


USE OF ENSEMBLE METHODS FOR SURVIVAL PREDICTION

Aslıhan ŞENTÜRK ACAR*, Department of Actuarial Sciences, Faculty of Science, Hacettepe University, Turkey

aslihans@hacettepe.edu.tr

( <https://orcid.org/0000-0002-1708-2028>)

Nihal ATA TUTKUN, Department of Statistics, Faculty of Science, Hacettepe University, Turkey

nihalata@hacettepe.edu.tr

( <https://orcid.org/0000-0001-5204-680X>)

Received: 02.10.2020, Accepted: 30.12.2020

*Corresponding author

Research Article

DOI: 10.22531/muglajsci.804566

Abstract

Cox regression model is used for modelling censored data to investigate the association between the survival time and covariates. It is important to assess the fit of Cox regression model since it has a key assumption called proportional hazards. Violation of this assumption induces an invalid model and changes the interpretation of the results. When the objective is the risk prediction, various machine learning methods can be good alternatives to Cox regression model due to their flexible structure. In this study, Turkish breast cancer data set is used to compare the predictive performance of Cox regression model and ensemble machine learning methods. Integrated Brier score is used to measure the predictive performance of candidate models. Based on case study results, machine learning methods are promising alternatives for survival prediction.

Keywords: Censored data, Cox regression model, Machine learning, Survival ensemble methods, Prediction

SAĞKALIM KESTİRİMİ İÇİN KOLEKTİF YÖNTEMLERİN KULLANILMASI

Özet

Cox regresyon modeli, yaşam süresi ve eşdeğişkenler arasındaki ilişkinin araştırılması için sansürlenmiş verinin modellenmesinde kullanılmaktadır. Orantılı tehlikeler gibi anahtar bir varsayıma sahip olması nedeniyle Cox regresyon model uyumunun değerlendirilmesi önemlidir. Bu varsayımın ihlali, geçersiz bir modele neden olur ve sonuçların yorumunu değiştirir. Amaç risk kestirimi olduğunda, esnek yapıları nedeniyle çeşitli makine öğrenimi yöntemleri Cox regresyon modeli için iyi alternatifler olabilir. Bu çalışmada Cox regresyon modeli ile kolektif makine öğrenimi yöntemlerinin kestirim performanslarının karşılaştırılması için Türk meme kanseri veri seti kullanılmıştır. Aday modellerin kestirim performanslarının ölçülmesinde bütünleşmiş Brier skor kullanılmıştır. Örnek çalışma sonuçlarına göre, makine öğrenimi yöntemleri sağkalım kestirimi için gelecek vaat eden alternatiflerdir.

Anahtar Kelimeler: Sansürlenmiş veri, Cox regresyon modeli, Makine öğrenimi, sağkalım kolektif yöntemleri, Kestirim

Cite

Acar Ş., A., Tutkun A., N. (2020). "Using Ensemble Methods for Survival Prediction", *Mugla Journal of Science and Technology*, 6(2), 158-164.

1. Introduction

Survival data includes the information of survival time that is defined as the time to the occurrence of a given event such as death and remission. When subjects have not experienced the event of interest at the end of the study, exact survival times are unknown and called as censored observations. If the event occurs for each individual in the sample, many classical statistical methods can be used. However, some of the individuals generally do not have the event of interest at the end of follow-up time and their true time to event is unknown. Therefore, methods for analyzing censored data differ

from classical approaches [1]. Although several statistical models have been proposed for analyzing survival data, frequently used one is Cox regression model (CRM).

Main assumption of CRM is the proportionality of the hazards. If this assumption is violated, usage of CRM may not be valid. Unfortunately, this important assumption has been checked and properly reported in very few scientific publications [2]. An alternative to CRM is the parametric survival models that involve stronger assumptions than semi-parametric models. Compared to CRM, there is extra requirement of checking the appropriateness of the selected distribution. To avoid the increased efforts of model checking and model

misspecification of parametric survival models, statisticians tend to prefer CRM [3].

CRM can be used to identify the variables that significantly affect the outcome of interest and present the results in terms of hazard ratio [4]. Exploration of presence of high order interactions needs inclusion of interaction terms in the model that makes model interpretation more difficult [5]. Alternative strategies handling these problems easily are machine learning (ML) algorithms that are applied to survival data in recent years [6]–[9]. Survival trees and survival ensembles are popular nonparametric ML methods used as an alternative to classical survival models.

Wang et al. [9] provide a brief overview of ML methods in addition to traditional methods for survival data analysis. They classify ML methods for survival data under five basic classes: survival trees, Bayesian methods, artificial neural networks, support vector machines and advanced ML methods. Survival trees [10] are extension of decision trees for survival data. Ensemble learning methods that are classified under advanced ML methods, are purposed to improve predictive performance of single decision trees. They combine the predictions of various base models (decision trees) to provide better predictions than a single model. These methods are nonparametric methods and have advantages such as easily handling interactions between variables and nonparametric relations. Predictions are obtained by averaging over individual trees in regression and by majority voting in classification tasks. Bagging [11], random forest [12] and boosting [13] methods are ensemble methods. Bagging is the earliest ensemble method that relies on bootstrap aggregation to reduce the variance of prediction [14]. Breiman [11] applied bagging procedure to solve the overfitting and stability problems that occur with single trees. Hothorn et al. [10] applied bagging procedure to right-censored data. They approximated bootstrap aggregated conditional survivor function using bootstrap learning samples. Breiman [12] purposed random forest approach to improve bagging procedure by choosing a random sample of predictors at a given tree node. Objective is to reduce the correlation among the trees and improve prediction accuracy. Random survival forest (RSF) is purposed by Ishwaran et al. (2008) by applying random forest approach to right censored survival data [15]. Boosting method combines the base learners iteratively to obtain strong learner. Hothorn et al. [16] introduced a random forest and a generic gradient boosting algorithms for censored data.

In this study, we aim to compare risk prediction performance of CRM and ensemble ML methods using a real data set related to breast cancer patients in Turkey. Here, risk is defined as recurrence of the illness after the operation time. We used bagging, random forest, gradient boosting method (GBM) to predict survival risk as an alternative to CRM. Predictive performances are compared using integrated Brier Score (IBS) criterion.

This paper is organized as follows. Basic information about survival analysis, Cox regression model and

ensemble learning methods are introduced in Section 2. Criterion to compare different methods is summarized in Section 3. Case study is applied in Section 4. Conclusion is given in Section 5.

2. Methods

Survival analysis is a class of statistical methods for studying the occurrence and timing of events. The distribution of survival time is characterized by probability density function, hazard function or survival function.

Hazard function is defined as,

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T \leq \Delta t / T > t)}{\Delta t} \quad (1)$$

for $t > 0$ and represents the probability that an individual alive at t experiences the event in the next period Δt . Survival function denotes the probability that failure will occur after time t and is given by,

$$S(t) = P(T > t) = \int_t^{\infty} f(x) dx, \quad 0 < t < \infty \quad (2)$$

A non-parametric estimator of $S(t)$ is Kaplan-Meier (KM) estimator which does not allow to evaluate the impact of more than one covariate. This disadvantage causes to prefer regression type models to make a more detailed analysis including covariates. Although several survival models are suggested for censored data, CRM is the most popular and applicable regression type within these models.

2.1. Cox Regression Model

Analyzing the impact of potential factors on patients' survival is typically based on the CRM that is a semi-parametric survival model. The model enables one to assess a relationship between patients' survival time and covariates.

CRM is used to obtain the covariate effects on hazard function. The survival data for unit i consists of $(t_i, \delta_i, \mathbf{x}_i, i = 1, 2, \dots, N)$ where t_i is the time on study, δ_i is the event indicator ($\delta_i = 1$ if the event has occurred and $\delta_i = 0$ if the survival time is censored) and \mathbf{x}_i is the vector of covariates. Under these notations hazard function is given by,

$$h_i(t) = h_0(t) \exp(\boldsymbol{\beta}' \mathbf{x}_i) \quad (3)$$

where $h_0(t)$ is the baseline hazard function and $\boldsymbol{\beta}$ is a $px1$ vector of unknown parameters. If the set of units who are at risk at time t_i is denoted by $R(t_i)$, likelihood for the CRM is given by,

$$(4)$$

$$L(\beta) = \prod_{i=1}^N \left[\frac{\exp(\beta' \mathbf{x}_i)}{\sum_{l \in R(t_i)} \exp(\beta' \mathbf{x}_l)} \right]^{\delta_i}$$

Parameter estimation is obtained by the Newton-Raphson procedure in the CRM.

2.2. Bagging

Bagging term was used by Breiman [11] as an acronym for bootstrap aggregation. Decision trees may suffer from high variance problem that indicates obtaining quite different predictions from model fits applied using randomly selected different training samples. Bootstrap aggregation method draws multiple bootstrap samples from the original data, fits decision tree to each bootstrap sample and averages prediction of those trees to reduce the variance and to improve prediction accuracy. In this procedure, approximately two-thirds of the observations are used to fit bagged trees and the remaining observations (out-of-bag) can be used for prediction error.

Hothorn et al. [10] purposed bagging survival trees that computes survival trees based on bootstrap samples. In this method, a survival tree is constructed on each bootstrap sample. For each subsample, bootstrap aggregated estimator of the survival function is the KM curve.

2.3. Random Forest

Breiman (2001) purposed random forest approach by incorporating a random feature selection process into the base learning process. In this method, a random bootstrap sample is drawn to grow a tree as in bagging process but at each node, randomly selected subset of the predictors are chosen for the split. In this way, correlation between the trees is reduced and more accurate predictions can be obtained. Number of randomly selected predictors is approximately chosen as the square root of total number of predictors.

Hothorn et al. [16] proposed random forest algorithm for the log-survival time. They used inverse probability of censoring (IPC) weights obtained using KM estimator for bootstrap sampling and a tree is constructed for each sample. IPC weights are used to deal with bias in population average. RSF approach is purposed by Ishwaran et al. (2008) by applying random forest approach to right censored survival data [15]. They used four different splitting rule; log-rank, conservation-of-events, log-rank score rule and random log-rank splitting rule. Algorithm can be summarized as follows,

- B bootstrap samples are drawn from the original data. On average 37% of observations is excluded from the sample as out-of-bag (OOB) data.
- Survival tree is constructed for each bootstrap sample. At each tree node, a subset of predictor variables is chosen randomly. The node is split into daughter nodes according to the survival difference.

- This process is repeated recursively for each node until a predetermined stopping criterion is ensured.
- Cumulative hazard rate functions are estimated using Nelson-Aalen estimators for each tree [15], [17] and these estimates are aggregated to obtain ensemble survivor function.
- Prediction error is calculated using OOB data.

2.4. Boosting

Boosting method is based on an iterative estimation process in that each weak estimator (tree) is grown using the information (residuals) of previous estimator to obtain a strong estimator. Objective is to minimize a loss function that is initially defined. Number of terminal nodes of the trees may be small and can be determined in the algorithm. Boosting method has three tuning parameters for the estimation; number of trees to fit, shrinkage parameter that controls the learning rate of algorithm and the number of splits in each tree [18]. Two different implementations of boosting approach are model based boosting (gradient boosting) and likelihood based boosting (offset boosting). These two methods differ with the updates of regression coefficients at each step [19]. In this study, we use gradient boosting algorithm [20] that can be summarized as follows,

Let y denotes the response variable, $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ is a set of explanatory variables and $L(y, F(\mathbf{x}))$ is a generic loss function. Objective is to estimate $F^*(\mathbf{x})$ that minimizes expected value of $L(y, F(X))$ over the joint distribution of (y, \mathbf{x}) ,

$$F^* = \arg \min_F E_{y, \mathbf{x}} L(y, F(\mathbf{x})) \quad (5)$$

Friedman [18] focused on additive form of $F(\mathbf{x})$ as $F(X, \{\beta_m, \mathbf{u}_m\}_1^M) = \sum_{m=1}^M \beta_m h(\mathbf{x}, \mathbf{u}_m)$ where $h(\mathbf{x}, \mathbf{u})$ is a base learner (e.g. classification tree) with parameters $\mathbf{u} = \{u_1, u_2, \dots\}$. Different numerical optimization methods can be to obtain parameters. Steepest descent numerical minimization method is one of them and summarized by Friedman [18] as follows,

Define an initial value $F_0(\mathbf{x}) = \arg \min_{\rho} \sum_{i=1}^N L(y_i, \rho)$ where, $\rho_m = \arg \min_{\rho} E_{y, \mathbf{x}} L(y, F_{m-1}(\mathbf{x}) - \rho g_m(\mathbf{x}))$ with $g_m(\mathbf{x}) = E_y \left(\frac{\partial L(y, F(\mathbf{x}))}{\partial F(\mathbf{x})} \middle| \mathbf{x} \right)_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}$ is the gradient.

For $m=1$ to M iterate:

1. Pseudo-responses are calculated, $\tilde{y}_i = - \left[\frac{\partial L(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)} \right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x}_i)}$, $i = 1, 2, \dots, N$
2. Base learner is fitted to pseudo-responses, $\mathbf{u}_m = \arg \min_{\mathbf{u}, \beta} \sum_{i=1}^N [\tilde{y}_i - \beta h(\mathbf{x}_i, \mathbf{u})]^2$
3. $\rho_m = \arg \min_{\rho} \sum_{i=1}^N L(y_i, F_{m-1}(\mathbf{x}_i) + \rho h(\mathbf{x}_i, \mathbf{u}_m))$
4. Model is updated, $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \rho_m h(\mathbf{x}, \mathbf{u}_m)$

For quadratic loss function, this method is called L_2 boosting method for censored data. Bühlmann [21] used L_2 boosting method with component-wise least squares. In this study we used gradient boosting approach with component-wise linear models.

3. Comparison of Methods

When the objective is prediction, predictive accuracy of models can be evaluated by discrimination and calibration. In the presence of censored data, evaluating predictive accuracy of the models is difficult due to the unknown failure times of observations. In recent years, time-dependent Brier score (BS) [22] is frequently used to compare predictive performance of survival models [23]. BS is a measure of both discrimination and calibration [24].

BS measures average difference between the observed outcome and the predicted survival probability [6]. For $i = 1, 2, \dots, N$ let T_i and C_i denote survival time and censoring time of subject i respectively. BS at time t is given by,

$$BS(t) = \frac{1}{N} \sum_{i=1}^N \left\{ \frac{(0 - \hat{S}(t/x_i))^2}{\hat{G}(t)} I(T_i \leq t, \delta_i = 1) + \frac{(1 - \hat{S}(t/x_i))^2}{\hat{G}(t)} I(T_i > t) \right\} \quad (6)$$

where $\hat{G}(t) = P(C_i > t)$ denotes the KM estimate of the censoring survival function ([22]; [25]). IBS is obtained by calculating BS across all available times. Lower BS indicates better fit of the model.

4. Case Study

4.1. About data

Breast cancer data [26] related to 124 patients are used for the case study. In the following analysis, recurrence of the illness after the operation time is the endpoint of interest (failure). This variable is measured in months and there is 146-month-follow-up period. Patients that did not experience disease again at the end of the follow-up period, are treated as censored observations. 35.5% of sample are censored. 5-year survival probability is 59.4% in the sample. Mean, median, 1st quartile and 3rd quartile of survival times are 74-months (67-83), 81-months (60-91), 36-months (30-48) and 108-months (96-120) with %95 confidence interval inside the parenthesis respectively.

Age, medical treatment type, radiotherapy, tumour size, type of intervention, stage of disease, toxicity, climacteric, number of axillary lymph nodes are used as prognostic factors which effect the survival time of breast cancer patients. Table 1 shows the summary information about predictor variables.

{Table 1 is at the bottom of the paper}

4.2. Analysis

Before modeling process, proportional hazards assumption is assessed by a statistical test. This test is accomplished by finding the correlation between the Schoenfeld residuals for a particular covariate and the ranking of individual failure times. It is shown by a correlation analysis of partial residuals with time. p -values obtained for all covariates are greater than 0.05 which shows the validity of PH assumption for this data set. Therefore, CRM is run for the data set and model fit results are given in Table 2. The model is statistically significant at 95% confidence interval.

Table 2. Results of CRM

Variable	Coef	S.E. (coef)	z	p	exp (coef)
Age	0.043	0.019	5.283	0.022	1.044
Clm1	0.12	0.397	0.092	0.762	1.128
Size1	-0.968	0.311	9.687	0.002	0.38
Size2	-1.118	0.397	7.926	0.005	0.327
Node1	-0.513	0.297	2.982	0.084	0.599
Node2	0.348	0.399	0.761	0.383	1.416
Rth1	-0.374	0.245	2.324	0.127	0.688
Tox1	-0.209	0.267	0.612	0.434	0.811
Int1	1.026	0.42	5.973	0.015	2.791
Int2	0.434	0.409	1.123	0.289	1.543
Stage1	0.326	0.274	1.407	0.236	1.385
Trt1	0.088	0.291	0.09	0.764	1.091

According to p -values given in Table 2, age, tumour size and type of intervention (modified radke mastectomy) are statistically significant at 95% confidence level. Estimated hazard of age is 1.044 that means the risk of failure increases 1.044 unit for 1-unit increase in age. The hazard for the patients with the tumor size smaller than 2 cm is 2.63 times the hazard for the patients with the tumor size of 2-5 cm and 3.06 times for the patients with the tumor size larger than 5 cm. The hazard for the patients with the intervention of modified radke mastectomy is 2.791 times the hazard for the patients with fixed mastectomy.

We also used backward stepwise variable selection (BSVS). According to variable selection, only age and tumour size are statistically significant variables. Results are given in Table 3.

Table 3. Results of CRM (BSVS)

Variable	Coef	S.E.(coef)	z	p	exp(coef)
Age	0.039	0.010	3.77	0.0002	1.040
Size 1	-0.842	0.271	-3.11	0.0019	0.431
Size 2	-0.917	0.361	-2.54	0.011	0.400

Our second model is bagging method that use conditional inference trees as base learners [27]. We used 1000 trees for fitting process. Third approach is Ishwaran's RSF approach that is fitted using 1000 trees [15] as in bagging and default parameters of randomForestSRC package in R [15] are used. Log-rank splitting rule is used for the algorithm [28]. According to RSF results, average terminal node size includes three observations and average number of terminal nodes per tree is 38. Three of nine predictors are randomly selected as candidate predictors for each node split. OOB error is 37.27%. Lastly, negative partial log-likelihood is used for Cox model inside GBM algorithm [29]. According to GBM estimation results age, radiotherapy, tumour size, type of intervention, stage of disease, toxicity and the number of axillary lymph nodes are selected as predictor variables.

Since the sample size is small, dividing data set into training and test set may induce bias. In this case, cross-validation (CV) methods are used to see how the model will perform on independent data set. We used 10-fold CV method to compare predictive performance of models. In this method, data set is randomly separated into 10 subsamples in equal size. One of subsample is separated as validation data and remaining ones are used for model training. Estimation results of 10 subsamples are averaged to calculate final estimation value. IPC weights are used to deal with right-censored data and KM method is used to estimate weights. CV and model comparison are performed using riskRegression R package [23]. Since gradient boosting method does not exist within applicable methods inside riskRegression package, we extended codes to implement it.

Table 4 shows IBS values calculated at mean (74), median (81), 1st quartile (36) and 3rd quartile (108) values of survival times (months) with %95 confidence intervals inside the parenthesis respectively.

{Table 4 is at the bottom of the paper}

According to prediction results given in Table 4, when variable selection is not applied to CRM, RF and GBM performed better than CRM and bagging method. But, when BSVS is applied, CRM performed best predictive performance at mean and median survival time points. At 1st quartile and 3rd quartile survival time points, GBM and RF performed best respectively. We can conclude that variable selection is a vital step for the predictive performance of CRM and ML methods are good alternatives to CRM. Important advantage of RF and GBM is the automatic variable selection process inside the algorithms.

5. Conclusion

CRM is one of the most widely used approach for censored data due to its easy interpretability and few assumptions. But it has some restrictions such as proportional hazards assumption. To predict risk with censored data, ML methods are good alternative since they do not have distributional restrictions and some algorithms handle variable

selection, interactions and nonlinear relationships among variables automatically.

Case study supports the idea that ML methods are good alternatives to classical survival methods when the objective is the risk prediction. RF and GBM performed better than bagging and CRM (without variable selection) at all prediction time points. This is an expected result since RF and GBM are improved versions of bagging method. Variable selection increased the predictive performance of CRM since CRM performed better than RF and GBM after BSVS at mean and median survival times. But we can not generalize this result because performances may change according to the data sets.

Restriction of the study is that sample size of data set is small and there are few predictor variables. ML methods are known to be effective for big data analytics. As a future study, predictive performance of ensemble methods and CRM can be compared on different larger data sets. Also different popular ML methods such as artificial neural networks [30] and support vector machines [31] can be used for survival prediction.

6. References

- [1] Clark, T. G., Bradburn, M. J., Love S. B., and Altman, D. G., "Survival analysis part I: basic concepts and first analyses," *Br. J. Cancer*, vol. 89, no. 2, pp. 232–238, Jul. 2003.
- [2] Babińska, M., Chudek, J., Elżbieta Chełmecka, Janik, M., Klimek, K., and Owczarek, A., "Limitations of Cox Proportional Hazards Analysis in Mortality Prediction of Patients with Acute Coronary Syndrome", *Studies in Logic, Grammar and Rhetoric*, 2018.
- [3] Nardi, A. and Schemper M., "Comparing Cox and parametric models in clinical studies," *Stat. Med.*, vol. 22, no. 23, pp. 3597–3610, Dec. 2003.
- [4] Kleinbaum, D.G., *Survival Analysis - A Self-Learning Text*, Springer, 2010.
- [5] Radespiel-Tröger, M., Rabenstein T., Schneider, H. T. and Lausena B., "Comparison of tree-based methods for prognostic stratification of survival data," *Artificial Intelligence in Medicine*, 28(3), pp.323-341, 2003.
- [6] Zhou, Y. and McArdle, J. J., "Rationale and Applications of Survival Tree and Survival Ensemble Methods," *Psychometrika*, vol. 80, no. 3, pp. 811–833, Sep. 2015.
- [7] Hu, C., and Steingrimsson, J. A., "Personalized Risk Prediction in Clinical Oncology Research: Applications and Practical Issues Using Survival Trees and Random Forests," *J. Biopharm. Stat.*, vol. 28, no. 2, pp. 333–349, Mar. 2018.
- [8] Paraschiakos, F., "Machine learning for survival analysis on clinical data," Master's thesis, 2016.
- [9] Wang, P., Li, Y. and Reddy C. K., "Machine Learning for Survival Analysis: A Survey", *ACM Computing Surveys (CSUR)*, 51(6):1-36, 2019.

- [10] Hothorn, T., Lausen, B., Benner, A. and Radespiel-Tröger, M. "Bagging survival trees," *Stat. Med.*, vol. 23, no. 1, pp. 77–91, Jan. 2004.
- [11] Breiman, L., "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.
- [12] Breiman, L., "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [13] Schapire, R.E., "A brief introduction to boosting," in *Ijcai*, vol. 99, pp. 1401–1406, 1999.
- [14] Kuhn, M. and Johnson, K., *Applied predictive modeling*, vol. 26. Springer, 2013.
- [15] Ishwaran, H., Kogalur, U. B., Blackstone, E. H. and Lauer, M. S. "Random survival forests," *Ann. Appl. Stat.*, vol. 2, no. 3, pp. 841–860, Sep. 2008.
- [16] Hothorn, T., Bühlmann P., Dudoit, S., Molinaro A. and van der Laan M. J., "Survival ensembles," *Biostat. Oxf. Engl.*, vol. 7, no. 3, pp. 355–373, Jul. 2006.
- [17] Mogensen, U. B., Ishwaran, H. and Gerds, T. A., "Evaluating random forests for survival analysis using prediction error curves," *J. Stat. Softw.*, vol. 50, no. 11, p. 1, 2012.
- [18] Friedman, J. H., "Greedy Function Approximation: A Gradient Boosting Machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001
- [19] Bin, R. D., Sauerbrei, W. and Boulesteix, A. L., "Investigating the prediction ability of survival models based on both clinical and omics data: two case studies," *Stat. Med.*, vol. 33, no. 30, pp. 5310–5329, 2014.
- [20] Bühlmann, P. and Hothorn, T., "Boosting algorithms: Regularization, prediction and model fitting," *Stat. Sci.*, pp. 477–505, 2007.
- [21] Bühlmann, P., "Boosting for high-dimensional linear models," *Ann. Stat.*, vol. 34, no. 2, pp. 559–583, Apr. 2006.
- [22] Graf, E., Schmoor, C., Sauerbrei, W. and Schumacher, M., "Assessment and comparison of prognostic classification schemes for survival data", *Statistics in medicine*, 18(17-18), pp.2529-2545, 1999.
- [23] Gerds, T. A. and Ozenne, B., "riskRegression: Risk Regression Models and Prediction Scores for Survival Analysis with Competing Risks," vol. *R package version 2019*.
- [24] Gerds, T. A., Andersen, P. K. and Kattan, M. W., "Calibration plots for risk prediction models in the presence of competing risks," *Stat. Med.*, vol. 33, no. 18, pp. 3191–3203, 2014.
- [25] Gerds, T. A. and Schumacher, M., "Consistent estimation of the expected Brier score in general survival models with right-censored event times," *Biom. J. Biom. Z.*, vol. 48, no. 6, pp. 1029–1040, Dec. 2006.
- [26] Erdoğan A., "Proportional Hazards Model," Unpublished MSc thesis, Hacettepe University, Ankara, 1993.
- [27] Hothorn, T., Hornik, K. and Zeileis, A., "Unbiased recursive partitioning: A conditional inference framework," *J. Comput. Graph. Stat.*, vol. 15, no. 3, pp. 651–674, 2006.
- [28] LeBlanc M. and Crowley, J., "Survival trees by goodness of fit," *J. Am. Stat. Assoc.*, vol. 88, pp. 457–467, 1993.
- [29] Hofner, B., Mayr, A., Robinzonov, N. and Schmid, M., "Model-based Boosting in R: A Hands-on Tutorial Using the R Package mboost," Feb. 14, 2012.
- [30] Biganzoli, E., Boracchi, P., Mariani, L. and Marubini, E., "Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach," *Statistics in medicine*, May 30, 1998.
- [31] Van Belle, V., Pelckmans, K., Van Huffel, S. and Suykens, J. A., "Support vector methods for survival analysis: a comparison between ranking and regression approaches," *Artificial intelligence in medicine*, Oct. 2011.

Table 1. Summary of predictor variables

Predictor variables (abbreviation)	Levels	Definition of levels	Numbers
Medical treatment type (trt)	0	Tamoxifen (TMX)	59
	1	Phenylalanine mustard (L-PAM)	65
Radiotherapy (rth)	0	Yes	66
	1	No	58
Tumour size (size)	0	<2 cm	31
	1	2-5 cm	66
	2	>5 cm	27
Type of intervention (int)	0	Fixed mastectopia	21
	1	Modified radke mastectopia	48
	2	Radke mastectopia	55
Stage of disease (stage)	0	Stage I	42
	1	Stage II	82
Toxicity (tox)	0	None	76
	1	Nausea/vomiting/ hot head	48
Climacteric (clm)	0	Cut	66
	1	Continue	58
Number of axillary lymph nodes	0	None	50
	1	1-2-3	52
	2	4+	22
Age (median (min-max))		49 (25-78)	

Table 4. IBS values (%95 confidence interval)

Model	Mean (74)	Median (81)	1st quartile (36)	3rd quartile (108)
CRM	0.227 (0.144,0.327)	0.236 (0.161,0.309)	0.221 (0.182,0.243)	0.245 (0.141,0.446)
CRM (BSVS)	0.214 (0.165,0.317)	0.214 (0.170,0.302)	0.197 (0.169,0.234)	0.208 (0.125,0.304)
Bagging	0.229 (0.161,0.356)	0.231 (0.162,0.332)	0.211 (0.169,0.256)	0.221 (0.140,0.367)
RF	0.225 (0.161,0.312)	0.233 (0.174,0.306)	0.216 (0.187,0.237)	0.202 (0.160,0.281)
GBM	0.218 (0.160,0.291)	0.223 (0.171,0.287)	0.196 (0.168, 0.224)	0.228 (0.141,0.411)