Expert Systems with Applications 36 (2009) 8582–8588

Contents lists available at ScienceDirect

Expert Systems with Applications

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

The comparisons of random survival forests and Cox regression analysis with simulation and an application related to breast cancer

Imran Kurt Omurlu [a,*], Mevlut Ture [a], Füsun Tokatli [b]

[a] Trakya University, Medical Faculty, Department of Biostatistics, 22030 Edirne, Turkey
[b] Trakya University, Medical Faculty, Department of Radiation Oncology, Edirne, Turkey

ARTICLE INFO

Keywords:
Cox regression
Random survival forests
Harrell's concordance index
Survival
Breast cancer
Simulation

ABSTRACT

The objective of this study was to compare the performances of Cox regression analysis (CRA) and random survival forests (RSF) methods with simulation and a real data set related to breast cancer. In the simulations, we compared across the methods under varying sample sizes by using Monte Carlo simulation method. The results showed that the performance of the CRA was a slightly better for analysis based on Harrell's concordance index than RSF approaches based on log-rank, conservation of events, log-rank score and approximate log-rank splitting rules. In the real data application, a retrospective analysis was performed in 279 breast cancer patients diagnosed. According to Harrell's concordance index, RSF based on approximate log-rank splitting rule to determined major risk factors for disease-free survival (DFS) showed a slightly better performance than other approaches. In general, performances of all the methods were almost similar. The predictive capability of CRA can be used for different sample sizes and potential future suitable survival data problems, whereas RSF provide interpretive results.

1. Introduction

Survival analyses are a collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs. Most popular of survival analyses is Cox regression analysis (CRA). Because it is a semiparametric and a method for investigating the effect of several variables upon the time a specified event takes to happen.

Recently, random survival forests (RSF) has been used for the analysis of survival data. It is an ensemble tree method for the analysis of right censored survival data. Constructing ensembles from tree structures can be significantly improve learning performance (Ishwaran & Kogalur, 2007).

Minn et al. (2007) investigated lung metastasis gene-expression signature (LMS) that mediated experimental breast cancer metastasis selectively to the lung and was expressed by primary human breast cancer with a high risk for developing lung metastasis. However, they used RSF for determining the influence factors for the risk of metastasis of patients from the NKI-295/EMC-344 cohort (two large cohorts of early stage breast cancer patients). Ishwaran and Kogalur (2007) used RSF method for analyzing data sets from a randomized clinical trial of patients with primary biliary cirrhosis of the liver and lung cancer. They compared log-rank, conservation of events, log-rank score and approximate log-rank splitting rules on these data sets.

The purpose of this study was to compare performances of RSF and CRA according to varying sample sizes using Monte Carlo simulation and to apply CRA and RSF method for disease-free survival (DFS) in breast cancer patients.

2. Materials and methods

2.1. Cox regression analysis

The CRA is the most general of the regression models because it is not based on any assumptions concerning the nature or shape of the underlying survival distribution. The CRA is the most widely used method of survival analysis.

Survival analysis typically examines the relationship of the survival distribution to covariates. Most commonly, this examination entails the specification of a linear-like model for the log hazard. The Cox model may be written as

$$h(t,\boldsymbol{x}) = h_0(t)e^{\beta' x},$$

where $\boldsymbol{x}$ is the covariate vector, $\boldsymbol{\beta}$ is the unknown parameter vector and the $h_0(t)$ is called the baseline hazard (it is the hazard for the respective individual when all independent variable values are equal to zero). $h(t,\boldsymbol{x})$ denotes the resultant hazard, given the values of the m covariates for the respective case and the respective survival time (t). This method uses the partial likelihood to estimate

* Corresponding author. Tel.: +90 284 2357641/1631; fax: +90 284 2357652.
E-mail address: ikurt@trakya.edu.tr (I. Kurt Omurlu).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.10.023

Table 1
Simulation results for 1000 replications according to sample size.

Method		Harrell's concordance error rates (*C*-index)			
		$n = 50$	$n = 100$	$n = 250$	$n = 500$
RSF	Log-rank	0.3717	0.3258	0.2972	0.2865
	Conservation of events	0.3970	0.3568	0.3331	0.3274
	Log-rank score	0.3824	0.3324	0.3021	0.2907
	Approximate log-rank	0.3585	0.3126	0.2866	0.2783
CRA		0.2915	0.2704	0.2509	0.2474

the parameters, and parameter estimates in the method are obtained by maximizing partial likelihood function. The partial likelihood is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{exp(\boldsymbol{\beta}'\boldsymbol{x}_{(i)})}{\sum_{l \in R(t_{(i)})} exp(\boldsymbol{\beta}'\boldsymbol{x}_{(l)})},$$

where the summation in the denominator is over all subjects in the risk set at time t_i, denoted by $R(t_{(i)})$, the product is over the k distinct ordered survival times and $\boldsymbol{x}_{(i)}$ denotes the value of the covariate for the subject with ordered survival time $t_{(i)}$ (Hosmer & Lemeshow, 1999; Kleinbaum & Klein, 2005).

The CRA has two assumptions, while no assumptions are made about the shape of the underlying hazard function. First, they specify a multiplicative relationship between the underlying hazard function and the log-linear function of the covariates. The second assumption is that there is a log-linear relationship between the independent variables and the underlying hazard function (Hosmer & Lemeshow, 1999; Kleinbaum & Klein, 2005).

2.2. Random survival forests

RSF is a survival based on tree method for the analysis of right censored survival data. It based on a splitting rule and bootstrap samples. In RSF, randomization is introduced in two forms. First, a randomly drawn bootstrap sample of the data is used for growing the tree. Second, the tree learner is grown by splitting nodes on randomly selected predictors. While at first glance Random Forest might seem an unusual procedure, considerable empirical evidence has shown it to be highly effective. Standard analyses often rely on restrictive assumption such as proportional hazards. Also, with such methods there is always the concern whether associations between predictors and hazards have been modelled appropriately, and whether or not non-linear effects or higher order interactions for predictors should be included. These problems are handled seamlessly and automatically in the RSF (Breiman, 2001; Ishwaran & Kogalur, 2007). RSF uses one of the survival splitting rules (log-rank splitting, conservation of events splitting, log-rank score splitting and approximate log-rank splitting).

2.2.1. Log-rank splitting rule

The log-rank splitting rule for a split at the value c for predictor x is

Table 3
Harrell's concordance error rates for methods.

Method		Error rate
RSF	Log-rank	0.2949
	Conservation of events	0.2939
	Log-rank score	0.2998
	Approximate log-rank	0.2832
CRA		0.3003

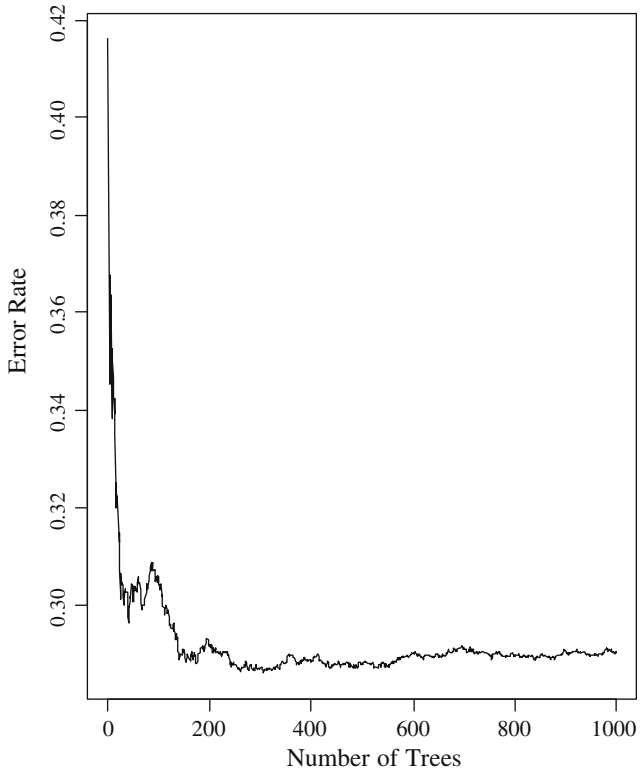

Fig. 1. Error rate of RSF for log-rank splitting rule.

$$L(x,c) = \frac{\sum_{i=1}^{n}\left(d_{i,1} - Y_{i,1}\frac{d_i}{Y_i}\right)}{\sqrt{\sum_{i=1}^{n}\frac{Y_{i,1}}{Y_i}\left(1-\frac{Y_{i,1}}{Y_i}\right)\left(\frac{Y_i-d_i}{Y_i-1}\right)d_i}},$$

where n is the number of individuals within h node, $d_{i,j}$ is the number of deaths at time t_i in the daughter nodes $j = 1, 2$, $Y_{i,j}$ is the number of individuals at risk at time t_i in the daughter nodes and $t_1 < t_2 < \cdots < t_N$ is the ordered survival time. The value $|L(x,c)|$ is the measure of node separation. The larger the value for $|L(x,c)|$, the greater the difference between the two groups, and the split is the better (Ishwaran & Kogalur, 2007).

Table 2
Estimated regression coefficients with standard errors hazard ratios with 95% confidence intervals and *p*-values from the stepwise CRA for DFS.

Independent variables		$\hat{\beta}$	Standard error	Hazard ratio	95% (CI)	*p*
Tumor size (cm)		0.248	0.040	1.281	(1.185–1.385)	<0.001
Age (years-old)		0.022	0.009	1.022	(1.003–1.041)	0.020
Pericapsular involvement of lymph nodes	Negative	0	–	1	–	
	Positive	0.700	0.225	2.015	(1.296–3.132)	0.002
Hormonal Therapy	Absent	0	–	1	–	
	Present	−0.801	0.230	0.449	(0.286–0.705)	<0.001

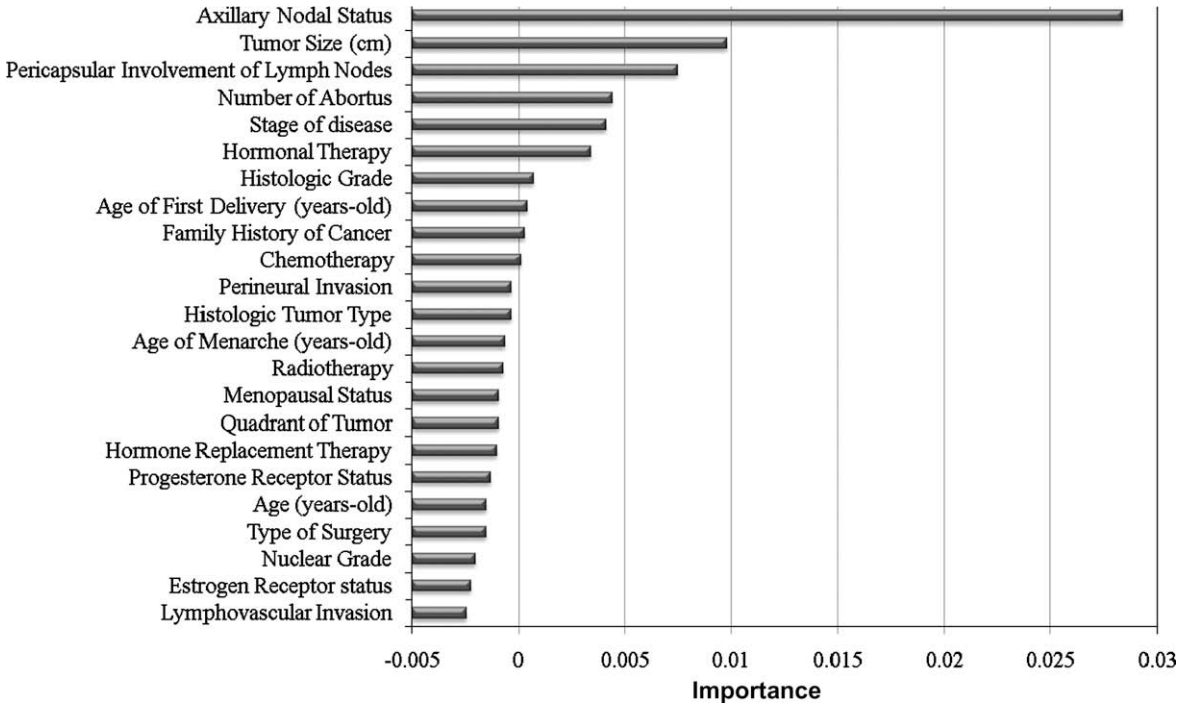

Fig. 2. Out-of-bag importance values of RSF for log-rank splitting rule.

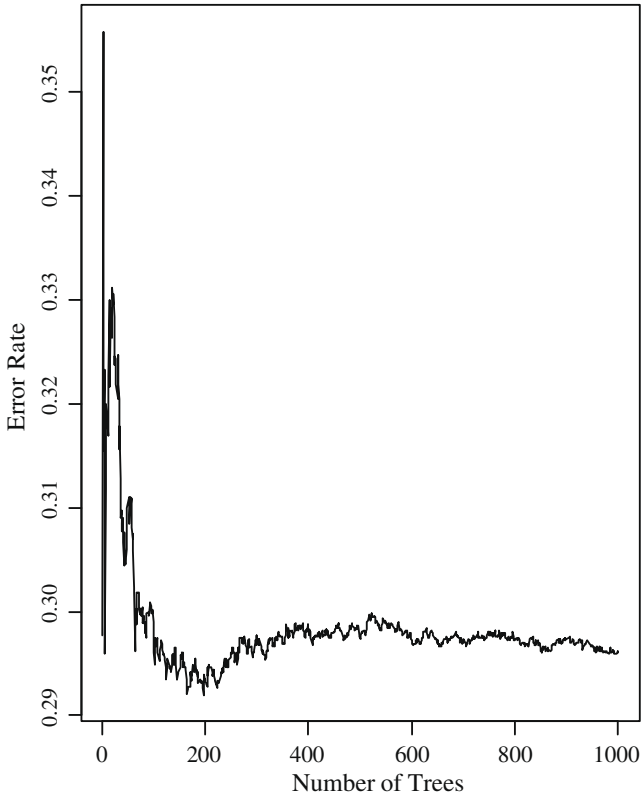

Fig. 3. Error rate of RSF for conservation of events splitting rule.

2.2.2. Conservation of events splitting rule

This splitting rule claims that the sum of the estimated cumulative hazard function over the observed time points must equal the total number of deaths. The measure of conservation of events for the split on x at the value c is

$$Conserve(x,c) = \frac{1}{Y_{1,1}+Y_{1,2}} \sum_{j=1}^{2} Y_{1,j} \sum_{k=1}^{n_j-1} M_{k,j},$$

where $M_{k,j}$ is the residuals that measure accuracy of conservation of events ($k = 1,\ldots,n_j$). This value is small if the two groups are well separated (Ishwaran & Kogalur, 2007).

2.2.3. Log-rank score splitting rule

This splitting rule is defined as

$$Score(x,c) = \frac{\sum_{x_l \leqslant c} a_l - n_1 \bar{a}}{\sqrt{n_1\left(1-\frac{n_1}{n}\right)s_a^2}},$$

where $\bar{a}$ and s_a^2 are the sample mean and sample variance ($l = 1,2,\ldots, n$), and a_l is the ranks for each survival time T_l. This rule defines the measure of node separation by $|Score(x,c)|$, and maximizing this value over x and c yields the best split (Horthorn & Lausen, 2003; Ishwaran & Kogalur, 2007).

2.2.4. Approximate log-rank splitting rule

This splitting rule can be used in the place of $L(x,c)$ to greatly reduce computations. The measure of approximate log-rank splitting rule is

$$L_A(x,c) = \frac{D^{1/2}\left(D_1 - \sum_{l=1}^{n} I\{x_l \leqslant c\}\widehat{H}(T_l)\right)}{\sqrt{\left\{\sum_{l=1}^{n} I\{x_l \leqslant c\}\widehat{H}(T_l)\right\}\left\{D - \sum_{l=1}^{n} I\{x_l \leqslant c\}\widehat{H}(T_l)\right\}}},$$

where $\widehat{H}(T_l)$ is the Nelson–Aalen estimator, $D_j = \sum_{i=1}^{N} d_{ij}$ for $j = 1, 2$, and $D = \sum_{i=1}^{N} d_i$ (Ishwaran & Kogalur, 2007).

In RSF, the importance of a variable is determined by comparing the performance of the predictor on the original data and that obtained from the out-of-bag data with the content of the variable masked off. Large positive importance values indicate informative variables while small positive or negative importance values indicate non-informative variables.

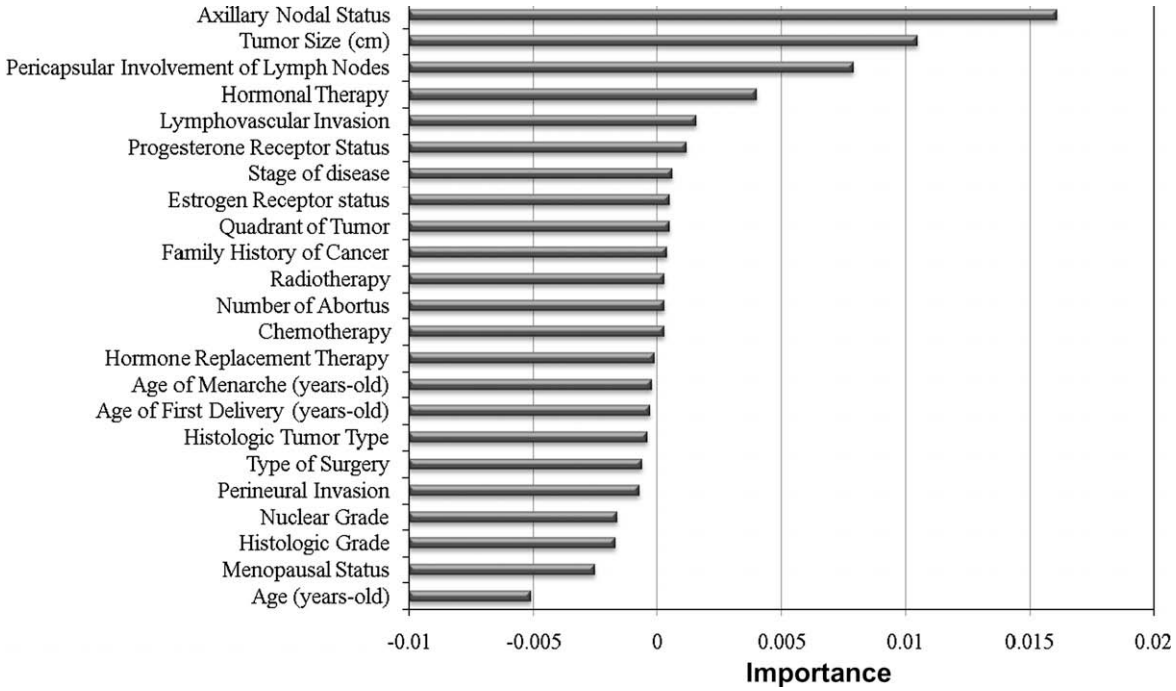

Fig. 4. Out-of-bag importance values of RSF for conservation of events splitting rule.

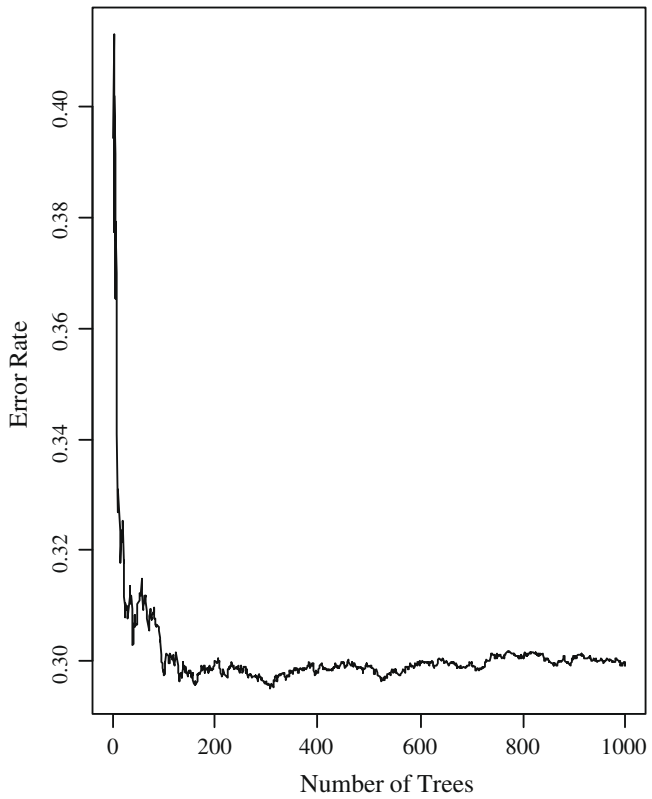

Fig. 5. Error rate of RSF for log-rank score splitting rule.

2.3. Harrell's concordance index

Harrell's concordance index (*C*-index) is a measure of survival performance. It does not depend on choosing a fixed time for evaluation of the model and specifically takes into account censoring of individuals. The error rate is computed as $1 - C$, where C is the Harrell's concordance index. Error rates are between 0 and 1, with 0.5 corresponds to a procedure doing no better than random guessing. A value of 0 is perfect accuracy (Harrell, Califf, Pryor, Lee, & Rosati, 1982; Ishwaran & Kogalur, 2007).

We used the Harrell's concordance index to quantify the accuracy of CRA and RSF.

2.4. Description of simulation methods

Our interest in this study was to compare the error rates from CRA and RSF. We performed a simulation study based on five prognostic factors. Prognostic factors were generated from different distributions related to each other. Three numeric variables were drawn from a uniform distribution on [0,1]. Two categorical variables were drawn from a binomial distribution on [2,0.5] and [1,0.4]. The survival time for this study was obtained from an exponential distribution.

We compared across the methods under varying sample sizes ($n = 50$, 100, 250, 500) by using the Monte Carlo simulation method. We did 1000 replications for each model using methods performed by R 2.6.0. The data were then analyzed using the randomSurvivalForest package (The randomSurvivalForest Package., 2008). In addition, RSF drew 1000 bootstrap samples from the generated data, grew a tree for each bootstrapped data set and split on a predictor using a survival splitting rule.

Concordance error rates were obtained from each method for 1000 replications and the mean of the concordance error rates were recorded for each sample size.

2.5. Breast cancer data

We used to analyze the breast cancer data from Ture, Tokatli, & Kurt, 2008. In total there were 23 predictors and 279 breast cancer patients diagnosed for DFS in the data. In all patients' recurrence, DFS, age, menopausal status, age of menarche, age of first delivery, number of abortus, hormone replacement therapy, family history of cancer, histologic tumor type, quadrant of tumor, tumor size,

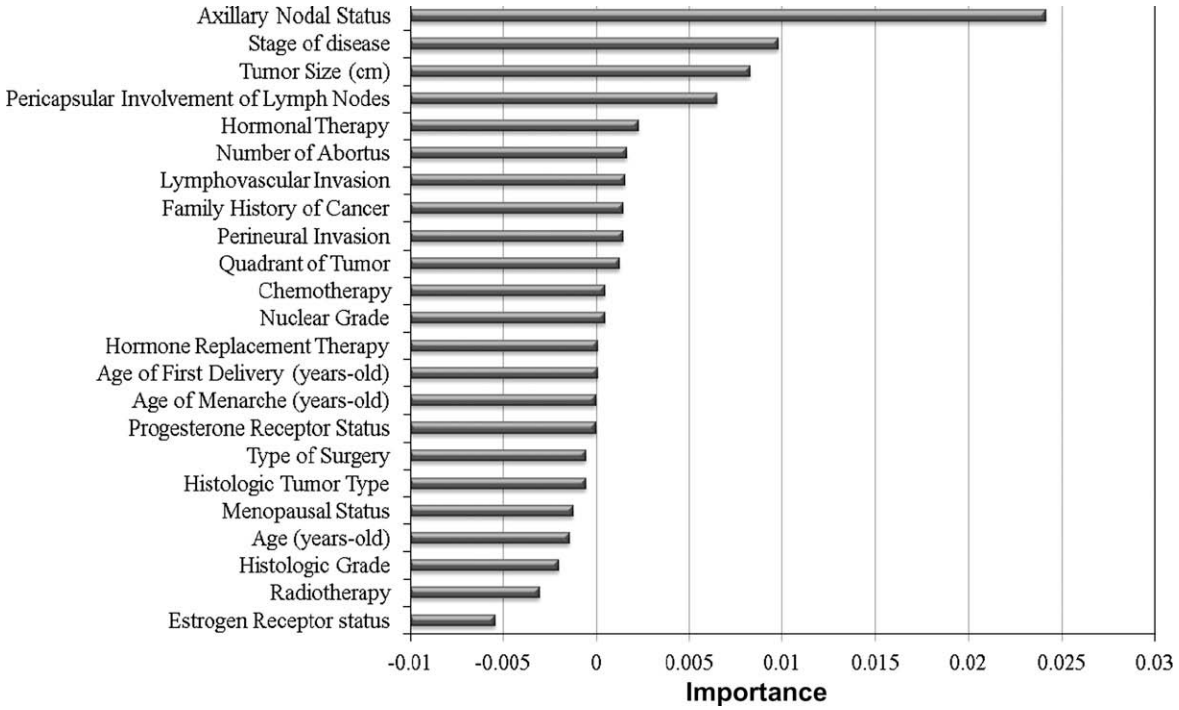

Fig. 6. Out-of-bag importance values of RSF for log-rank score splitting rule.

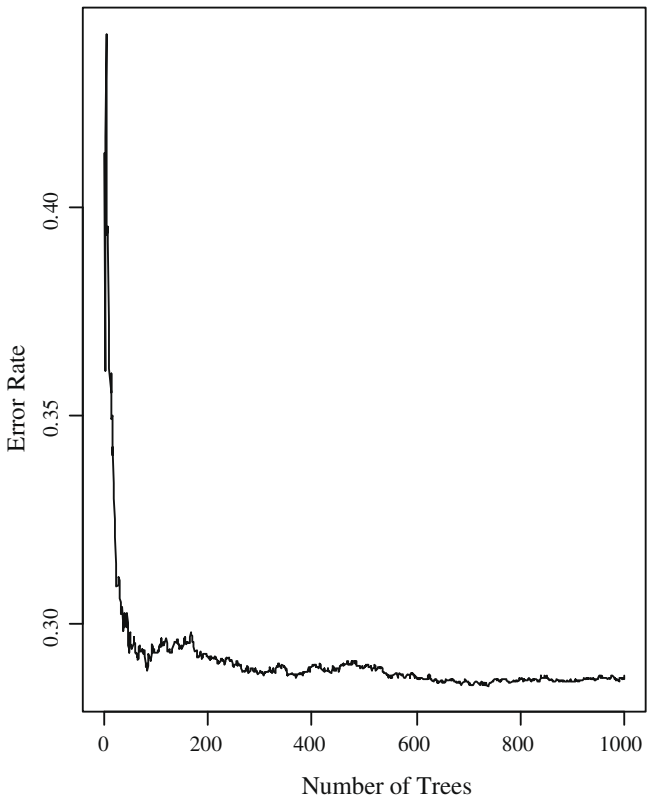

Fig. 7. Error rate of RSF for approximate log-rank splitting rule.

estrogen receptor status, progesterone receptor status, histologic grade, nuclear grade, type of surgery, axillary nodal status, pericapsular involvement of lymph nodes, stage of disease, lymphovascular invasion, perineural invasion, radiotherapy, chemotherapy and hormonal therapy were assessed and documented.

3. Results

3.1. Simulations

We simulated the generated data by running for each of the four RSF splitting rules and CRA. The averaged values over the 1000 simulations were reported in Table 1 for varying sample sizes. As can be seen, the error rates were between 0.2915 and 0.3970 for $n = 50$, 0.2704–0.3568 for $n = 100$, 0.2509–0.3331 for $n = 250$ and 0.2474–0.3274 for $n = 500$.

As it can be seen from Table 1, CRA took on the smallest error rate for all of the sample sizes. RSF approximate log-rank splitting rule ranked second, RSF log-rank splitting rule ranked third, RSF log-rank score splitting rule ranked fourth, followed by RSF conservation of events splitting rule. Besides performance of all the methods was almost similar for four sample sizes, and there was not significantly different in performance of all the methods according to two proportions t test ($p > 0.05$). As a result simulations showed that CRA had a slightly better predictive performance for four sample sizes.

3.2. Evaluation on breast cancer data

In Table 2, we gave estimates of the regression coefficients in the stepwise CRA with forward elimination for DFS time. In this model, tumor size, age, pericapsular involvement of lymph nodes and hormonal therapy had significant effects. Concordance error rate of CRA was 0.3003 (Table 3).

We considered the informativeness of each predictor under the log-rank splitting rule. In Figs. 1 and 2, we reported the error rate for the RSF log-rank model as a function of the number of trees and the out-of-bag importance values for predictors. Fig. 2 depicted the importance values for all 23 predictors. From the plot, we found

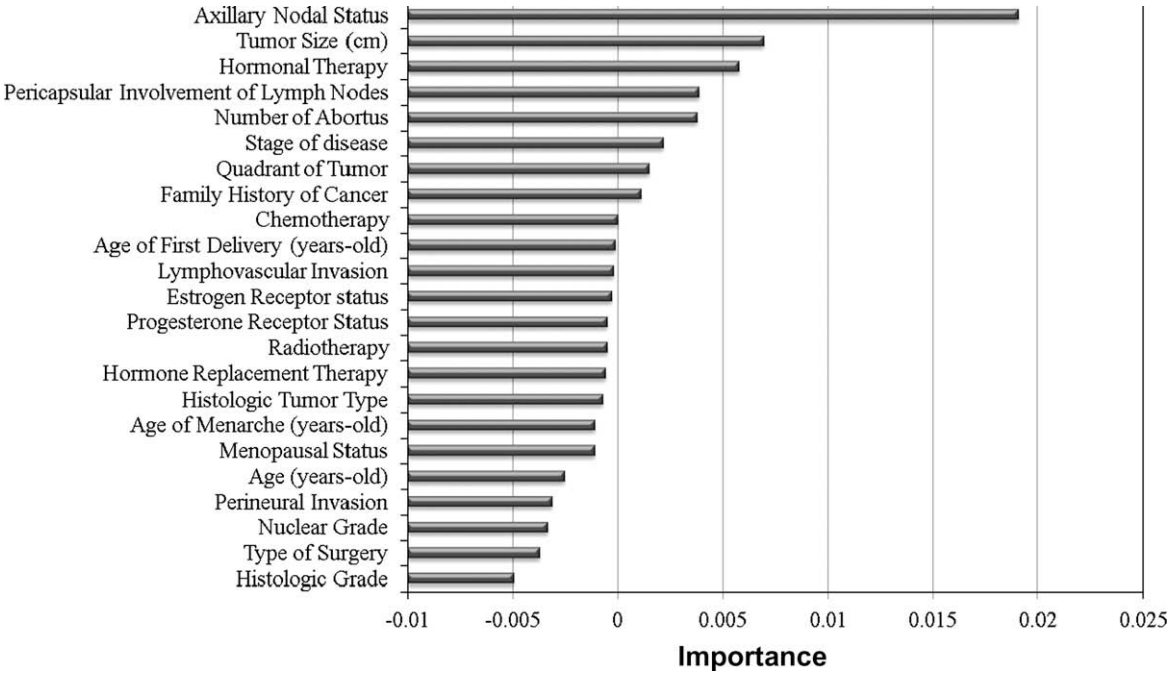

Fig. 8. Out-of-bag importance values of RSF for approximate log-rank splitting rule.

that the six prognostic factors (axillary nodal status, tumor size, pericapsular involvement of lymph nodes, number of abortus, stage of disease and hormonal therapy, respectively) had effect on DFS. However, we see that axillary nodal status, tumor size and pericapsular involvement of lymph nodes were clearly predictive and had substantially larger importance values than all other prognostic factors. Tumor size, hormonal therapy and pericapsular involvement of lymph nodes were found significant at both RSF log-rank splitting rule and CRA. Concordance error rate of this RSF model was 0.2949 (Table 3).

In the RSF for conservation of events splitting rule, we report in Fig. 3 the error rate for the RSF model as a function of the number of trees and Fig. 4 the out-of-bag importance value for predictors. Fig. 4 showed that the four prognostic factors (axillary nodal status, tumor size, pericapsular involvement of lymph nodes and hormonal therapy, respectively) had effect on DFS. Tumor size, hormonal therapy and pericapsular involvement of lymph nodes were found significant by the CRA and RSF log-rank splitting rule. Concordance error rate of this RSF model was 0.2939 (Table 3).

In the RSF for log-rank score splitting rule, we report in Fig. 5 the error rate for the RSF model as a function of the number of trees and Fig. 6 the out-of-bag importance values for predictors. Fig. 6 showed that the four prognostic factors (axillary nodal status, stage of disease, tumor size and pericapsular involvement of lymph nodes, respectively) were clearly predictive and had substantially larger importance values than all other prognostic factors. Tumor size and pericapsular involvement of lymph nodes were found significant by the CRA, RSF log-rank splitting rule and RSF conservation of events splitting rule. Concordance error rate of this RSF model was 0.2998 (Table 3).

In the RSF for approximate log-rank splitting rule, we report in Fig. 7 the error rate for the RSF model as a function of the number of trees and Fig. 8 the out-of-bag importance values for predictors. From the plot, we found that the five prognostic factors (axillary nodal status, tumor size, hormonal therapy, pericapsular involvement of lymph nodes and number of abortus, respectively) had effect on DFS. Tumor size and pericapsular involvement of lymph nodes were found significant by the CRA, RSF log-rank splitting rule, RSF conservation of events splitting rule and log-rank score splitting rule. Concordance error rate of this RSF model was 0.2832 (Table 3).

4. Discussion

We tried to compare across the methods under varying sample sizes by using the Monte Carlo simulation method and to discover the risk factors and make decision rules for the management of DFS. In addition, we evaluated performance of methods using Harrell's concordance index.

Minn et al. (2007) reported that RSF was used to examine how tumor size was predicted to influence metastasis among LMS^{+} primary tumors, and the concordance index from a RSF modeling the influence of the LMS, tumor size, and other breast cancer prognostic gene expression signatures on the risk for lung metastasis was calculated by using the NKI-295 cohort. They found that tumor size was an important and often independent variable associated with metastasis in clinical studies and studies on poor prognosis gene-expression signatures. Besides they used CRA for lung metastasis-free survival and obtained the predictive ability of the LMS was independent of other standard prognostic markers. Ishwaran and Kogalur (2007) compared across four survival splitting rules of RSF on data sets from a randomized clinical trial of patients with primary biliary cirrhosis of the liver and lung cancer. They did the analysis 100 replications independently and drew 1000 bootstrap samples for each approach. Log-rank score splitting rule had the best predictive performance for lung cancer data while approximate log-rank splitting rule had the best predictive performance for primary biliary cirrhosis of the liver data. In our simulation study, we run an analysis with five predictors from generated data under each of RSF approaches and CRA for varying sample sizes, repeating the 1000 replications independently. Performances of all the methods were almost similar for four sample sizes. The difference among the highest and the smallest error rate was 0.1055, 0.0864, 0.0822, and 0.0800 for $n = 50$, 100, 250, and 500, respectively. Also, in the breast cancer application, performances of all

the methods were almost similar. All of the methods found that specially, tumor size and pericapsular involvement of lymph nodes were major risk factors for DFS in breast cancer patients.

The CRA is the most common tool for investigating simultaneously the influence of several factors on the survival time of patients. It was a good method for predicting DFS in breast cancer patients. CRA was more advantageous than RSF approaches because it is capable of extracting patterns and relationships hidden deep into medical datasets. Although RSF approaches are ensemble methods based on survival trees, they become more of a black box when interpreting the model due to the sheer number of trees generated.

As a result, the predictive capability of CRA can be used for different sample sizes and potential future suitable survival data problems, whereas RSF provide interpretive results. These methods can be used to extrapolate any risk factor collected for survival and to understand what predictors are affecting the survival with a higher level of confidence than with other methods.

References

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

Harrell, F., Califf, R., Pryor, D., Lee, K., & Rosati, R. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association, 247*, 2543–2546.

Horthorn, T., & Lausen, B. (2003). On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis, 43*, 121–137.

Hosmer, D. W., & Lemeshow, S. (1999). *Applied survival analysis: Regression modeling of time to event data*. New York: John Wiley.

Ishwaran, H., & Kogalur, U. B. (2007). Random survival forests for R. *R News, 7*(2), 25–31.

Kleinbaum, D. G., & Klein, M. (2005). *Survival analysis: A self-learning text (statistics for biology and health)*. New York: Springer.

Minn, A. J., Gupta, G. P., Pauda, D., Bos, P., Nguyen, D. X., Nuyten, D., et al. (2007). Lung metastasis genes couple breast tumor size and metastatic spread. *PNAS, 104*(16), 6740–6745.

The randomSurvivalForest Package. (2008) <http://cran.r-project.org/web/packages/randomSurvivalForest/randomSurvivalForest.pdf>.

Ture, M., Tokatli, F., & Kurt, İ. (2008). Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients. *Expert Systems with Applications, 36*(2P1), 2017–2026.