STATISTICS IN MEDICINE
Statist. Med. 2004; **23**:77–91 (DOI: 10.1002/sim.1593)

Bagging survival trees

Torsten Hothorn[1], Berthold Lausen[1,*,†], Axel Benner[2] and Martin Radespiel-Tröger[3]

[1]*Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander-University, Erlangen-Nuremberg, Waldstraße 6, D-91054 Erlangen, Germany*
[2]*German Cancer Research Center, Central Unit Biostatistics, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany*
[3]*Population-based Cancer Registry Bavaria, Carl-Thiersch-Straße 7, D-91052 Erlangen, Germany*

SUMMARY

Predicted survival probability functions of censored event free survival are improved by bagging survival trees. We suggest a new method to aggregate survival trees in order to obtain better predictions for breast cancer and lymphoma patients. A set of survival trees based on B bootstrap samples is computed. We define the aggregated Kaplan–Meier curve of a new observation by the Kaplan–Meier curve of all observations identified by the B leaves containing the new observation. The integrated Brier score is used for the evaluation of predictive models. We analyse data of a large trial on node positive breast cancer patients conducted by the German Breast Cancer Study Group and a smaller 'pilot' study on diffuse large B-cell lymphoma, where prognostic factors are derived from microarray expression values. In addition, simulation experiments underline the predictive power of our proposal. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: bootstrap aggregation; censored data; prognostic factors; Brier score

1. INTRODUCTION

The development of prognostic models is one of the major tasks in clinical oncology. The identification of risk groups of patients defined by values of certain prognostic factors is of special importance. A medical decision, e.g. the application of a special treatment, may depend on the patient's prognosis. Prognostic factor studies for censored data are the basis for the development of prognostic models in oncology [1].

Recursive partitioning procedures have been used for prediction of survival probabilities in cancer research by various authors [1–3]. Several algorithms for recursive partitioning of a censored response have been suggested [4–10]. Instability and variable selection bias are well known problems of tree based methods. Lausen *et al.* [11] suggest a methodology for trees that account for variables measured on different scales. The adjusted P-value of a maximally

*Correspondence to: Berthold Lausen, Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander-University, Erlangen-Nuremberg, Waldstraße 6, D-91054 Erlangen, Germany.
†E-mail: berthold.lausen@rzmail.uni-erlangen.de

Received June 2002
Accepted May 2003

selected logrank statistic [12] is used to adjust for the variable selection bias due to different scales and allows the trees to be stabilised. Kim and Loh [13] propose multiway splits to adjust for the variable selection bias.

Bootstrap aggregation of classification and regression trees ('bagging' [14, 15]) stabilizes predictors in many applications. Bootstrapped Kaplan–Meier curves are studied by Efron [16] and Akritas [17]. Sauerbrei [18] provides a recent overview on bootstrapping in survival analysis. Dannegger [19] investigates the instability of trees with an application to survival data of node negative breast cancer patients and averages the point predictions of survival times in the nodes of multiple trees to stabilize the procedure. However, Graf *et al.* [2] claim that point predictions of event free time in cancer patients did not lead to satisfactory results in the past. Instead, the predicted survival probability function, that is, the predicted probability of being event free up to time z, proved to be a useful prediction in oncology. Here, we propose to derive the predicted survival probability function for a new patient as follows: first a set of survival trees based on B bootstrap samples of the observations is constructed, and second the bootstrap aggregated Kaplan–Meier curve of a new patient is computed for all bootstrap observations identified by the B leaves containing the new observation.

We study the improvement of predicted survival probability functions in cancer patients for two types of cancer. First, a large study on node positive breast cancer patients with seven prognostic factors [20] is used for the construction and evaluation of improved predictors. Second, we analyse prognostic factors derived from gene expression profiling data for 38 patients suffering diffuse large B-cell lymphoma [21]. The choice of an appropriate criterion to evaluate the predicted survival probability functions is not obvious. A comprehensive discussion on this issue can be found in Henderson [22]. Based on a measure originally developed for weather forecasts by Brier [23], Graf *et al.* [2] propose the integrated Brier score for censored data which we use as a measure for goodness of prediction for the breast cancer and lymphoma data sets.

The paper is structured as follows. The statistical and recursive partitioning framework is reviewed in Sections 2 and 3 as much as required in the sequel. Bootstrap aggregated survival trees are introduced in Section 4. Section 5 comments on the integrated Brier score and we investigate the possible rates of improvement by means of a simulation study in Section 6. The gain achieved for predicted survival probability functions in breast cancer and lymphoma patients by bagging survival trees is analysed in Section 7.

2. STATISTICAL MODEL

Throughout this paper Z^0 denotes the true survival time and C denotes the true censoring time with distribution functions F and G, respectively. We observe $Z = \min(Z^0, C)$, the time under observation until either an event or censoring occurs. The variable $\delta = I(Z^0 \leqslant C)$ denotes the indicator for an observed event, called the 'censoring indicator' and $\mathbf{X} = (X_1, \ldots, X_p)$ denotes a set of p covariables from a sample space χ. The covariables can be measured at different scales. The conditional distribution function of true survival time given the covariables $\mathbf{X}$ is denoted by $F_{\mathbf{X}}$. We observe a learning sample $\mathscr{L} = \{(z_i, \delta_i, \mathbf{x}_i); i = 1, \ldots, N\}$ of N independent observations consisting of p-dimensional covariables $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, survival time z_i and censoring indicator δ_i. The observations in the learning sample are independent and identically distributed random samples from the joint distribution function H of time under observation, censoring indicator and covariables:

$$(z_1, \delta_1, \mathbf{x}_1), \ldots, (z_N, \delta_N, \mathbf{x}_N) \overset{\text{iid}}{\sim} H$$

The marginal probability of being event free up to a time $z > 0$ is $S(z) = P(Z^0 > z) = 1 - F(z)$. The Kaplan–Meier product limit estimator [24] can be used to estimate $S(\cdot)$. The Kaplan–Meier curve computed using observations from a learning sample $\mathscr{L}$ is denoted by $\hat{S}_{\mathscr{L}}(\cdot)$. We assume random censorship, that is the independence of Z^0 and C given the covariables $\mathbf{X}$.

The 'patient specific' survival probability function, i.e. the probability of being event free up to time z, conditional on the vector of covariables $\mathbf{X} = \mathbf{x}$ is

$$S(z|\mathbf{x}) = P(Z^0 > z|\mathbf{X} = \mathbf{x}) = 1 - F_{\mathbf{X}}(z)$$

With $S(\cdot|\mathbf{x})$ we denote the true conditional survival function. For a new patient with covariable status $\mathbf{x}_{\text{new}}$, the predicted survival probability function $\hat{S}_{\mathscr{L}}(z|\mathbf{x}_{\text{new}})$ of being event free up to time z estimated from a learning sample $\mathscr{L}$ is of special interest.

3. SURVIVAL TREES

We review parts of the terminology of binary trees following the framework given in Breiman *et al.* [25]. A binary tree is a set of q nodes and their edges. The nodes t_j are subsets of the sample space χ. Based on a learning sample $\mathscr{L}$, a splitting and a stopping rule, a tree $T(\mathscr{L}) = \{1, \ldots, q\}$ is constructed. The elements of $T(\mathscr{L})$ represent the nodes by their indices. The set of terminal nodes or leaves, i.e. nodes that are not splitted, is denoted by a subset of the tree $\tilde{T}(\mathscr{L}) \subset T(\mathscr{L})$. The leaves are disjoint partitions of the whole sample space χ of the covariables:

$$\chi = \bigcup_{j \in \tilde{T}(\mathscr{L})} t_j \text{ and } t_j \bigcap t_k = \emptyset \quad \text{for all leaves } j \neq k \in \tilde{T}(\mathscr{L})$$

In addition, we define a function which identifies the leaf of a tree $T(\mathscr{L})$ for an observation with covariable status $\mathbf{x}$:

$$\tau(\mathbf{x}; \mathscr{L}) = t_j \quad \text{where } j \in \tilde{T}(\mathscr{L}) \text{ and } \mathbf{x} \in t_j \tag{1}$$

A predictor of the conditional survival function for a new patient with covariables $\mathbf{x}_{\text{new}}$ is the Kaplan–Meier curve based on all observations from the learning sample $\mathscr{L}$ which are part of the same leaf as $\mathbf{x}_{\text{new}}$ itself:

$$\hat{S}_{\mathscr{L}}(\cdot|\mathbf{x}_{\text{new}}) = \hat{S}_{\mathscr{L}(\mathbf{x}_{\text{new}})}(\cdot)$$

$\mathscr{L}(\mathbf{x}_{\text{new}})$ denotes a subset of the learning sample, i.e. the subset of observations from $\mathscr{L}$ within the same leaf of the tree $T_{\mathscr{L}}$ as an observation with covariables $\mathbf{x}_{\text{new}}$:

$$\mathscr{L}(\mathbf{x}_{\text{new}}) = \{(z_i, \delta_i, \mathbf{x}_i) \in \mathscr{L} | \mathbf{x}_i \in \tau(\mathbf{x}_{\text{new}}; \mathscr{L})\} \tag{2}$$

In general, two approaches are used for building survival trees. The first one uses a measure of within-node homogeneity, for example, the distance between Kaplan–Meier estimates of the survival curves [4] or a measure based on Poisson deviance residuals [8]. The second approach is based on a measure of between-node separation by using a test statistic to distinguish between survival times [6, 9, 11]. In this paper, we do not focus on special splitting or stopping criteria but define bootstrap aggregated (bagged) survival trees for arbitrary tree growing algorithms. For the computations in Sections 6 and 7, a criterion equivalent to the one suggested by LeBlanc and Crowley [8] is used.

4. BAGGING SURVIVAL TREES

Breiman [15] calls a tree unstable if small perturbations in the learning sample $\mathscr{L}$ induce a large change in the resulting predictor $T(\mathscr{L})$. A more formal definition of stability is given by Bühlmann and Yu [26]. A stable predictor converges to some fixed value as the sample size N tends to infinity, whereas an unstable predictor does not. The stability of a predicted survival probability function derived from a survival tree may be affected by small learning samples, a large number of covariables or a small effect to noise ratio. The aggregation of multiple unstable predictors leads to a stabilization in many classification and regression problems. Here, we propose an aggregation scheme for survival trees.

Under a completely specified statistical model we are able to draw infinitely many learning samples $\mathscr{L}^1, \mathscr{L}^2, \ldots$ of size N from H. We are interested in an estimate of the true conditional survival function $S(\cdot|\mathbf{x}_{\text{new}})$, where $\mathbf{x}_{\text{new}}$ is a fixed vector of p covariables. It is therefore natural to compute an estimate $\hat{S}(\cdot|\mathbf{x}_{\text{new}})$ based on observations with covariable values that are 'close' to $\mathbf{x}_{\text{new}}$, i.e. observations which are elements of the same leaf of a survival tree as $\mathbf{x}_{\text{new}}$ itself. Consequently, for each of those learning samples, a survival tree $T(\mathscr{L}^b), b = 1, 2, \ldots$ is constructed. The subset of the learning sample $\mathscr{L}^b(\mathbf{x}_{\text{new}})$ as defined in (2) is the set of all observations from $\mathscr{L}^b$ which are elements of the leaf $\tau(\mathbf{x}_{\text{new}}; \mathscr{L}^b)$. Now, we have a number of subgroups $\mathscr{L}^b(\mathbf{x}_{\text{new}}), b = 1, 2, \ldots$ and join their matrix representations into the aggregated sample $\mathscr{L}_A$

$$\mathscr{L}_A(\mathbf{x}_{\text{new}}) = [\mathscr{L}^1(\mathbf{x}_{\text{new}}); \mathscr{L}^2(\mathbf{x}_{\text{new}}); \ldots]$$

An aggregated estimator of $S(\cdot|\mathbf{x}_{\text{new}})$ is the Kaplan–Meier curve for the observations of the aggregated subgroups

$$\hat{S}_A(\cdot|\mathbf{x}_{\text{new}}) = \hat{S}_{\mathscr{L}_A(\mathbf{x}_{\text{new}})}(\cdot)$$

In contrast to aggregated classification trees [14, 15], where the predictions of each tree are aggregated by majority voting, we aggregate the observations from each leaf directly and compute one single predictor for the aggregated sample only.

Having one learning sample $\mathscr{L}$ of size N we estimate the aggregated survival function $\hat{S}_A(\cdot|\mathbf{x}_{\text{new}})$ using the bootstrap. First, we consider bootstrap learning samples

$$\mathscr{L}^{*(b)} = \{(z_i^{*(b)}, \delta_i^{*(b)}, \mathbf{x}_i^{*(b)}); i = 1, \ldots, N\}, \;\; b = 1, 2, \ldots$$

from the empirical joint distribution $\hat{H}$. The subsets of the bootstrap samples

$$\mathscr{L}^{*(b)}(\mathbf{x}_{\text{new}}) = \{(z_i^{*(b)}, \delta_i^{*(b)}, \mathbf{x}_i^{*(b)}) \in \mathscr{L}^{*(b)} | \mathbf{x}_i^{*(b)} \in \tau(\mathbf{x}_{\text{new}}; \mathscr{L}^{*(b)})\}$$

are the subsets of observations from $\mathscr{L}^{*(b)}$ within the same leaf as $\mathbf{x}_{\text{new}}$. Here $\tau(\mathbf{x}_{\text{new}}; \mathscr{L}^{*(b)})$ refers to a leaf of the survival tree $T(\mathscr{L}^{*(b)})$, constructed on the bootstrap sample $\mathscr{L}^{*(b)}$ as defined in (1).

A bootstrap aggregated version of the estimated conditional survival function $\hat{S}$ can be defined as

$$\hat{S}_A^*(\cdot|\mathbf{x}_{\text{new}}) = \hat{S}_{\mathscr{L}_A^*(\mathbf{x}_{\text{new}})}(\cdot)$$

i.e. the Kaplan–Meier curve based on the bootstrap aggregated subgroups

$$\mathscr{L}_A^*(\mathbf{x}_{\text{new}}) = [\mathscr{L}^{*(1)}(\mathbf{x}_{\text{new}}); \mathscr{L}^{*(2)}(\mathbf{x}_{\text{new}}); \ldots]$$

The bootstrap aggregated conditional survival function $\hat{S}_A^*(\cdot|\mathbf{x}_{\text{new}})$ is approximated by a finite number B of bootstrap learning samples as follows:

(1) Draw B bootstrap samples of size N with replacement from $\mathscr{L}$ and denote them by $\mathscr{L}^{*(1)}, \ldots, \mathscr{L}^{*(B)}$.
(2) Construct a survival tree $T(\mathscr{L}^{*(b)})$ based on each bootstrap sample $\mathscr{L}^{*(b)}$.
(3) Compute the bootstrap aggregated survival function for a new observation $\mathbf{x}_{\text{new}}$ by

$$\hat{S}_A^B(\cdot|\mathbf{x}_{\text{new}}) = \hat{S}_{\mathscr{L}_A^B(\mathbf{x}_{\text{new}})}(\cdot)$$

where

$$\mathscr{L}_A^B(\mathbf{x}_{\text{new}}) = [\mathscr{L}^{*(1)}(\mathbf{x}_{\text{new}}); \mathscr{L}^{*(2)}(\mathbf{x}_{\text{new}}); \ldots; \mathscr{L}^{*(B)}(\mathbf{x}_{\text{new}})]$$

5. GOODNESS OF PREDICTION

In the classification and regression framework, the goodness of prediction is measured by the misclassification error or mean squared error. However, there is no obvious goodness of prediction criterion for predicted survival probability functions. Several proposals have been studied [27, 28], see Henderson [22] for a review. We use the integrated Brier score for censored data as introduced by Graf *et al.* [2]. $\hat{G}(z)$ denotes the Kaplan–Meier estimate of the censoring distribution, that is the Kaplan–Meier estimate based on observations $(z_i, 1-\delta_i)$, $i=1,\ldots,N$ (in the absence of ties between censored and uncensored observations). The Brier score as a function of time $z>0$ is defined by

$$\text{BS}(z) = \frac{1}{N}\sum_{i=1}^{N} (\hat{S}(z|\mathbf{x}_i)^2 I(z_i \leqslant z \wedge \delta_i = 1)\hat{G}(z_i)^{-1} + (1-\hat{S}(z|\mathbf{x}_i))^2 I(z_i > z)\hat{G}(z)^{-1})$$

and the integrated Brier score is given by

$$\text{IBS} = \max(z_i)^{-1} \int_0^{\max(z_i)} \text{BS}(z)\,\mathrm{d}z$$

We use cross-validation to estimate the integrated Brier score for the data in Section 7.

For the simulation experiments in the next section, where the true conditional survival function $S(z|\mathbf{x})$ is known, we measure the distance between the predicted and true conditional survival probability function by a similar criterion. In analogy to the Brier score, the squared difference between the two curves is integrated with respect to time and averaged over all observations (mean integrated squared error, MIE):

$$\text{MIE} = \frac{1}{N}\sum_{i=1}^{N}(z_i^*)^{-1}\int_0^{z_i^*}(\hat{S}_{\mathscr{L}}(z|\mathbf{x}_i) - S(z|\mathbf{x}_i))^2\,\mathrm{d}z$$

where z_i^* denotes the time of the last event in $\mathscr{L}(\mathbf{x}_i)$.

6. SIMULATION STUDY

6.1. Setup

The rate of improvement with respect to prediction error by bagging survival trees compared to single survival trees is investigated in two simulation setups. The first setup is similar to configurations used in LeBlanc and Crowley [9] or Keleş and Segal [10]. Five independent predictors $X_1,\dots,X_5$ are uniformly distributed on $[0,1]$. Survival times are exponentially distributed with conditional survival function $S(z|\mathbf{x}) = \exp(-z\phi_{\mathbf{x}})$ according to three models with logarithms of the hazards $\vartheta_{\mathbf{x}} = \log(\phi_{\mathbf{x}})$:

$$\begin{aligned} &\text{A:} \quad \vartheta_{\mathbf{x}} = 0 \\ &\text{B:} \quad \vartheta_{\mathbf{x}} = 3I(X_1 \leqslant 0.5 \cap X_2 > 0.5) \\ &\text{C:} \quad \vartheta_{\mathbf{x}} = 3X_1 + X_2 \end{aligned}$$

In model B, two risk groups are defined in terms of a tree with three leaves induced by the predictors X_1 and X_2. The risk in model C depends on a linear combination of X_1 and X_2. Censoring times are distributed uniformly on $[0,\gamma]$. Uncensored learning samples as well as learning samples with approximately 25 and 50 per cent censored observations are used, the sample size is $N = 200$. See Table I for values of the censoring parameter γ used in the different setups.

In the second setup we study the performance of bagging survival trees for seven risk groups defined by the predictors $X_1,\dots,X_5$ from above. The underlying tree is displayed in Figure 1. Different numbers of non-informative covariables are added to the five predictors in the learning sample (none, 10 and 20), each of them uniformly distributed on $[0,1]$. The number of observations in the learning sample is $N = 200$, approximately 50 per cent of them censored. The logarithms of the hazards of the risk groups are scaled by an additional

Table I. Values of the censoring parameter γ used in the simulations.

Censoring	Setup 1			Setup 2		
(per cent)	A	B	C	$c = 0.5$	$c = 1$	$c = 2$
25	4.02	2.87	0.68	—	—	—
50	1.59	0.94	0.21	0.90	0.51	0.18

Censoring times are uniformly distributed on $[0,\gamma]$ with γ computed depending on the model and amount of censoring.

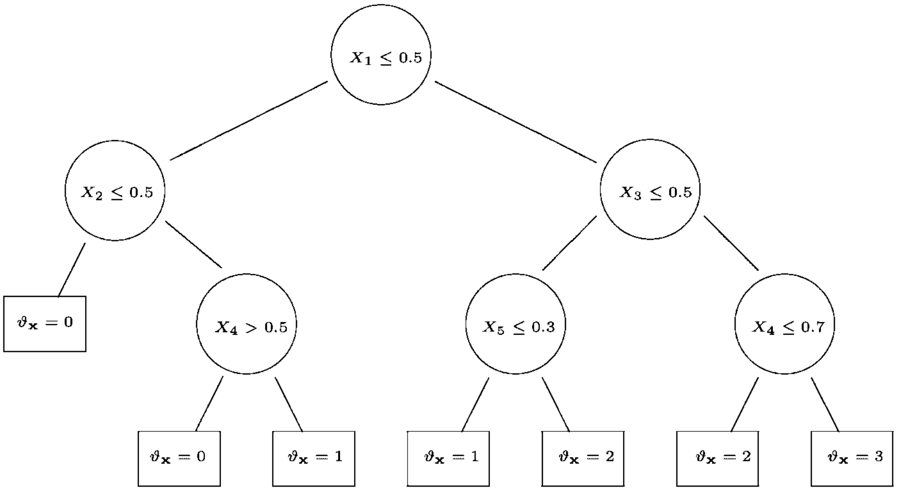
$X_1 \leq 0.5$
$X_2 \leq 0.5$
$X_3 \leq 0.5$
$\vartheta_{\mathbf{x}} = 0$
$X_4 > 0.5$
$X_5 \leq 0.3$
$X_4 \leq 0.7$
$\vartheta_{\mathbf{x}} = 0$
$\vartheta_{\mathbf{x}} = 1$
$\vartheta_{\mathbf{x}} = 1$
$\vartheta_{\mathbf{x}} = 2$
$\vartheta_{\mathbf{x}} = 2$
$\vartheta_{\mathbf{x}} = 3$

Figure 1. Tree model: seven risk groups based on five informative covariables $X_1,\dots,X_5$.

parameter $c>0$ and the conditional survival function is

$$S(z|\mathbf{x}) = \exp(-z\phi_{\mathbf{x},c}) \quad \text{with } c\cdot\vartheta_{\mathbf{X}} = \log(\phi_{\mathbf{x},c})$$

Since the conditional survival function $S(z|\mathbf{x})$ is known in simulation experiments, the mean integrated squared error as a measurement of goodness of prediction is computed by numerical integration. The goodness of prediction for single and aggregated trees is evaluated using an independent test sample of size 100. In addition, the mean integrated squared error is reported for the simple Kaplan–Meier curve, i.e. an estimator without knowledge of the information in the covariables. We report the median of the mean integrated squared error as well as the interquartile range for 1000 Monte-Carlo replications of learning and test sample in Tables II and III.

The rpart package [29] in the R system for statistical computing [30] is used for the construction of survival trees. The splitting criterion implemented in rpart is equivalent to the one described in LeBlanc and Crowley [8]. Single trees are pruned to an appropriate size in order to avoid instability induced by over sized trees. For models A–C, splits that improve the full likelihood deviance by an amount of less than 10 per cent are removed from the tree. For the tree model, an amount of 2 per cent improvement is required for a split to be done; the average number of nodes of the single survival trees is reported in Table III. In contrast, each of the $B=50$ trees for bagging is relatively large. We stop the tree growth when less than 20 observations are element of a node or if a split does not improve the full likelihood deviance by an amount of at least 1 per cent. For models A and B, less extreme trees are used for aggregation: splits with an improvement of less than 10 per cent (model A) and 5 per cent (model B) are removed from each of the multiple trees. Those values are obtained by independent simulation experiments.

Table II. Median of mean integrated squared errors (×100 for model A and ×10 for B and C) for the three models of the first setup for 1000 pairs of learning and test samples of size 200 and 100.

		Censoring		
Model		None	25 per cent	50 per cent
A	Kaplan–Meier	0.031 (0.037)	0.096 (0.104)	0.127 (0.151)
	Survival tree	0.031 (0.038)	0.096 (0.104)	0.127 (0.151)
	Bagging	0.032 (0.037)	0.097 (0.107)	0.132 (0.160)
	Relative improvement (per cent)	−6	−1	−4
B	Kaplan–Meier	0.152 (0.045)	0.314 (0.053)	0.727 (0.114)
	Survival tree	0.053 (0.068)	0.077 (0.085)	0.099 (0.106)
	Bagging	0.026 (0.021)	0.054 (0.039)	0.096 (0.070)
	Relative improvement (per cent)	50	30	3
C	Kaplan–Meier	0.130 (0.056)	0.392 (0.061)	0.581 (0.068)
	Survival tree	0.109 (0.039)	0.204 (0.054)	0.231 (0.070)
	Bagging	0.051 (0.023)	0.115 (0.049)	0.183 (0.074)
	Relative improvement (per cent)	54	44	21

Interquartile range given in parentheses. Bagging with $B = 50$ bootstrap samples.

Table III. Median of mean integrated squared error (×10) for the tree model of the second setup for 1000 pairs of learning and test samples of size 200 and 100.

	$c = 0.5$	$c = 1$	$c = 2$
Kaplan–Meier	0.234 (0.024)	0.697 (0.048)	1.482 (0.072)
Five informative covariables only ($p = 5$)			
Median number of nodes	6	6	6
Survival tree	0.363 (0.215)	0.359 (0.166)	0.318 (0.149)
Bagging	0.226 (0.071)	0.246 (0.079)	0.230 (0.098)
Relative improvement (per cent)	38	31	28
10 Non-informative covariables added ($p = 15$)			
Median number of nodes	10	9	7
Survival tree	0.689 (0.236)	0.587 (0.219)	0.397 (0.200)
Bagging	0.231 (0.085)	0.281 (0.083)	0.281 (0.110)
Relative improvement (per cent)	66	52	29
20 Non-informative covariables added ($p = 25$)			
Median number of nodes	12	10	8
Survival tree	0.831 (0.227)	0.705 (0.224)	0.457 (0.230)
Bagging	0.225 (0.088)	0.282 (0.089)	0.300 (0.121)
Relative improvement (per cent)	73	60	34

Interquartile range given in parentheses. Bagging with $B = 50$ bootstrap samples. In addition, the median number of nodes in each of the single trees in reported.

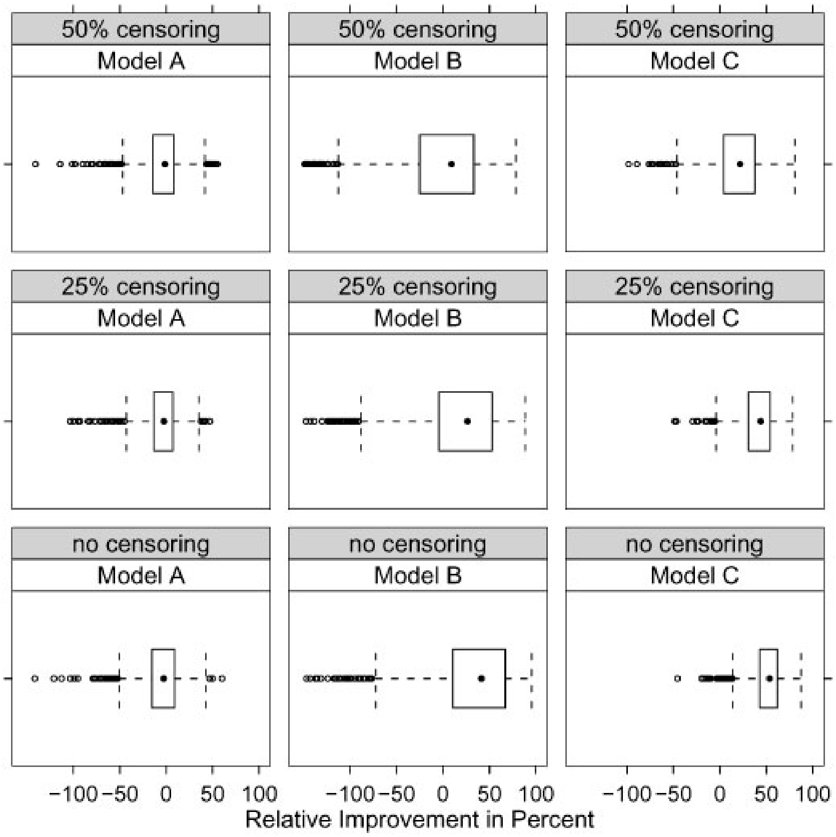

Figure 2. Boxplots of the relative improvement of mean integrated squared error of 1000 simulation runs for models A–C and three levels of censoring. The relative improvement is given by the difference of the error of single survival trees and bagging relative to the error of survival trees. The learning samples are of size $N = 200$. One hundred and fifty five extreme values with less then -150 per cent improvement are not shown here.

6.2. Results

In all simulation setups described above, the mean integrated squared error of bagging survival trees is either better or as good as the error for single survival trees. The relative reduction depends on the model, the amount of censoring and the number of non-informative covariables in the learning sample.

For model A, where the survival time does not depend on any of the covariables, both single survival trees and bagging are as good as the Kaplan–Meier curve of the learning sample. The majority of the single trees are trivial, i.e. no split. The trees for bagging are pruned to a medium depth.

The relative improvement by bagging compared to single trees ranges between 3 and 50 per cent for model B. The single survival trees usually split into three terminal nodes and therefore are able to identify the underlying tree structure and improve upon the simple Kaplan–Meier curve. Aggregating multiple trees leads to further improvement: the difference between the mean integrated squared error of a single survival tree and bagging relative to the error of a single survival tree is up to 50 per cent smaller for bagging. Moreover, the variance of the prediction error estimated by the interquartile range is much smaller for bagging.

When the risk depends on a linear combination of covariables as in model C, the relative improvement is between 21 and 54 per cent. The single trees are not able to search in linear combinations while the aggregation of multiple large trees is able to adapt itself to this situation. Figure 2 provides a graphical representation of the relative gain by bagging survival trees for models A–C.

In the second setup, where the risk groups depend on the tree displayed in Figure 1, the separation between the risk groups is scaled by $c = 0.5, 1$ and 2. The sample size of $N = 200$ is relatively small for a complex tree. The amount of 50 per cent censoring is relatively large but this setup is likely to be relevant in practical situations. The effect of instability is studied by adding up to 20 non-informative covariables to the learning sample. In contrast to unbagged survival trees, the goodness of fit of bagged survival trees is scarcely affected by the number of non-informative covariables (0.226 for five informative covariables compared to 0.231 with 10 and 0.225 with 20 additional non-informative covariables, $c = 0.5$). The integrated squared error of single survival trees is always larger than that of the Kaplan–Meier curve (0.234), i.e. the prediction is even worse. However, bagging performs comparable to the Kaplan–Meier curve for $c = 0.5$. Detailed results, including the median number of terminal nodes of the single trees, are given in Table III. The number of leaves ranges between 6 and 12 and shows that the single trees are able to identify the tree shown in Figure 1.

If the risk groups are better separated by $c = 1$ or 2, survival trees improve upon the Kaplan–Meier curve. The relative improvement for bagging compared to survival trees ranges from 28 per cent ($c = 2$, 5 informative covariables only) to 73 per cent ($c = 1$, 20 non-informative covariables). Boxplots of the relative improvement for the tree model in all three situations under test are displayed in Figure 3. It should be noted that both the median prediction error and its variance are reduced by bagging, indicating the effect of stabilisation achieved here.

7. DATA

The prediction of survival probability functions based on the values of certain prognostic factors is a major issue in oncology. We therefore investigate the gain of bagging survival trees for data of breast cancer patients with positive lymph nodes and for gene expression profiling data of patients with diffuse large B-cell lymphoma (DLBCL). The learning sample in the breast cancer study has 686 observations. In contrast, we analyse a small data set of 38 patients suffering from DLBCL.

7.1. Breast cancer: GBSG-2 study

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 [20]. Patients not older than 65 years with positive regional lymph nodes but no distant metastases were included in the study. Complete data of

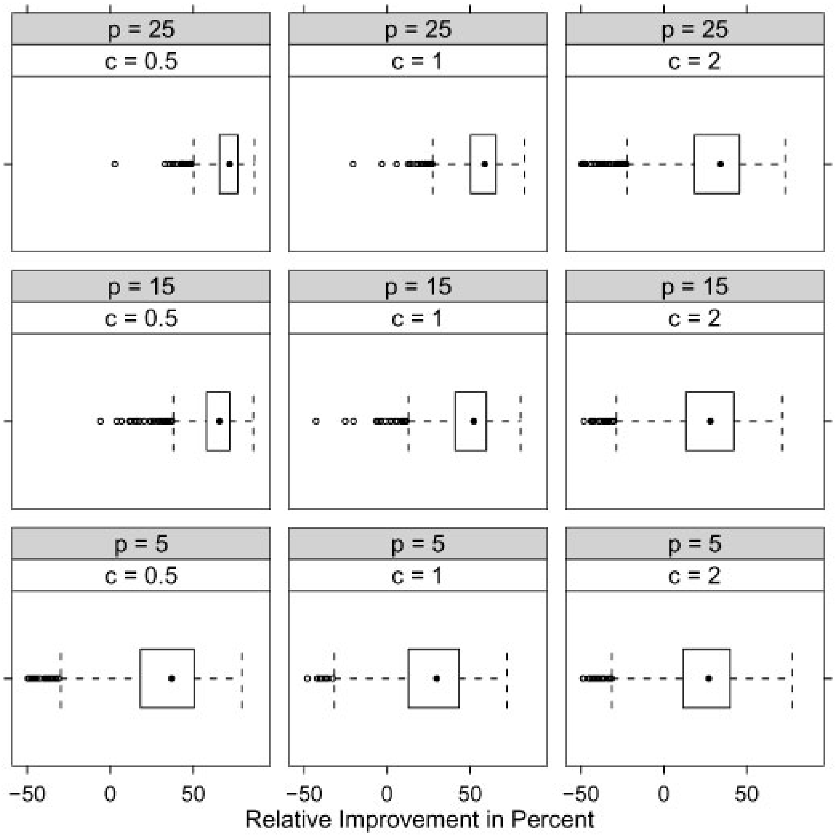

Figure 3. Boxplots of the relative improvement of mean integrated squared error of 1000 simulation runs for the tree model (cf. Figure 1). The hazards are scaled by three different values of c and different numbers of covariables p are investigated. The learning samples are of size $N = 200$ and roughly 50 per cent of the observations are censored. One hundred and sixty seven extreme values with less then -50 per cent improvement are not shown here.

seven prognostic factors of 686 women are used in Sauerbrei and Royston [31] for prognostic modelling, the data set is available online at `http://www.blackwellpublishers.com/rss/`.

Observed hypothetical prognostic factors are age, menopausal status, tumour size, tumour grade, number of positive lymph nodes, progesterone receptor, estrogen receptor and the information of whether or not a hormonal therapy was applied. Table IV gives characteristics of the patients in the study.

Survival trees using p-value adjusted logrank statistics [11] are used for the evaluation of prognostic factors in the GBSG-2 study by Schumacher *et al.* [1]. The goodness of prediction of several prognostic models including survival trees for a similar study on node negative breast cancer patients is investigated in Graf et al. [2]. Ten independent runs of the 10-fold cross-validated integrated Brier score are averaged and reported here. The single survival trees are pruned to trees with, on average, six terminal nodes. For bagging, $B = 50$ unpruned large trees are used. The integrated Brier score for survival trees is 0.173. Bagging survival trees

Table IV. Patient characteristics in GBSG2-Study.

Prognostic factor	Median	First and	Third quartile
Age (years)	53	46	61
Tumor size (mm)	25	20	35
No. of positive lymph nodes	3	1	7
Progesterone receptor (fmol)	32.5	7	131.75
Estrogen receptor (fmol)	36	8	114
	Levels	No. of observations	
Menopausal status	Pre	290	
	Post	396	
Tumor grade	I	81	
	II	444	
	III	161	
Hormonal therapy	Yes	246	
	No	440	

For ordered covariables, the median as well as the first and third quartile are given whereas the number of observations for each level is reported for unordered factors.

improve the accuracy by 6 per cent: the integrated Brier score for bagged survival trees is 0.163. The integrated Brier score when the overall Kaplan–Meier curve is used as predictor for each observation is 0.189.

7.2. DLBCL

Using different expression patterns of the genome it is an interesting and straightforward hypothesis to assess types of a disease on the molecular level. Alizadeh *et al.* [21] analyse the overall survival of patients with diffuse large B-cell lymphoma (DLBCL) and show that it is possible to identify two distinct types of DLBCL by gene expression profiling. They use a microarray with genes that are expressed in lymphoid cells and genes with known or suspected roles in processes important in immunology and cancer. Overall they use 17 856 cDNA clones and 96 normal and malignant lymphocyte tissue samples. The first step of the data analysis (cf. Eisen [32]) is to identify each spot on the microarray and to classify the spot pixels and the background pixels. The ratio of the red and green fluorescence can be estimated. We use the simple robust estimate MRAT of Eisen *et al.* [33] which is defined as the median of the ratio for each spot pixel. Each spot value is centred by the median of the background pixels. The data is available at `http://llmpp.nih.gov/lymphoma/`. We use the mean expression values for genes which are expressed by two or more different cDNA clones. Restricting the analysis to complete data, i.e. expression spot available for each combination of gene and tissue sample and survival times observed, we analyse data of 7680 marker genes and 38 tissue samples. Analysing the response lymphoma classification, Dudoit *et al.* [34] use 81 tissue samples to compare classification trees, bagged classification trees and other discrimination methods.

We derive hypothetical prognostic factors by using 10 clusters estimated with the agglomerative average linkage hierarchical cluster analysis. We compute the Euclidean distance for the logarithm of the MRAT values $(\text{MRAT}+0.01)$ for the marker genes. We analyse a partition of the marker genes in 10 clusters, which is defined by a cluster level of the dendrogram resulting in a partition of 10. The mean gene expression for each cluster is the computed value of each prognostic factor. Because of the small number of observations, the nodes of the single and multiple trees are allowed to split if they contain more than four observations. The integrated Brier score of the overall Kaplan–Meier curve itself is 0.231. We obtain the integrated Brier score 0.225 for survival trees and 0.233 for our proposal of bagging survival trees (with $B=50$ bootstrap samples): neither single survival trees nor bagging lead to an improvement with respect to the integrated Brier score in this situations. If the International Prognostic Index is used as additional covariable, the integrated Brier score is 0.279 for survival trees and 0.214 for bagged survival trees, which is a reduction of about 25 per cent.

8. DISCUSSION

We suggest a method to aggregate survival trees. In contrast to aggregation by majority voting or averaging of the predictions in classification or regression problems, averaged point predictions, e.g. the median of an estimated survival probability function, are of minor interest in clinical oncology [2]. Instead, the predicted conditional survival probability function for a new patient is more informative. Therefore, we do not aggregate point predictions but predict the conditional survival probability function by computing one single Kaplan–Meier curve based on observations identified by the leaves of B bootstrap survival trees. Although aggregation leads to improved predictions for cancer patients as shown in this paper, it is a ‘black box’ of multiple trees. As a remedy, graphical methods like importance plots [25] are used for aggregated predictors by Breiman [35] and Friedman [36] and may be used for bagged survival trees as well. However, our predictor is based on similar patients and the degree of similarity can be described by repetitions of a patient from a learning sample in the aggregated set.

Bagging survival trees depends on two parameters: the number of bootstrap samples B and the size of the multiple trees. The mean integrated squared errors for different numbers of bootstrap samples for one configuration of the tree model as given in Figure 4 indicate that more than $B=50$ trees do not lead to further improvements here.

Except for the artificial models A and B, we use rather large trees for aggregation. However, the choice of the appropriate tree size, i.e. small, medium or large, for bagging survival trees remains a matter of debate.

For survival trees, the estimated conditional survival function is consistent as the number of observations tends to infinity as shown by LeBlanc and Crowley [9]. For a different resampling plan (‘subagging’: m-out-of-N without replacement), Bühlmann and Yu [26] show analytically that subagging reduces the mean squared error due to smoothing split points in regression trees. Although the general setup is the same, it is unclear how predictions for censored data can fit into this framework.

The procedure suggested in this paper is implemented in the `ipred` package [37]. Besides bagging survival trees, the package implements functions for the computation of the Brier score and its integrated version and includes example calculations for the data used in this paper. The `ipred` package is available at `http://CRAN.R-project.org`.

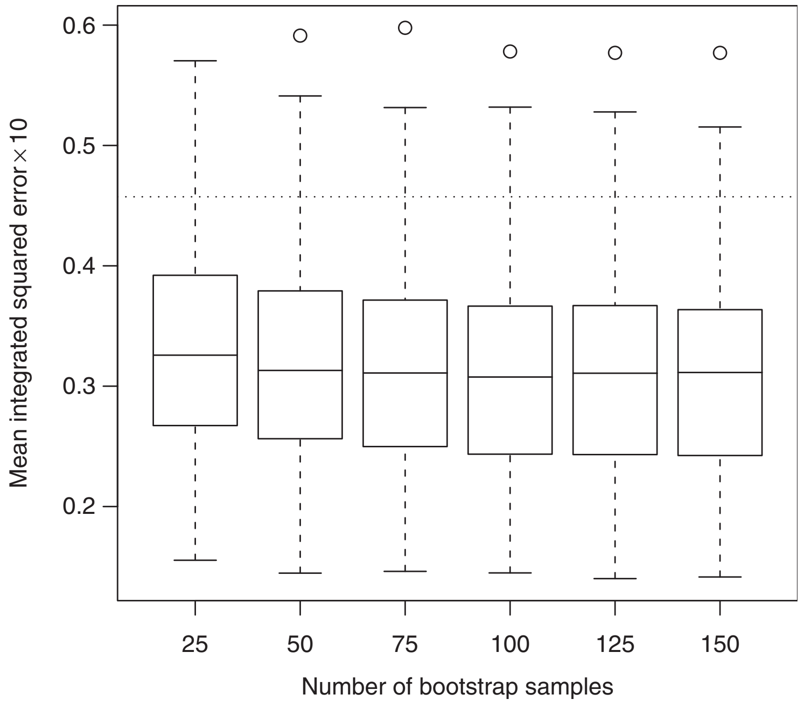

Figure 4. Boxplots of the mean integrated squared errors (×10) for the tree model ($c = 2$, $p = 25$, i.e. 20 non-informative variables) for different values of B. The dotted line gives the estimated error for a single survival tree.

ACKNOWLEDGEMENTS

Comments by an anonymous referee helped us to improve the design of the simulation experiments and the presentation of the results. We thank Terry Therneau, Robin Henderson, Norbert Höllander, Willi Sauerbrei, Martin Schumacher and Werner Vach for fruitful discussions of the paper and underlying ideas. Hints by Janice Hegewald led to an improved readability.

REFERENCES

1. Schumacher M, Holländer N, Schwarzer G, Sauerbrei W. Prognostic factor studies. In *Statistics in Oncology*, Crowley J (ed.). Marcel Dekker: New York, Basel, 2001; 321–378.
2. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* 1999; **18**:2529–2545.
3. LeBlanc M. Tree-based methods for prognostic stratification. In *Statistics in Oncology*, Crowley J (ed.). Marcel Dekker: New York, Basel, 2001; 457–472.
4. Gordon L, Olshen R. Tree-structured survival analysis. *Cancer Treatment Reports* 1985; **69**:1065–1069.
5. Ciampi A, Chang C, Hogg S, McKinney S. Recursive partition: a versatile method for exploratory data analysis in biostatistics. In *Proceedings from Joshi Festschrift*, Umphrey G (ed). North-Holland: Amsterdam, 1987; 23–50.
6. Segal MR. Regression trees for censored data. *Biometrics* 1988; **44**:35–47.
7. Davis RB, Anderson JR. Exponential survival trees. *Statistics in Medicine* 1989; **8**:947–961.
8. LeBlanc M, Crowley J. Relative risk trees for censored survival data. *Biometrics* 1992; **48**:411–425.
9. LeBlanc M, Crowley J. Survival trees by goodness of split. *Journal of the American Statistical Association* 1993; **88**:457–467.

10. Keleş S, Segal MR. Residual-based tree-structured survival analysis. *Statistics in Medicine* 2002; **21**:213–326.
11. Lausen B, Sauerbrei W, Schumacher M. Classification and regression trees (CART) used for the exploration of prognostic factors measured on different scales. In *Computational Statistics*, Dirschedl P, Ostermann R (eds). Physica-Verlag: Heidelberg, 1994; 483–496.
12. Lausen B, Schumacher M. Maximally selected rank statistics. *Biometrics* 1992; **48**:73–85.
13. Kim H, Loh WY. Classification trees with unbiased multiway splits. *Journal of the American Statistical Association* 2001; **96**:589–604.
14. Breiman L. Bagging predictors. *Machine Learning* 1996; **24**:123–140.
15. Breiman L. Arcing classifiers. *The Annals of Statistics* 1998; **26**:801–824.
16. Efron B. Censored data and the bootstrap. *Journal of the American Statistical Association* 1981; **76**:312–319.
17. Akritas MG. Bootstrapping the Kaplan–Meier estimator. *Journal of American Statistical Association* 1986; **81**:1032–1038.
18. Sauerbrei W. Bootstrapping in survival analysis. In *Encyclopedia of Biostatistics*, Armitage P, Colton T (eds). Wiley: Chichester, New York, 1998; 433–436.
19. Dannegger F. Tree stability diagnostics and some remedies for instability. *Statistics in Medicine* 2000; **19**: 475–491.
20. Schumacher M, Basert G, Bojar H, Hübner K, Olschewski M, Sauerbrei W, Schmoor C, Beyerle C, Neumann RLA, Rauschecker HF for the German breast cancer study group. Randomized 2×2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology* 1994; **12**:2086–2093.
21. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM. Distinct types of diffuse large B-cell lymphoma identified by expression profiling. *Nature* 2000; **403**:503–511.
22. Henderson R. Problems and prediction in survival-data analysis. *Statistics in Medicine* 1995; **14**:161–184.
23. Brier GW. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 1950; **78**:1–3.
24. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958; **53**:457–481.
25. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Wadsworth: California, 1984.
26. Bühlmann P, Yu B. Analyzing bagging. *The Annals of Statistics* 2002; **30**:927–961.
27. Korn EL, Simon R. Measures of explained variation for survival data. *Statistics in Medicine* 1990; **9**:487–503.
28. Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine* 2000; **19**:453–473.
29. Therneau T, Atkinson E. An introduction to recursive partitioning using the rpart routine. *Technical Report* 61, Section of Biostatistics, Mayo Clinic, Rochester, 1997.
30. Ihaka R, Gentleman R. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 1996; **5**:299–314.
31. Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society*, *Series A* 1999; **162**: 71–94.
32. Eisen M. *ScanAlyse User Manual*. Stanford University: Stanford, CA, U.S.A., 1998.
33. Eisen M, Spellman P, Brown P, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Science* 1998; **95**:14 863–14 868.
34. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumours using gene expression data. *Journal of the American Statistical Association* 2002; **97**:77–87.
35. Breiman L. Random forests. *Machine Learning* 2001; **45**:5–32.
36. Friedman JH. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics* 2001; **29**:1189–1202.
37. Peters A, Hothorn T, Lausen B. ipred: improved predictors. *R News* 2002; **2**:33–36 (ISSN 1609–3631).