mk}); i = 1, \ldots, n.$$

The prediction weight $a_i(\mathbf{x})$ measures the 'similarity' of $\mathbf{x}$ to $\mathbf{X}_i$ $(i = 1, \ldots, n)$ by counting how many times the value $\mathbf{x}$ falls into the same cell as the ith observation in the learning sample. This is essentially an extension of the classical (unweighted) average of the predictions extracted from each single partition (cf. Breiman, 1996) as used also in Hothorn et al. (2004). The prediction $\hat{f}(\mathbf{x})$ can be computed as the solution of

$$\hat{Y} = \hat{f}(\mathbf{x}) = \operatorname*{argmin}_{y \in \mathbb{R}} \sum_{i=1}^{n} L(\tilde{Y}_i, y) a_i(\mathbf{x}).$$

For quadratic loss $L(Y, \psi(\mathbf{X})) = (Y - \psi(\mathbf{X}))^2$, the prediction is simply the weighted average of the observed log-survival times

$$\hat{Y} = \hat{\mathbb{E}}(Y|\mathbf{X} = \mathbf{x}) = \hat{f}(\mathbf{x}) = \left(\sum_{i=1}^{n} a_i(\mathbf{x}) \right)^{-1} \sum_{i=1}^{n} a_i(\mathbf{x}) \tilde{Y}_i.$$

The full data loss function can be evaluated here because the weights w_i and thus the case counts v_{mi} as well as the prediction weights $a_i(\mathbf{x})$ are zero for censored observations by definition.

In step 3 of the algorithm the partitions are usually induced by some form of recursive partitioning with additional randomization. This can be implemented by using only a small

number of randomly selected covariates for further splitting of every node of the tree. Note that random forest for censored data reduces to the original random forest procedure (Amit and Geman, 1997; Breiman, 2001a) when all events have been observed. Conceptually, the algorithm is not restricted to (randomized) trees as base learners, every other regression model can be applied as well. However, survival times need to be estimated via unweighted averages of the predictions extracted from all M base learners similar to the original bagging approach. A drawback of the random forest algorithm for censored data is that out-of-bag predictions and thus out-of-bag error rate estimates can't be computed when some observations are given a very large weight and are thus appearing in nearly every bootstrap sample.

3.2 Gradient Boosting - Full Data World

In the full data world, the generic boosting algorithm sketched in the sequel can be applied to pursue minimization of $\sum_{i=1}^{n} L(Y_i, \psi(\mathbf{X}_i))$ via functional gradient descent (for the details we refer to Friedman, 2001 and Bühlmann and Yu, 2003). Let U denote a pseudo response variable. A base learner regressing the pseudo response U on the covariates $\mathbf{X}$ is denoted by $h(\cdot|\vartheta_{U,\mathbf{X}})$, where $\vartheta_{U,\mathbf{X}}$ is a vector of parameters. Fitting the base learner can be performed by minimizing any loss function, for example solving the least squares problem

$$\hat{\vartheta}_{U,\mathbf{X}} = \underset{\vartheta}{\operatorname{argmin}} \sum_{i=1}^{n} (U_i - h(\mathbf{X}_i|\vartheta))^2. \tag{2}$$

Algorithm: Generic Gradient Boosting

Step 1 (Initialization). Define $U_i = Y_i$ $(i = 1, \ldots, n)$, set $m = 0$ and $\hat{f}_0(\cdot) = h(\cdot|\hat{\vartheta}_{U,\mathbf{X}})$. Fix $M > 1$.

Step 2 (Gradient). Compute the residuals

$$U_i = -\left.\frac{\partial L(Y_i, \psi)}{\partial \psi}\right|_{\psi = \hat{f}_m(\mathbf{X}_i)}$$

and fit the base learner $h(\cdot|\hat{\vartheta}_{U,\mathbf{X}})$ to the new 'responses' U_i as in (2).

Step 3 (Update). Update $\hat{f}_{m+1}(\cdot) = \hat{f}_m(\cdot) + \nu h(\cdot|\hat{\vartheta}_{U,\mathbf{X}})$ with step size $0 < \nu \leq 1$, for example $\nu = 0.1$.

Step 4 (Iteration). Increase m by one and repeat steps 2 and 3 until $m = M$.

Note that, unlike for the random forest algorithm, the number of iterations M is a tuning parameter which needs to be determined via cross-validation. Internal stop criteria are available for special cases, which we will discuss in Section 3.4.

3.3 Gradient Boosting - Observed Data World

In the observed data world, we cannot solve the least squares problem (2) for fitting the base learner since we do not have access to U_i which is a function of Y_i. But the right hand side of (2) can be replaced by an empirical risk as in (1) and we then get the weighted least squares problem

$$\hat{\vartheta}_{\tilde{U},\mathbf{X}} = \underset{\vartheta}{\operatorname{argmin}} \sum_{i=1}^{n} w_i(\tilde{U}_i - h(\mathbf{X}_i|\vartheta))^2 \text{ with } \tilde{U}_i = -\left.\frac{\partial L(\tilde{Y}_i, \psi)}{\partial \psi}\right|_{\psi=\hat{f}_m(\mathbf{X}_i)}.$$

Thus, the following algorithm can be applied to minimize (1).

Algorithm: Generic Gradient Boosting for Censored Data

Step 1 (Initialization). Define $\tilde{U}_i = \tilde{Y}_i$ $(i = 1, \ldots, n)$, set $m = 0$ and $\hat{f}_0(\cdot) = h(\cdot|\hat{\vartheta}_{\tilde{U},\mathbf{X}})$. Fix $M > 1$.

Step 2 (Gradient). Compute the residuals

$$\tilde{U}_i = -\left.\frac{\partial L(\tilde{Y}_i, \psi)}{\partial \psi}\right|_{\psi=\hat{f}_m(\mathbf{X}_i)}$$

and fit the base learner $h(\cdot|\hat{\vartheta}_{\tilde{U},\mathbf{X}})$ to the new 'responses' $\tilde{U}_i$ by weighted least squares.

Step 3 (Update). Update $\hat{f}_{m+1}(\cdot) = \hat{f}_m(\cdot) + \nu h(\cdot|\hat{\vartheta}_{\tilde{U},\mathbf{X}})$ with step size $0 < \nu \leq 1$.

Step 4 (Iteration). Increase m by one and repeat steps 2 and 3 until $m = M$.

The boosting estimator is $\hat{f}_M(\mathbf{x})$ and the predicted log-survival time for an observation with covariate status $\mathbf{x}$ is $\hat{Y} = \hat{f}_M(\mathbf{x})$. The algorithm proposed here reduces to the original form of gradient boosting in the absence of censoring. For quadratic loss $L(Y, \psi(\mathbf{X})) = (Y - \psi(\mathbf{X}))^2/2$, the algorithm is obtained by residuals $\tilde{U}_i = \tilde{Y}_i - \hat{f}_m(\mathbf{X}_i)$ in the mth boosting iteration and we call this method L_2-boosting for censored data.

3.4 Choice of Base Learners and Stop Criterion

The base learner h needs to be able to take weights $\mathbf{w}$ into account. Recursive partitioning procedures are popular choices of such base learners and the methodology of Molinaro et al. (2004) can be applied directly. Bühlmann and Yu (2003) suggested univariate smoothing splines: In each boosting iteration, one of the p covariates is selected and the relationship between the residuals U and the selected covariate is modeled by a smoothing spline with low degrees of freedom.

Another possibility which is studied here is the application of component-wise least squares (Bühlmann, 2004b). This choice is computationally attractive and allows for the definition of an AIC-based internal stop criterion. Let $\mathbf{X}^{(j)}$ denote the design matrix associated with the jth covariate. In case the jth covariate is a factor, the matrix $\mathbf{X}^{(j)}$ is a dummy matrix. A column for the intercept term could be included. $\mathbf{W}$ denotes the $n \times n$ diagonal matrix with diagonal elements $\mathbf{W}_{ii} = \sqrt{w_i}, i = 1, \ldots, n$. Then

$$\mathbf{H}^{(j)} = \mathbf{X}^{(j)} \left(\left(\mathbf{W}\mathbf{X}^{(j)} \right)^\top \left(\mathbf{W}\mathbf{X}^{(j)} \right) \right)^{-1} \left(\mathbf{W}\mathbf{X}^{(j)} \right)^\top \mathbf{W}$$

is the usual hat matrix for computing predictions of a simple linear model with covariate j alone. In the mth boosting iteration, we select the covariate with minimum empirical risk, i.e.,

$$k_m = \underset{j=1,\ldots,p}{\operatorname{argmin}} \sum_{i=1}^{n} w_i (\tilde{U}_i - (\mathbf{H}^{(j)} \tilde{\mathbf{U}})_i)^2$$

where $\tilde{\mathbf{U}} = (\tilde{U}_1, \ldots, \tilde{U}_n)^\top$ is the vector of pseudo responses in the m step. The fit in the mth

step can be written in terms of the boosting hat operator $\left(\hat{f}_m(\mathbf{X}_1), \ldots, \hat{f}_m(\mathbf{X}_n)\right)^\top = \mathbf{B}\tilde{\mathbf{Y}}$ as introduced by Bühlmann and Yu (2003), where $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)^\top$ denotes the n-vector of responses extracted from $\mathcal{L}$. In the first boosting iteration, the boosting operator is $\mathbf{B}_0 = \nu\mathbf{H}^{(k_0)}$ and the update step 3 can be written as $\mathbf{B}_{m+1} = \mathbf{B}_m + \nu\mathbf{H}^{(k_m)}(\mathbf{I}_n - \mathbf{B}_m)$ where the $n \times n$ matrix $\mathbf{I}_n$ denotes the identity matrix. This formulation of boosting in terms of a boosting operator opens up the way to an AIC-based internal stop criterion (Bühlmann, 2004b). The trace of the boosting operator $\mathbf{B}_m$ is interpreted as degrees of freedom and a corrected version of AIC can be computed by

$$AIC(m) = \log(\hat{\sigma}^2) + \frac{1 + \text{trace}(\mathbf{B}_m)/n}{1 - (\text{trace}(\mathbf{B}_m) + 2)/n} \text{ with } \hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n} w_i'(\tilde{Y}_i - (\mathbf{B}_m\tilde{\mathbf{Y}})_i)^2$$

where the weights have been rescaled to $w_i' = w_i(\sum_i w_i)^{-1}n$. An estimate of the optimal number of boosting iterations is $\hat{M} = \text{argmin}_{m=1,\ldots,M} AIC(m)$.

4 Illustrations and Applications

Predictive modeling is the primary domain of ensemble methods, especially in situations where the number of covariates is large relative to the number of (uncensored) observations. A typical application is the construction of novel tumor classification schemes based on gene expression profiling data. One representative of such investigations is a study on acute myeliod leukemia (AML) patients recently published by Bullinger et al. (2004). The main focus of this study was on the differentiation of previously unknown tumor subclasses by means of genetic information. Here, we try to construct 'black box' predictors for the survival time of AML patients incorporating both clinical and genetic information. Although the random forest or boosting estimate of the regression function f may be arbitrarily complex, some insight into the nature of the regression relationship is necessary in order to compare the fitted model with subject matter knowledge. In our second application, random forest and boosting are applied to data of a well-analyzed study on node positive breast cancer, and we compare the

estimated flexible regression functions with previously published findings. All analyses were performed within the R system for statistical computing (R Development Core Team, 2004), version 2.0.1. Until published on CRAN, implementations of the algorithms applied here are available from the authors upon request.

4.1 Acute Myeloid Leukemia

The treatment of patients suffering from acute myeloid leukemia (AML) is determined by a tumor classification scheme taking the status of various cytogenetic aberrations into account. Bullinger et al. (2004) investigate an extended tumor classification scheme incorporating molecular subgroups of the disease obtained by gene expression profiling. A combination of unsupervised and supervised techniques is applied to define a binary outcome predictor (good vs. poor prognosis) taking into account the expression measures of 133 selected genes (which are represented by 149 cDNAs). This binary surrogate variable is shown to discriminate between patients with short and longer survival in an independent sample of patients.

Instead of using a binary variable summarizing expression levels of 149 cDNAs, random forest and L_2-boosting are applied to construct predictors based on both the clinical data and the expression levels of the genes selected by Bullinger et al. (2004). The results reported here are based on clinical and gene expression data published online at http://www.ncbi.nlm.nih.gov/geo, accession number GSE425. The overall survival time and censoring indicator as well as the clinical variables age, sex, lactic dehydrogenase level (LDH), white blood cell count (WBC), and treatment group are taken from Supplementary Table 1. In addition, two molecular markers, the fms-like tyrosine kinase 3 (FLT3) and the mixed-lineage leukemia (MLL) gene, are available from this table as well as cytogenetic information helpful to define a risk score ('low': karyotype t(8;21), t(15;17) and inv(16); 'intermediate': normal karyotype and t(9;11); and 'high': all other forms). The Supplementary Table 6 gives a list of 149 cDNAs selected by Bullinger et al. (2004) for building a binary prognostic factor, 147 of them

have corresponding expression levels in Supplementary Table 3. Our analysis utilizes one single learning sample of $n = 116$ patients, 68 patients died during the study period. The IPC weights are derived from a simple Kaplan-Meier estimate $\hat{G}$ of the censoring survivor function. For one patient a very late event was observed and we restrict the IPC weight for this patient to a value of five. Missing values in the expression matrix of all 6283 cDNAs and 116 patients are imputed using $k = 10$ nearest neighbor averaging (Troyanskaya et al., 2001) as implemented in package `pamr` (Hastie et al., 2004). In total, 62 patients with IPC weights greater than zero had complete observations for the clinical variables and are used in the sequel.

Random forest for censored data (RF) with 10 covariates randomly selected in each node of $M = 250$ trees and L_2-boosting for censored data (L2B) with component-wise linear models and AIC-based stopping criterion ($\hat{M} = 350$) were trained using both the eight clinical variables and the information covered by the expression levels ($p = 155$). The fit of both learners is depicted in Figure 1 and indicates a reasonable agreement between observed and predicted (log)-survival times for both algorithms.

Both candidate models are compared with the naive prediction by means of a benchmark experiment following Hothorn et al. (2005). From the learning sample $\mathcal{L}$, 100 bootstrap samples are drawn and the performance measures of all candidate models, i.e., the empirical risk defined in terms of the IPC weights, are evaluated on the same sample of out-of-bootstrap observations in an unreplicated complete block design. The benchmark experiments are performed conditional on the IPC weights, since we are interested in a comparison between the candidate models only. In order to investigate whether the molecular information of the expression levels helps to predict the survival time we study in addition the performance of both algorithms when faced with a learning sample consisting of the clinical variables only (cRF and cL2B with $p = 8$). The joint and marginal distributions of the performance measures evaluated on the out-of-bootstrap observations are displayed in Figure 2, with median out-of-

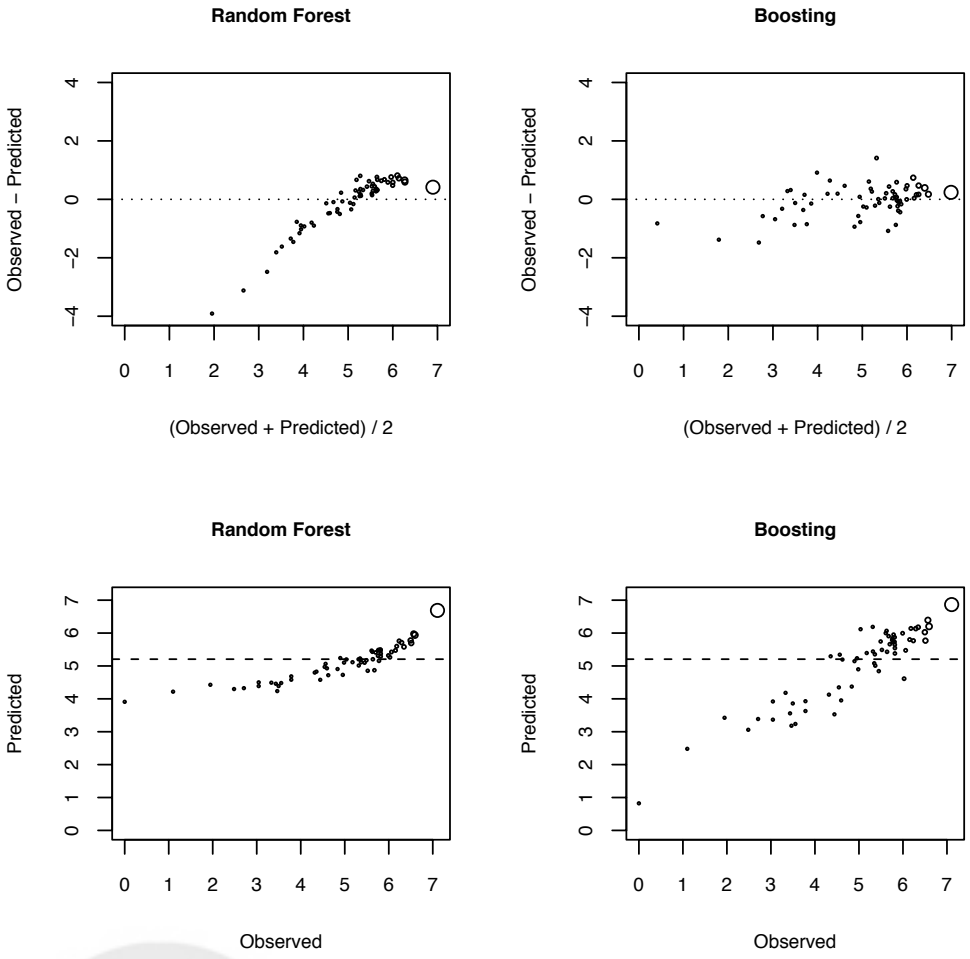

Figure 1: AML data: Mean-difference plots (top) and scatterplots (bottom) of observed and predicted log-survival time of random forest and L_2-boosting for censored data. The radius of the circles is proportional to the IPC weights and the dashed horizontal line is the weighted mean (with IPC weights) of the log-survival times, i.e., the prediction without any knowledge of the covariates.

bootstrap errors of 2.451 (mean), 2.382 (random forest) and 1.769 (L_2-boosting). In general, the performance distributions of the five candidate models show a global difference (asymptotic p-value < 0.0001, Friedman test). All pair-wise multiple comparisons based on Friedman rank sums (Wilcoxon–Nemenyi–McDonald–Thompson, see Hollander and Wolfe, 1999, Chapter 7.3) indicate that the naive prediction of the weighted mean is outperformed by AIC-based L_2-boosting (adjusted p-value < 0.0001). There is no evidence that the performance distributions of random forest and the weighted mean differ (adjusted p-value $= 0.4909$).

However, the distribution of the empirical risk of both ensemble methods is lower when only the eight clinical covariates are used (all adjusted p-values < 0.0001). This supports the hypothesis that the raw gene expression levels do not help to improve the prediction of survival time. Bullinger et al. (2004) argue that the 'likelihood and the duration of survival are likely to be fairly crude surrogates for the underlying biologic characteristics distinguishing prognostically relevant tumor subclasses' and therefore propose an alternative strategy utilizing a prognostic variable obtained from a mix of cluster analysis and binary classification.

4.2 Node Positive Breast Cancer

A prospective, controlled clinical trial on the treatment of node positive breast cancer patients was conducted by the German Breast Cancer Study Group (GBSG-2), a detailed description of the study is given in Schumacher et al. (1994). Patients not older than 65 years with positive regional lymph nodes but no distant metastases were included in the study. Complete data on $p = 7$ prognostic factors for $n = 686$ women are used in Sauerbrei and Royston (1999) for prognostic modeling by means of multivariate fractional polynomials, i.e., flexible linear regression models based on transformed covariates. These findings will serve as the basis for the assessment of the diagnostic capabilities of survival ensembles.

Observed hypothetical prognostic factors are age, menopausal status, tumor size, tumor grade, number of positive lymph nodes, progesterone receptor, estrogen receptor, and the

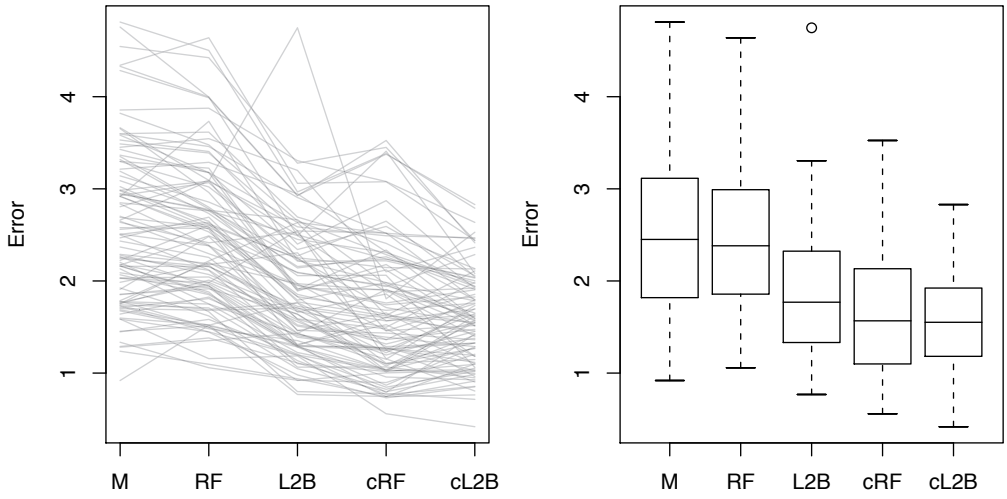

Figure 2: AML data: Parallel coordinate plot and boxplots of the joint and marginal distribution of the error evaluated on 100 out-of-bootstrap samples for the simple weighted mean (M), random forest (RF), and L_2-boosting for censored data with component-wise least squares (L2B). In addition, the bootstrap errors for both ensemble methods based on the learning sample of the eight clinical covariates only are given (cRF and cL2B).

information of whether or not a hormonal therapy was applied. The recurrence free survival time is the response variable of interest. The data are available in the R-package `ipred` (Peters et al., 2002) and the IPC weights are derived from a simple Kaplan-Meier estimate $\hat{G}$ of the censoring survivor function. The weights are restricted to a maximal value of five because of three very late events. The performance of four candidate algorithms is investigated: an ordinary linear model fitted via IPC-weighted least squares (LM), regression trees based on the IPC weights (RP) as suggested by Molinaro et al. (2004) using the implementation in package `rpart` (Therneau and Atkinson, 1997), random forest for censored data (RF, with five covariates randomly selected in each node of 100 trees) and L_2-boosting for censored data (L2B) with component-wise linear models and AIC-based stopping criterion.

The AIC-criterion for L_2-boosting suggests to stop after the 86th boosting iteration. Figure 3 depicts a mean-difference plot of observed and predicted logarithms of recurrence free survival for all four models. The figure leads to the impression that the relationship between the covariates and the recurrence free survival time is relatively weak, a finding supported by an analysis with the Brier score in Hothorn et al. (2004).

The performance of the four candidate models is compared by means of a benchmark experiment utilizing the framework given by Hothorn et al. (2005) as described above. In order to study the stability of the models in high-dimensional situations, we choose a strategy in-between an analysis of the original data and a simulation experiment. We add $p_+ = (10, 50, 100)$ uncorrelated covariates drawn from a uniform distribution to the observed learning sample $\mathcal{L}$ and evaluate the performance using the out-of-bootstrap observations as described earlier. The results are depicted in Figure 4. Many-to-one comparisons with the weighted mean based on Friedman rank sums indicate that for the learning sample with only the original covariates ($p_+ = 0$) the linear model, boosting and random forest perform better than the weighted mean (all adjusted p-values < 0.0001). There is no evidence that the performance distributions of regression trees and the weighted mean differ (adjusted

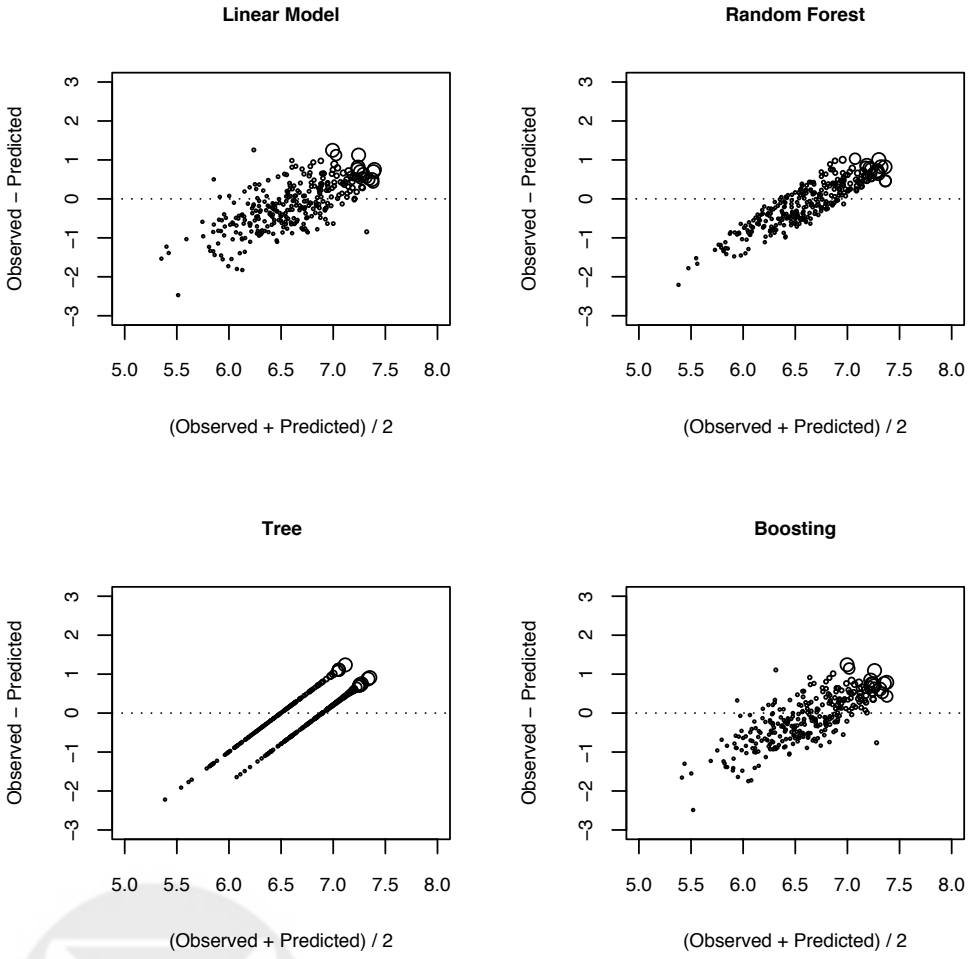

Figure 3: GBSG-2 data: Mean-difference plots of observed and predicted log recurrence free survival for all four candidate methods. The radius of the circles is proportional to the IPC weights.

p-value = 0.9653). Again, the relative improvement compared with the weighted mean is relatively small. For an increasing number of random covariates the linear model is heavily affected by overfitting but the ensemble methods are rather stable. For $p_+ = 50$ additional random covariates, the bootstrap test set error of random forest and boosting is smaller than that for the weighted mean (both adjusted p-values = 0.0001). However, there is only weak evidence that random forest performs better than the weighted mean for learning samples with $p_+ = 100$ additional random covariates added (adjusted p-value = 0.0303); boosting cannot outperform the mean (adjusted p-value = 0.5830) in this situation. The relative stability of regression trees is caused by the fact that the trees are pruned back to stumps or the root node most of the time.

	M	RP	LM	RF	L2B
$p_+ = 0$	0.311	0.311	0.291	0.293	0.289
$p_+ = 10$	0.311	0.311	0.321	0.296	0.299
$p_+ = 50$	0.311	0.311	0.423	0.305	0.303
$p_+ = 100$	0.311	0.311	0.647	0.308	0.310

Table 1: Benchmark experiments for the GBSG-2 data: Median performance for 100 bootstrap samples for the weighted mean (M), recursive partitioning (RP), a linear model (LM), random forest (RF), and L_2-boosting (L2B) for censored data with component-wise least squares.

Sauerbrei and Royston (1999) provide an in-depth analysis of the GBSG-2 data focusing on fractional polynomials as interpretable but flexible regression models. We compare the estimated regression function f represented by random forest and boosting with the findings reported in their paper, where a non-linear influence of the number of positive nodes, age, and progesterone receptor was identified by visualization of the covariates and the corresponding (partial) linear predictors. With Figures 5 and 6 we proceed in a similar way by plotting the covariates against the predictions (such strategies were also applied for classification problems

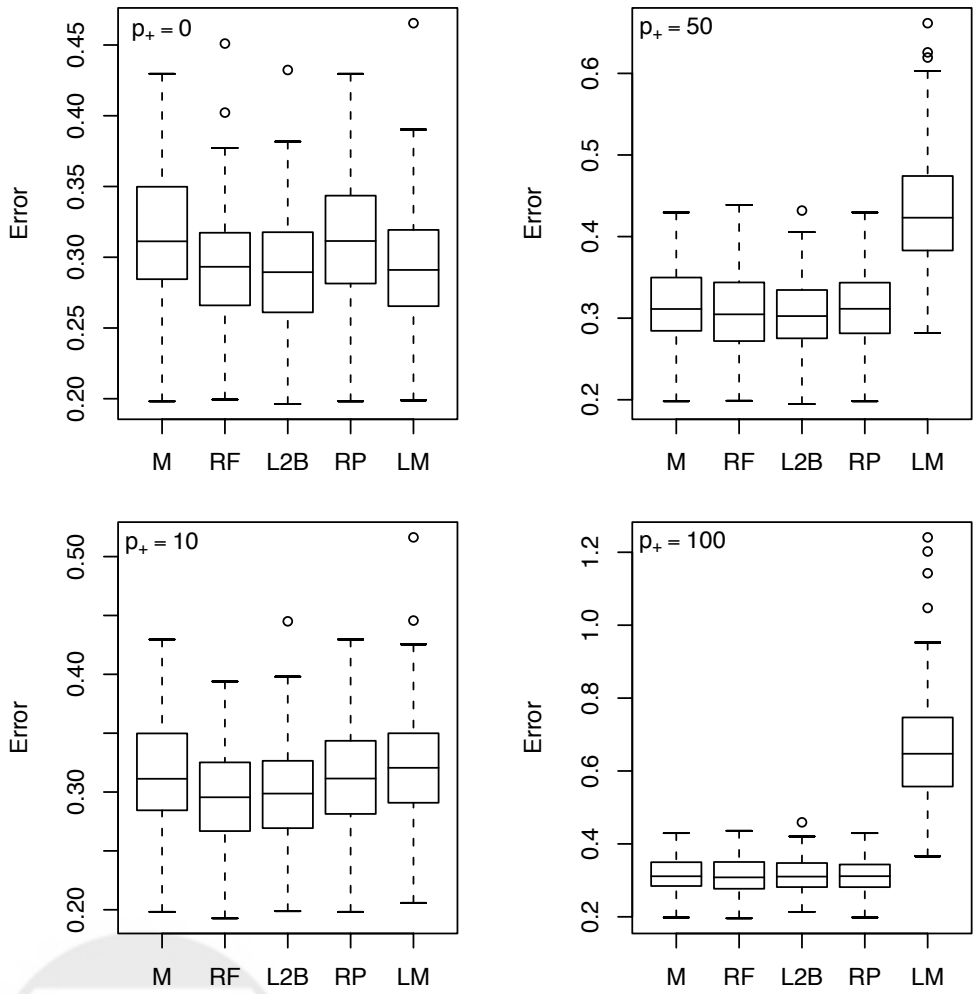

Figure 4: GBSG-2 data: The marginal distribution of the error evaluated on 100 out-of-bootstrap samples for the weighted mean (M), random forest (RF), L_2-boosting for censored data with component-wise least squares (L2B), recursive partitioning (RP) and a simple linear model (LM) for a number of additional random covariates p_+.

by Breiman, 2001b and Garczarek and Weihs, 2003).

The predicted log recurrence free survival time decreases with increasing number of positive lymph nodes (up to about 15 positive lymph nodes) for both random forest and boosting in a way nearly identical to the finding reported by Sauerbrei and Royston (1999). Both boosting and random forest suggest a relationship between age and survival time, namely a decreasing risk for women up to an age of 40 to 45 years and a nearly constant risk for older women, as in Sauerbrei and Royston (1999). A strong influence of the estrogen receptor is indicated by both ensemble methods, however, estrogen receptor measurements were not included in any of the models studied by Sauerbrei and Royston (1999). Progesterone receptor values (restricted to values less than 100 fmol/l) indicate a relationship to recurrence free survival: Very small values (less than about 10, say) are associated with short recurrence free survival times whereas higher values indicate longer recurrence free survival times. A similar finding is reported by Sauerbrei and Royston (1999).

5 Discussion

The two algorithms presented in this paper extend ensemble prediction to censored data problems. Ensemble techniques have been developed at the borderline between machine learning and statistics in the past decade; previous attempts to apply the main ideas to survival time data were bound to established key ingredients such as the partial likelihood (Ridgeway, 1999), the Brier score for censored data (Benner, 2002), or survival trees (Hothorn et al., 2004) and, consequently, inherited the associated difficulties.

The general estimation framework of van der Laan and Robins (2003) allows for a sound theoretical formulation of the underlying risk optimization problems which can be solved with the new algorithms. Moreover, the framework enables us to apply well-known cross-validation techniques for model evaluation (Keleş et al., 2004). Both ensemble algorithms are generic in the sense that arbitrary loss functions, for example absolute loss, and other base learners

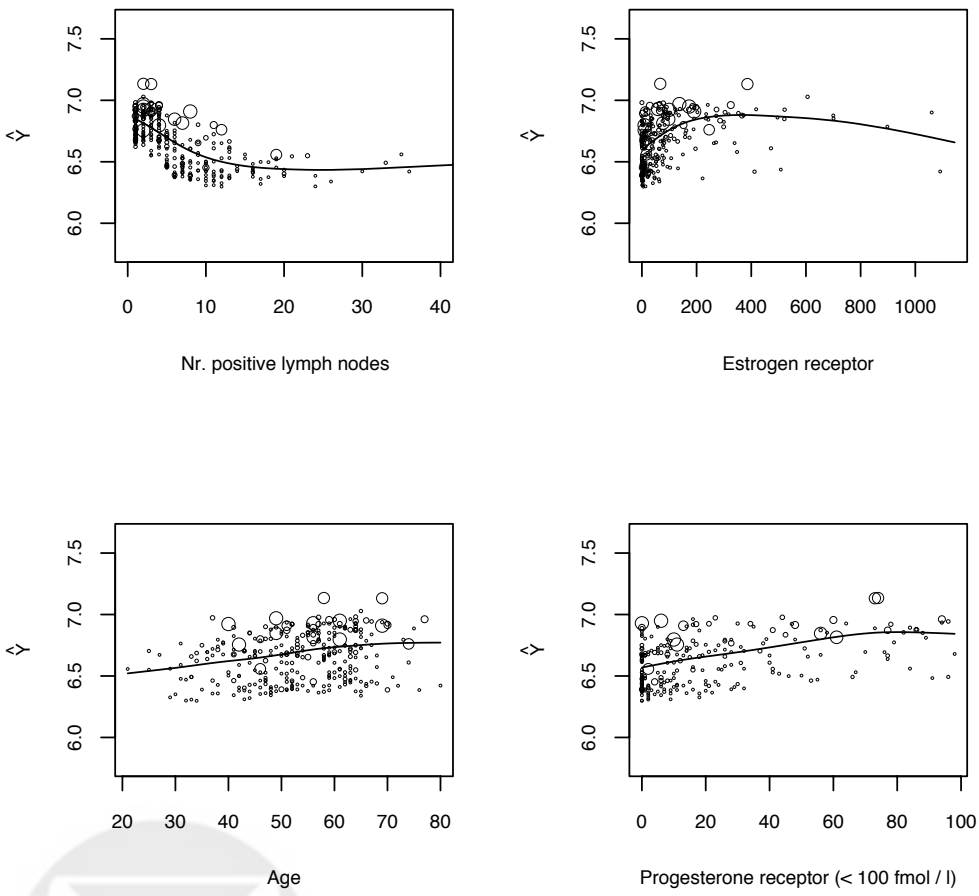

Figure 5: GBSG-2 data: Scatterplots of selected covariates and predicted log recurrence free survival time obtained from random forest for censored data. A smoothing spline with four degrees of freedom is plotted. The radius of the circles is proportional to the IPC weights.

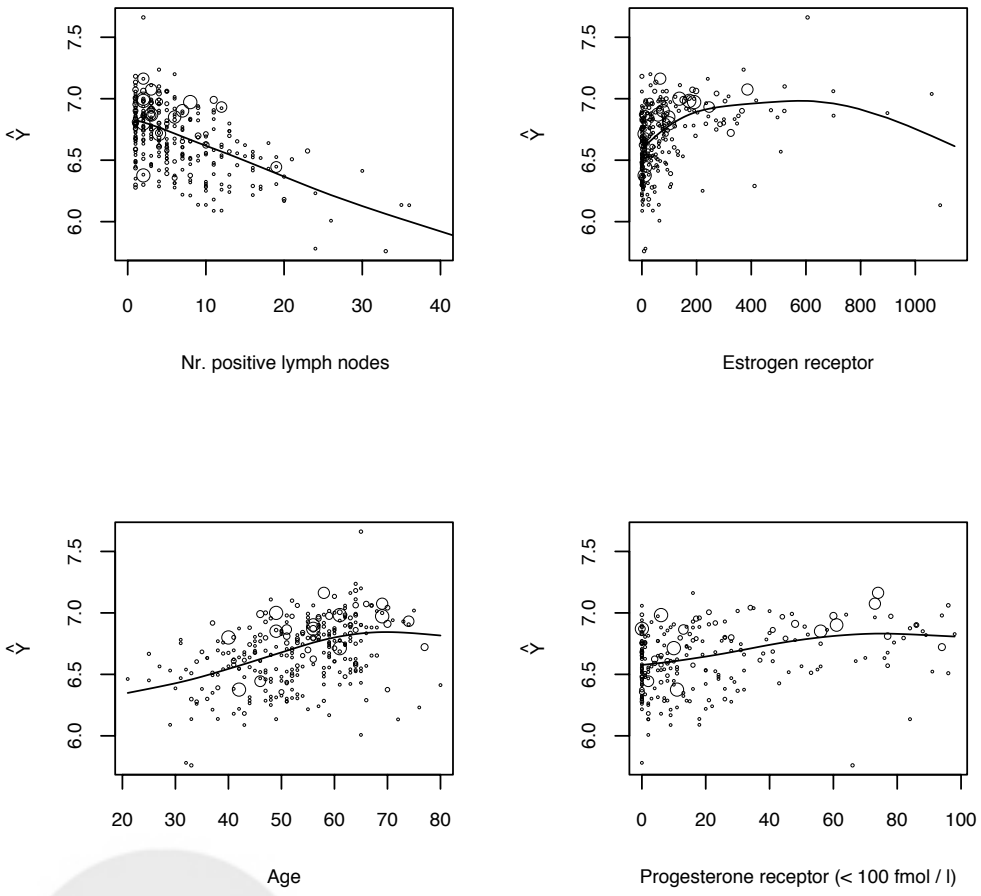

Figure 6: GBSG-2 data: Scatterplots of selected covariates and predicted log recurrence free survival time obtained from L_2-boosting for censored data with component-wise least squares. A smoothing spline with four degrees of freedom is plotted. The radius of the circles is proportional to the IPC weights.

can be implemented easily. It should be noted that our implementations do not require an external choice of hyper parameters. Another important issue is the fact that the random forest and the boosting algorithm reduce to their original complete data form in the absence of censoring.

In this situation with uncensored data, the flexibility and stability of both the random forest and the boosting approach have been demonstrated in many benchmark experiments; we therefore restricted ourself to a semi-artificial benchmark experiment with varying number of covariates based on the GBSG-2 data. The main focus of our analysis of the AML and the GBSG-2 data is on the practical advantages of the methodology in terms of prediction accuracy and diagnostic ability. The results of flexible diagnostic modeling with fractional polynomials published by Sauerbrei and Royston (1999) could be reproduced for the GBSG-2 data. Thus, ensemble techniques are not just superb 'black boxes' in terms of prediction accuracy but can be used to investigate the nature of the regression relationship inherent in the data. We depicted simple partial relationships between one covariate and the predicted survival times, more advanced approaches for the visualization of complex regression relationships (Nason et al., 2004) are applicable as well.

The definition of the observed data loss function is the basis of all subsequent calculations. For the analysis of the AML and the GBSG-2 data we used inverse probability of censoring weights obtained from a Kaplan-Meier estimate $\hat{G}$ of the censoring survivor function, i.e., an estimate based on $\tilde{T}_i$ and $1-\Delta_i$ for observations $i = 1, \ldots, n$. Molinaro et al. (2004) applied a Cox model to estimate the weights which allows for modeling the censoring survivor function based on information covered by a subset of the covariates. Robustness properties are studied theoretically in van der Laan and Robins (2003) and lead to double robust inverse probability of censoring weights (DR-IPC weights) as an alternative scheme. However, the practical implications of a misspecification of the weights, for example by omitting an important covariate when estimating the censoring distribution, and advantages or disadvantages of parametric,

semi-parametric, or non-parametric modeling strategies need to be investigated by means of artificial simulation experiments. Another idea is to stabilize the estimate of the censoring distribution, and thus to stabilize the weights, by some form of ensemble technique prior to modeling or even simultaneously with the estimation of the regression function. Those issues are to be addressed in future research.

References

Altman, D. G. and Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine* **19**, 453–473.

Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation* **9**, 1545–1588.

Bair, E. and Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology* **2**, 0511–0522. URL http://www.plosbiology.org.

Bedrick, E. J., Exuzides, A., Johnson, W. O. and Thurmond, M. C. (2002). Predictive influence in the accelerated failure time model. *Biostatistics* **3**, 331–346.

Benner, A. (2002). Application of "aggregated classifiers" in survival time studies. In Härdle, W. and Rönz, B. (eds.), *Proceedings in Computational Statistics: COMPSTAT 2002*. Heidelberg: Physica-Verlag.

Breiman, L. (1996). Bagging predictors. *Machine Learning* **24**, 123–140.

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

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science* **16**, 199–231. With discussion.

Breiman, L. (2002). *How to use survival forests*. URL http://stat-www.berkeley.edu/

users/breiman/.

Bryan, J., Yu, Z. and van der Laan, M. J. (2004). Analysis of longitudinal marginal structural models. *Biostatistics* **5**, 361–380.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika* **66**, 429–436.

Bühlmann, P. (2004a). Bagging, boosting and ensemble methods. In Gentle, J. E., Härdle, W. and Mori, Y. (eds.), *Handbook of Computational Statistics*. Springer-Verlag, Berlin, Heidelberg, pp. 877–907.

Bühlmann, P. (2004b). Boosting for high-dimensional linear models. Technical report, ETH Zürich, Switzerland. URL http://www.stat.math.ethz.ch.

Bühlmann, P. and Yu, B. (2003). Boosting with L_2 loss: Regression and classification. *Journal of the American Statistical Association* **98**, 324–338.

Bullinger, L., Döhner, K., Bair, E., Fröhlich, S., Schlenk, R. F., Tibshirani, R., Döhner, H. and Pollack, J. R. (2004). Use of gene-expression profiling to identify prognostic subclasses in adult acute myloid leukemia. *New England Journal of Medicine* **350**, 1605–1616.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B* **34**, 187–202. With discussion.

Dudoit, S. and van der Laan, M. J. (2003). Unified cross-validation methodology for selection among estimators: Finite sample results, asymptotic optimality, and applications. Technical Report 130, Devision of Biostatistics, University of California, Berkeley, Berkeley, California. URL http://www.bepress.com/ucbbiostat/paper130.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **29**, 1189–1202.

Garczarek, U. M. and Weihs, C. (2003). Standardizing the comparison of partitions. *Computational Statistics* **18**, 143–162.

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

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association* **87**, 942–951.

Hastie, T. and Tibshirani, R. (2004). Efficient quadratic regularization for expression arrays. *Biostatistics* **5**, 329–340.

Hastie, T., Tibshirani, R., Narasimhan, B. and Chu, G. (2004). *pamr: Prediction analysis for microarrays*. URL http://CRAN.R-project.org. R package version 1.24.

Henderson, R. (1995). Problems and prediction in survival-data analysis. *Statistics in Medicine* **14**, 161–184.

Hollander, M. and Wolfe, D. A. (1999). *Nonparametric statistical inference*. New York: John Wiley & Sons, 2nd edition.

Hothorn, T., Lausen, B., Benner, A. and Radespiel-Tröger, M. (2004). Bagging survival trees. *Statistics in Medicine* **23**, 77–91.

Hothorn, T., Leisch, F., Zeileis, A. and Hornik, K. (2005). The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics* URL http://www.wu-wien.ac.at/am/reports.htm#82. (accepted).

Huang, J. and Harrington, D. (2005). Iterative partial least squares with right-censored data analysis: A comparison to other dimension reduction techniques. *Biometrics* **61**, 17–24.

James, I. (1998). Accelerated failure-time models. In Armitage, P. and Colton, T. (eds.), *Encyclopedia of Biostatistics*. John Wiley & Sons, Chichester, pp. 26–30.

Keleş, S., van der Laan, M. J. and Dudoit, S. (2004). Asymptotically optimal model selection method for regression on censored outcomes. *Bernoulli* **10**, 1011–1037.

Kooperberg, C., Stone, C. J. and Truong, Y. K. (1996). Hazard regression. *Journal of the American Statistical Association* **90**, 78–94.

LeBlanc, M. and Crowley, J. (1999). Adaptive regression splines in the Cox model. *Biometrics* **55**, 204–213.

Li, L. and Li, H. (2004). Dimension reduction methods for microarrays with applications to censored survival data. *Bioinformatics* **20**, 3406–3412.

Molinaro, A. M., Dudoit, S. and van der Laan, M. J. (2004). Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis* **90**, 154–177.

Nason, M., Emerson, S. and LeBlanc, M. (2004). CARTscans: A tool for visualizing complex models. *Journal of Computational and Graphical Statistics* **14**, 807–825.

Orbe, J., Ferreira, E. and Núñez-Antón, V. (2003). Censored partial regression. *Biostatistics* **4**, 109–121.

Peters, A., Hothorn, T. and Lausen, B. (2002). ipred: Improved predictors. *R News* **2**, 33–36. URL http://CRAN.R-project.org/doc/Rnews/. ISSN 1609-3631.

R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org. ISBN 3-900051-00-3.

Ridgeway, G. (1999). The state of boosting. *Computing Science and Statistics* **31**, 172–181.

Ripley, R. M., Harris, A. L. and Tarassenko, L. (2004). Non-linear survival analysis

using neural networks. *Statistics in Medicine* **23**, 825–842.

Sauerbrei, W. (1999). The use of resampling methods to simplify regression models in medical statistics. *Journal of the Royal Statistical Society, Series C* **48**, 313–329.

Sauerbrei, W. and Royston, P. (1999). Building multivariable prognostic and diagnostic models: Transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society, Series A* **162**, 71–94.

Schemper, M. (2003). Predictive accuracy and explained variation. *Statistics in Medicine* **22**, 2299–2308.

Schumacher, M., Basert, G., Bojar, H., Hübner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., RLA Neumann and Rauschecker HF for the German Breast Cancer Study Group (1994). Randomized 2×2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology* **12**, 2086–2093.

Sinisi, S. E. and van der Laan, M. J. (2004). Deletion/substitution/addition algorithm in learning with applications in genomics. *Statistical Applications in Genetics and Molecular Biology* **3**, Article 18.

Stute, W. (1999). Nonlinear censored regression. *Statistica Sinica* **9**, 1089–1102.

Therneau, T. M. and Atkinson, E. J. (1997). An introduction to recursive partitioning using the rpart routine. Technical Report 61, Section of Biostatistics, Mayo Clinic, Rochester. URL http://www.mayo.edu/hsr/techrpt/61.pdf.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525.

van der Laan, M. J., Dudoit, S. and van der Vaart, A. W. (2004). The cross-validated adaptive epsilon-net estimator. Technical Report 142, Division of Biostatistics, University of California, Berkeley, California. URL http://www.bepress.com/ucbbiostat/paper142.

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