Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2020;6(4):267-274

Original Article

Predicting Epithelial Ovarian Cancer First Recurrence with Random Survival Forest: Comparison Parametric, Semi-Parametric, and Random Survival Forest Methods

Maryam Deldar[1], Robab Anbiaee[2], Kourosh Sayehmiri[1]*

[1]Department of Biostatistics, Faculty of health, Ilam University of Medical Sciences, Ilam, Iran.
[2]Department of Radiotherapy, School of Medicine, Imam Hossein Hospital, Shahid Beheshti University of Medical Sciences.

ARTICLE INFO

Received 05.10.2020
Revised 03.11.2020
Accepted 11.11.2020
Published 10.12.2020

Key words:
Epithelial ovarian cancer;
Cox;
Weibull;
First recurrent;
Random survival

ABSTRACT

Objective: Rapid technological advances in the last century and the large amount of information have made it difficult to analyze a large number of independent variables. In such circumstances, the existence of interactions of different degrees in the model is expected, in this case, the Cox model cannot be useful and the nonparametric method of random survival forest can be a useful alternative. This study compares the prediction error of random survival forest with Cox and Weibull models in predicting the time to the first recurrence in patients with epithelial ovarian cancer.
Method: In this retrospective study, the records of patients with epithelial ovarian cancer who referred to Imam Hossein Hospital in Tehran from 2007 to 2018 were used. To investigate the factors affecting the first recurrence of these patients, RSF was fitted to the data. Finally, prediction error of Cox, Weibull and RSF were compared using C-Index and Brier score.
Results: Brier score was calculated 0.16 for RSF, and 0.24 for Cox, also C-Index was calculated 0.34 for RSF and 0.42 for Cox. Brier score was calculated 0.092 for Cox and 0.089 for Weibull, so the prediction error of RSF was lower than both Cox and Weibull models.
Conclusion: Random survival forest with a suitable fit on many variables and without the need for a special default with a prediction error less than the Weibull and Cox methods can predict the response variable when confronted with high-dimensional data.

Introduction

Ovarian cancer is one of the leading causes of death due to gynecological malignancies in women around the world, and despite the improvement of treatment methods in recent decades, most women with this disease relapse and eventually die (1). The five-year survival rate of ovarian cancer is approximately 47%, and because it is a rare disease, it is difficult to conduct retrospective cohort studies (2). The disease has no specific symptoms, so it is often diagnosed when the disease has progressed (3). Known factors that affect the survival of patients with

* .Corresponding Author's Email: Sayehmiri@razi.tums.ac.ir

epithelial ovarian cancer include: patient age, disease grade, disease stage, histologic type, and outcome of treatments including surgery (4).

The most common model used for survival data is the Cox model. When the sample size is high and the number of independent variables is large, it is difficult to indicate that the Cox proportional hazards are appropriate for each variable. In addition, in this method, variables with low effects may be removed in the model, while they may have large effects in contrast to other variables. In this case, the Cox model cannot be useful and non-parametric statistical method of random survival forest can be a useful alternative to the Cox model (5). The efficiency and validity of random survival forest are evaluated by indicators such as C-Index and Brier score. The non-parametric statistical method of random survival forest has been developed in order to solve the problems mentioned in Cox model and other classical models in the analysis of survival data.

Rachel Lipson(2014) in Canada predicted end-stage ovarian cancer survival times using random survival forests and parametric methods across a range of risk factors and compared the results (6). In addition, in this work, we also considered the Cox proportional hazards model as a predictive model. Robin Myte (2013) for colorectal cancer survival variables made a comparison between random survival forests and the Cox proportional hazards models (5). Furthermore, Vinnie Liu (2017) in Canada, for predicting ovarian cancer survival times based on the C-Index and the Brier score, compared random forest survival in terms of predictive power with Cox and Weibull methods (7).

This project investigates the performance of parametric and Cox methods with random survival forests in predicting first recurrence in patients with epithelial ovarian cancer. These patients referred to Imam Hossein Hospital in Tehran during 2007-2018.

Material and method

In this cohort study, the medical records of 141 patients with epithelial ovarian cancer who referred to the oncology and radiotherapy department of Imam Hossein Hospital in Tehran between 2007 and 2018 were used. The required data were extracted from the patients' files with proper education. If there was a recurrence, the time of the first recurrence of these patients and if there was no recurrence or withdraw, the patient's withdrawal time or the patient's censorship time were recorded.

Patient characteristics such as age, body mass index, white blood cell, hemoglobin, platelets, stage and degree of disease, metastatic tumor, histologic type, type of chemotherapy, chemotherapy courses and presence of ascites at diagnosis entered the random survival forest model, as factors to determine the effect on the time to the first recurrence of epithelial ovarian cancer patients. Finally, the coefficients of prediction of Cox, Weibull, and random survival Forest models were compared using the C-Index and the Brier score. Statistical analyses were performed using R (Ver 3.6.2) and relevant packages.

Statistical analysis

Random Survival Forest

The key elements in random survival forest are random growth of survival trees and

calculation of the cumulative hazard function on these trees. After the full growth of survival forest, the cumulative hazard function is an estimate for measuring error rate and as a result, enables us to compare different survival analysis methods that can be found (8). For each forest tree, the cumulative hazard function is determined by grouping the cumulative hazard estimates in each of the M-terminal nodes. Cumulative hazard estimation can be determined not only for the end nodes but for all tree nodes (8).

The RSF algorithm

(1) B bootstrap samples are extracted from the original data, with each bootstrap sample excludes an average 37% of the data, that is, out-of-bag data (OOB data). B is defined as 1000 in our study.

(2) A survival tree is grown for each bootstrap sample to develop a comprehensive model composed of all 15 variables. 4 candidate variables are randomly selected for each node. The node split using the candidate variable that maximized survival difference between child nodes.

(3) Every tree was grown to full size until each terminal node is with less than $d > 0$ unique deaths.

(4) Calculate CHF for each tree. To obtain ensemble CHF was averaged across all of the trees.

(5) At last, using OOB data, prediction error for ensemble CHF can be calculated.

Harrell's concordance index (C-Index) and Brier score

The prediction accuracy of the RSF model can be determined based on the Harrell C-Index and the Brier score. For this purpose, are used samples out-of-bag of each tree are used. The RSF prediction error is between zero and one so that, the lower prediction error of the random survival forest corresponds to greater prediction accuracy (9). C can be interpreted as the ratio of predicted rankings right from survival times to allowed pairs. The prediction error can be calculated as 1-C. The Brier score is a number between zero and one. The lower the Brier score, the more accurate it is.

Results

In terms of significance of predictor variables, random survival forest using log-rank split rule showed that tumor stage using variable importance criterion (VIMP)*100 with importance value of 1.993 Metastatic tumor with importance of 2.665, and maximum platelet count with importance of 2.132 were the most important variables affecting random survival forest model (Figure1). The software output is shown in Table 1 based on the VIMP criterion.

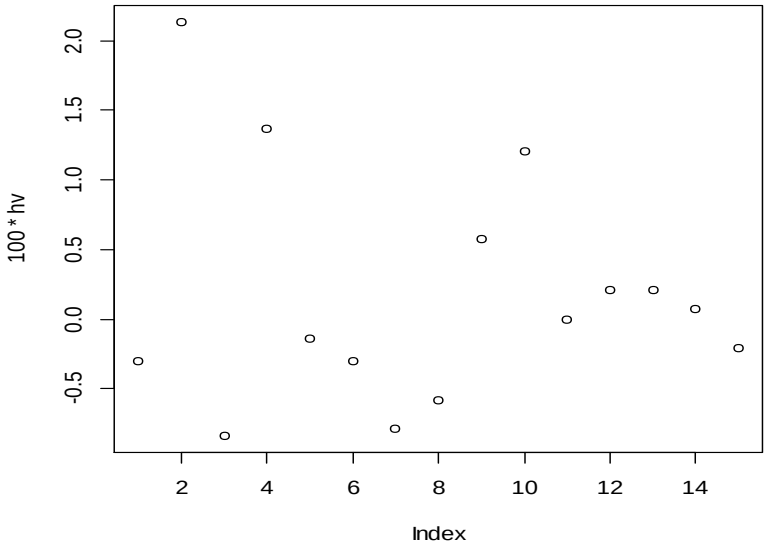

Figure1. Random Survival Forest variable importance (VIMP)*100 based on log-rank rule.

Table1. Variable importance (VIMP), for selected variables in RSF

Variables	Variable importance (VIMP)*100	Variables	Variable importance (VIMP) *100
Metastatic tumor	2.665	Age at diagnosis	-0.904
Figo stage at diagnosis	1.993	BMI(kg/m^2)	-0.811
Maximum platelet count	2.132	Ascites at diagnosis	-0.440
Mean platelet count	0.046	Chemotherapy course	0.301
Minimum platelet count	-0.509	Chemotherapy type	0.602
Mean white blood cells	-1.738	Tumor grade at diagnosis	-0.231
Mean hemoglobin	0.254	Histologic type	0.463
Minimum hemoglobin	-0.625		

Our results showed that the Brier score in the random survival forest model based on log rank split criterion had lower error and higher predictive power than the Cox model.

According to Table (2), Brier score was 0.16 for the RSF and 0.24 for the Cox model. Figure (2) compares the prediction error of these two models. According to the Harrel index, the out-of-bag error was 0.34 for the RSF and 0.42 for the Cox model, which the random survival forest error was again lower than the Cox error.

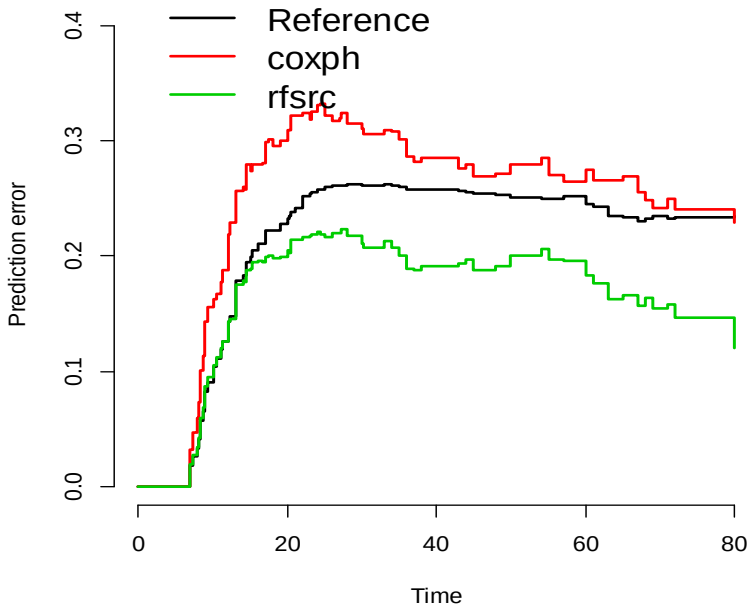

Figure2. Comparison of prediction error curves for the ovarian cancer data based on Brier score.

Comparison of Cox semi-parametric model prediction error and Weibull parametric model

We use the Brier score index to evaluate the efficiency of comparing the Cox model prediction error and the Weibull parametric model. According to our results, the Briar score for the Cox and Weibull models was approximately equal. The Cox model prediction error was 0.092 and the Weibull prediction error was 0.089 (Table 3). Figure 3 compares the prediction errors of the two Cox and Weibull methods.

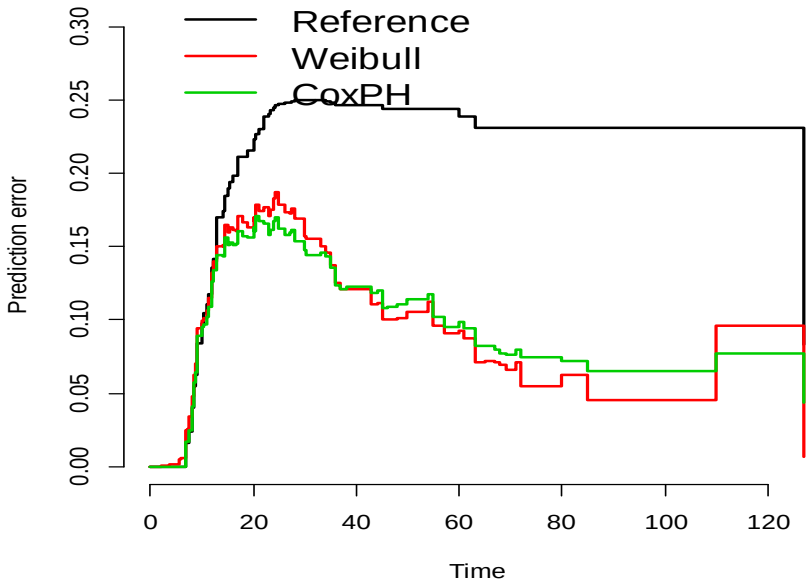

Figure3. Comparison of Brier score for the Cox and Weibull methods on ovarian cancer data.

Table2. Estimates of prediction error for Cox and RSF methods.

Selected model	Brier score	Based on C-Index
Random survival forest	0.16	0.34
Cox	0.24	0.42

Table3. Estimates of prediction error for Cox and Weibull methods based on Brier score.

Model	Cox	Weibull
Brier score	0.092	0.089

Discussion

In our study, the random survival forest based on the Harrell index had a lower out-of-bag error than the Cox semi-parametric model, and the cumulative error of the random survival forest based on the Brier score was less than both Cox and Weibull models. Rachel Lipson(2014) in Canada, which compared the random survival forest model with the Weibull model in predicting ovarian cancer survival times, using C-Index received the same error rates for each model but, the Lawless-Yuan prediction error estimates for the random survival forest model were lower (6). In another study conducted by Vinnie Liu (2017) in Canada, in predicting ovarian cancer survival times based on C-Index, random survival forest

was better predictive power and based on Briar score Cox and Weibull models were similarly better than random survival forest (7).

Ghodratollah Roshanaee (2018) determined the factors affecting on survival of kidney transplant in living donor patients using a random survival forest, based on Berier score the prediction error of random survival forest were lower than both Cox and Kaplan-Meier methods (10). Mohaddeseh Mohebbi (2015) using the Cox and Forest randomized survival models to predict the first metastasis in breast cancer, based on Brier score, two models had more predictive power than 70%, but random survival forest had more predictive power than Cox model (11). Robin Myte (2013) used both Cox and Random survival forest models to select colorectal cancer survival variables, that based on C-Index prediction error RSF was less than the Cox model (5).

Conclusion

In this work, we show the ease of using RSF in real data to discover highly complex relationships between variables. Our study included epithelial ovarian cancer that found significant relationships between disease stage, metastatic tumor and maximum platelet count with shorter disease-free survival times. Such complex relationships are easily found using tools such as VIMP in combination with the highly adapted nature of RSF. In contrast, conventional classical methods are not automated and require significant user input from data settings in which the variables are highly correlated.

Acknowledgements:

Not applicable

Conflicts of interest:

The authors declare that they have no competing interests

References

1.Lindemann K, Beale PJ, Rossi E, Goh JC, Vaughan MM, Tenney ME, et al. Phase I study of BNC105P, carboplatin and gemcitabine in partially platinum-sensitive ovarian cancer patients in first or second relapse (ANZGOG-1103). Cancer chemotherapy and pharmacology. 2019;83(1):97-105.

2.Clarke CL, Kushi LH, Chubak J, Pawloski PA, Bulkley JE, Epstein MM, et al. Predictors of long-term survival among high-grade serous ovarian cancer patients. Cancer Epidemiology and Prevention Biomarkers. 2019;28(5):996-9.

3.Sun Y, Liu S, Feng Z, Cheng J, Lu L, Wang M, et al. Preoperative omental metastasis-related maximum standardized fluorine-18-fluorodeoxyglucose uptake value can predict chemosensitivity and recurrence in advanced high-grade serous ovarian cancer patients. Nuclear medicine communications. 2018;39(8):761-7.

4.Li Z, Hong N, Robertson M, Wang C, Jiang G. Preoperative red cell distribution width and neutrophil-to-lymphocyte ratio predict survival in patients with epithelial ovarian cancer. Scientific reports. 2017;7:43001.

5.Myte R. Covariate Selection for Colorectal Cancer Survival Data: A comparison case study between Random Survival Forests and the Cox Proportional-Hazards model. 2013.

6.Lipson R. Predicting Ovarian Cancer Survival Times: Performance of Parametric Methods and Random Survival Forests. 2014.

7.Liu V. Predicting ovarian cancer survival times: Feature selection and performance of

parametric, semi-parametric, and random survival forest methods. 2019.
8.Weathers B. Comparision of Survival Curves Between Cox Proportional Hazards, Random Forests, and Conditional Inference Forests in Survival Analysis. 2017.
9.Dietrich S. Investigation of the machine learning method Random Survival Forest as an exploratory analysis tool for the identification of variables associated with disease risks in complex survival data. 2016.
10.Roshanaei G. Determining affected factors on survival of kidney transplant in living donor patients using a random survivalforest. Koomesh. 2018;20(3):523-17.
11.Mohebbi M. Application of random survival forest model in prediction of the first metastasis in breast cancer patients and comparison with cox regression analysis. 2015.