Predicting Factors Affecting the First Recurrence of Epithelial Ovarian Cancer Using Random Survival Forest

Maryam Deldar[1], Robab Anbiaee[2,3], Kourosh Sayehmiri[1]

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

Received: 22 Dec. 2020; Accepted: 18 Jul. 2021

Abstract- Predicting survival time has many Effective implications in life quality management for the remainder of the patient's life. Also, survival data are highly variable and make accurate predictions difficult or impossible. Random Survival Forest by repeated tree construction on Bootstrap samples and averaging on the results of these trees reduce the prediction error and cause further generalization of these results. In this retrospective study, the records of 141 patients with epithelial ovarian cancer who were referred to the oncology and radiotherapy ward of Imam Hossein Hospital in Tehran from 2007 to 2018 were used. Random Survival Forest was fitted to the data to investigate the key factors affecting the first recurrence of epithelial ovarian cancer. The mean age of the patients in our study was 52 (23-82) years and the median time to the first recurrence in these was 17 (0.5-127) months, respectively. According to RSF results, using variable importance criterion (VIMP) metastatic tumor with relative importance 2.665 and also using minimal (MD) by depth 2.349, tumor stage with relative importance 1.993 and depth 2.678, and maximum platelet count with relative importance 2.132 and depth 2.683 were the most important variables affecting in the first recurrence of Epithelial Ovarian Cancer. One of the disadvantages of classical methods is the inappropriate fitting of many variables and the need for specific assumptions. More advanced methods such as RSF without the need for any specific assumptions with less prediction error can well explain event variations when exposed to high-dimensional data.

Acta Med Iran 2021;59(8):504-509.

Keywords: Epithelial ovarian cancer; First recurrent; Random survival forest

Introduction

Ovarian cancer is one of the leading causes of death in gynecological cancers worldwide, and although surgical and chemotherapy outcomes have improved in recent decades, still most women with ovarian cancer have recurrent disease, and Ineffective treatment can lead to death (1). 40 to 60 percent of all epithelial ovarian cancer patients and 75 percent of patients with advanced disease eventually recurred. After surgery, the first course of chemotherapy begins with platinum. Neoadjuvant chemotherapy is used for late stages patients and patients who are poor for surgery (2).

Anemia is a common side effect of surgery and chemotherapy, and patients with anemia may have more problems than patients with sufficient hemoglobin. Also, chemotherapy can affect the body's ability to produce neutrophils and white blood cells, and there is evidence that high levels of white blood cells have benefits for patient survival (3). Although the Cox and Weibull models are interpretable and the inference about the regression coefficients as well as the relationship between the response and the predictor variables is specified, it is a simpler option for the prediction that does not require any distributional properties, is random survival forest. This work introduces random survival forests as a group tree method for analyzing survival data with right censoring, as it is well known that

Corresponding Author: K. Sayehmiri
Department of Biostatistics, Faculty of Health, Ilam University of Medical Sciences, Ilam, Iran
Tel: +98 9183410782, Fax: +98 8412227120, E-mail address: Sayehmiri@razi.tums.ac.ir

building sets of basic learners such as trees can improve prediction function (4).

Okunade in 2020 in a retrospective cohort study involving all histological EOC patients administered at the University of Lagos, Nigeria Teaching Hospital over a 7-year period, the relationship between the variables was tested and multivariate analysis was performed to adjust for all possible features that predict early EOC recurrence (5). In a 2016 study, Xiaoyan evaluated the risk factors for the recurrence of epithelial ovarian cancer using logistic regression. Out of 91 epithelial ovarian cancer patients, 33 had recurrences in the case group and 58 no had recurrences as a control group and clinical characteristics, method of operation, type of pathology, and postoperative treatment were evaluated between the two groups (6). Ditto in 2019 to identify predictive factors of recurrence and survival in patients with early-stage epithelial ovarian cancer, investigate 429 patients underwent primary surgery followed by either adjuvant chemotherapy or observation alone for apparent eEOC (7).

This study investigates 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.

Materials and Methods

This retrospective study used data from 141 ovarian cancer patients who referred to the oncology and radiotherapy ward of Imam Hossein Hospital in Tehran from 2007 to 2018 years. Data were extracted from patients' medical records; patient's first recurrence status was extracted from patient's chemotherapy and pathology records. Patients who were excluded from our follow-up for any reason were considered as right censored data. Demographic characteristics of patients such as age at diagnosis, body mass index, patient's blood parameters such as white blood cell, hemoglobin, platelet count, clinical data such as stage and grade of tumor, histologic type, metastatic tumor, Type of chemotherapy, chemotherapy courses, and presence ascites at diagnosis were entered into random survival forest.

Statistical analysis

Random survival forest

Random survival forest is a multiple machine learning method that computes a set of survival trees that can be used to select important variables for our event. Computing a group of trees by random survival forest based on two random processes, including bootstrapping and random node splitting. Each Bootstrap sample, on average, consists of two-thirds of the original data. The remaining one-third of the data is excluded from the study and is called out of the bag (OOB); the number of Bootstrap samples is named as the number of trees (8).

The importance of a variable in relation to the time to the incident in the random survival forest can be estimated using the minimum depth so that the minimum depth of the variable is determined by the distance from the root node to the closest node limited to that variable. This value is recorded for each variable and repeated in each tree, finally averaging across all forest trees (9).

The RSF algorithm

1. Bootstrap B samples are selected from the original data. Note that each Bootstrap sample removes about 37% of the data called out-of-bag data (OOB data).

2. A survival tree is grown for each sample of Bootstrap. At each node of the tree, p variables are randomly selected from all variables. The node is split using the candidate variable to maximize the survival difference between the sub-nodes.

3. Grow the tree to the maximum extent so that the terminal node has not less than d >0 unique events.

4. Calculate the cumulative hazard function (CHF) for each tree. To obtain ensemble CHF, the average across all of the trees.

5. Using OOB data, the prediction error for the ensemble CHF is calculated.

Results

Patient demographics

This retrospective study was including 141 patients with epithelial ovarian cancer that 58 patients (41%) had the first recurrence in our follow uptime, so the median time to the first recurrence in these was 17 (0.5-127) months, the median age of the patients in our follow up was 52 (23-82) years. Table 1 shows the properties of the selected continuous and discrete variables, respectively.

Random survival forest analyses

In order to increase the accuracy and efficiency of the forest, the missing data were imputed using output splitting. After imputation in the missing variables, we fitted the random survival forest to the data using the R package randomForestSRC. The number of Bootstraps was B=1000, and the number of randomly selected variables in each node was $q = \sqrt{15} \sim 4$ and log-rank split criterion. The predictive error rate after the growth of 200 trees was 34% for the log-rank splitting rule. As the number of trees increased, the error rate in the domain decreased (Figure 1).

The results of Table 2 show that the log-rank score split rule has the highest error rate and the log-rank criterion performs better than the others. In terms of the significance of predictor variables, random survival forest using log-rank split rule showed that tumor stage using variable importance criterion (VIMP) with a relative importance value of 1.993 and also using minimal depth criterion (Md) with a depth of 2.678, Metastatic tumor with the relative importance of 2.665 and depth of 2.349, and maximum platelet count with the relative importance of 2.132 and depth of 2.683 were the most important variables affecting random survival forest model. The software output is shown in Table 3 based on the minimum depth of one and two orders as well as the VIMP criterion. Strong variables have at least a value less than or equal to the threshold set by the software (Figure 2).

To indicate the graphical relationship between our eight selected variables and the first recurrence of epithelial ovarian cancer, partial plots were calculated based on the random survival forest model (Figure 3).

Table 1. Descriptive statistics for the available covariates

Continuous Covariate	Range	Mean	Median	Categorical Covariate	Category	N	Frequency (%)
Age at diagnosis (years)	23-82	52.7	52	**Metastatic tumor**	yes	52	36.9
					no	80	56.7
				Tumor grade at diagnosis	gradeI	24	17
BMI (kg/m^2)	12.66-39.45	27.15	27		grade II	32	22.7
					gradeIII	37	26.2
Minimum Platelet count	76000-410000	152950	135000	**FIGO Stage at diagnosis**	stage1	34	24.1
					stageII	16	11.3
					stageIII	44	31.2
					stageIV	14	9.9
Maximum Platelet count	178000-819000	381000	353000	**Ascites at diagnosis**	presence of ascite	58	41.1
					no presence of ascite	78	55.3
Mean Platelet count	126000-516000	251190	232750	**Chemotherapy course**	three weeks	64	45.4
					one week	37	26.2
Mean White blood cells	3065-10627	5192	4965	**Neoadjuvant chemotherapy**	adjuvant	64	13.5
					neoadjvant	19	45.4
Mean Hemoglobin	8.36-13.23	10.87	11	**Histologic type**	papillary serous	85	60.3
					others (Endometrioid, Clear Cell,Mucinous)	30	21.2
Minimum Hemoglobin	7.3-12.8	9.66	9.6	--	--	**--**	**--**

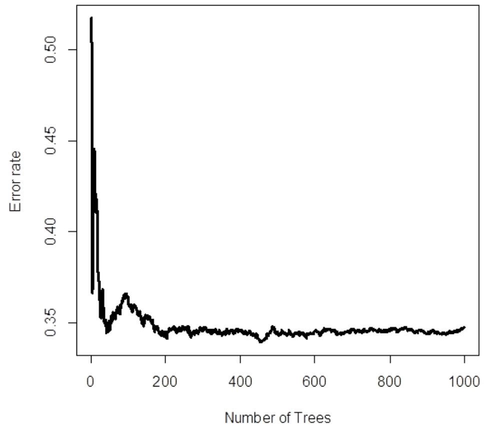

Figure 1. Random Survival Forest OOB prediction error estimates as a function of the number of forest trees

Table 2. Harrell's Concordance error rates for methods

Method		Error rate	
		SE	Mean
RSF	Log-rank	0.008	0.349
	Log-rank score	0.017	0.397
	Random	0.014	0.379

Table 3. Variable importance (VIMP), one and two minimal depth (Md) for selected variables in RSF

Variables	Variable importance (VIMP)	Minimal depth1(MD1)	Minimal depth2(MD2)
Metastatic tumor	2.665	2.428	4.028
Figo stage at diagnosis	1.993	2.793	4.010
Maximum platelet count	2.132	2.683	3.982
Mean platelet count	0.046	3.060	4.001
Minimum platelet count	-0.509	3.508	4.037
Mean white blood cells	-1.738	2.561	4.041
Mean hemoglobin	0.254	3.621	4.060
Minimum hemoglobin	-0.625	3.477	4.038
Age at diagnosis	-0.904	3.246	4.038
BMI(kg/m^2)	-0.811	3.293	4.051
Ascites at diagnosis	-0.440	3.949	4.068
Chemotherapy course	0.301	3.078	4.056
Chemotherapy type	0.602	3.711	4.068
Tumor grade at diagnosis	-0.231	3.636	4.062
Histologic type	0.463	3.843	4.070

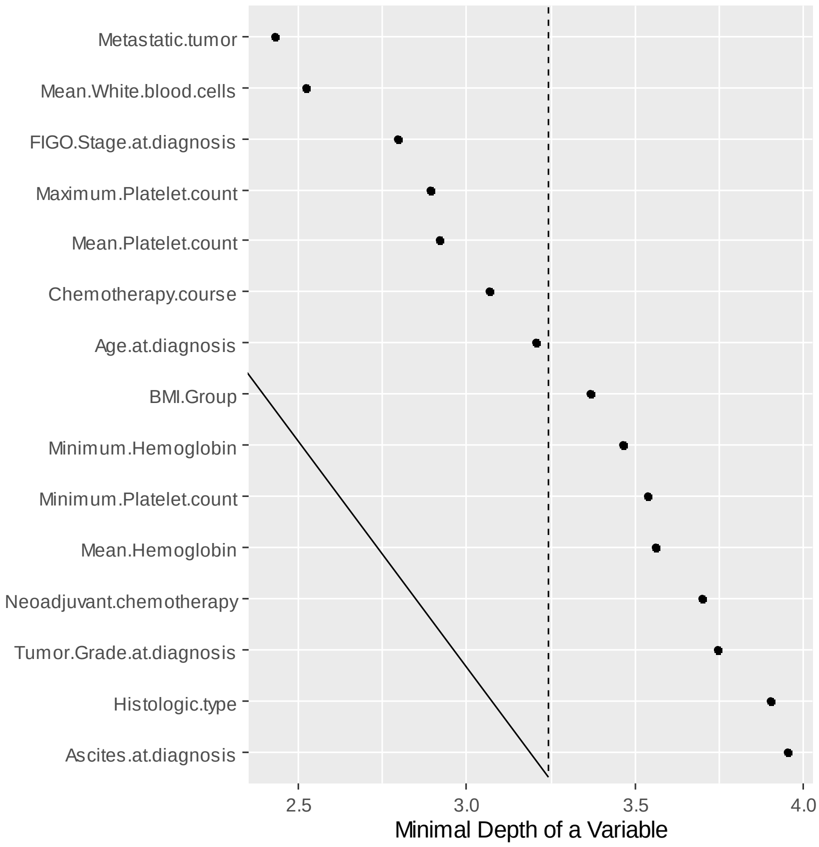

Figure 2. Minimal depth of selected variables. Lower minimal depth indicates variable importance. Dashed line indicates maximum count threshold limit

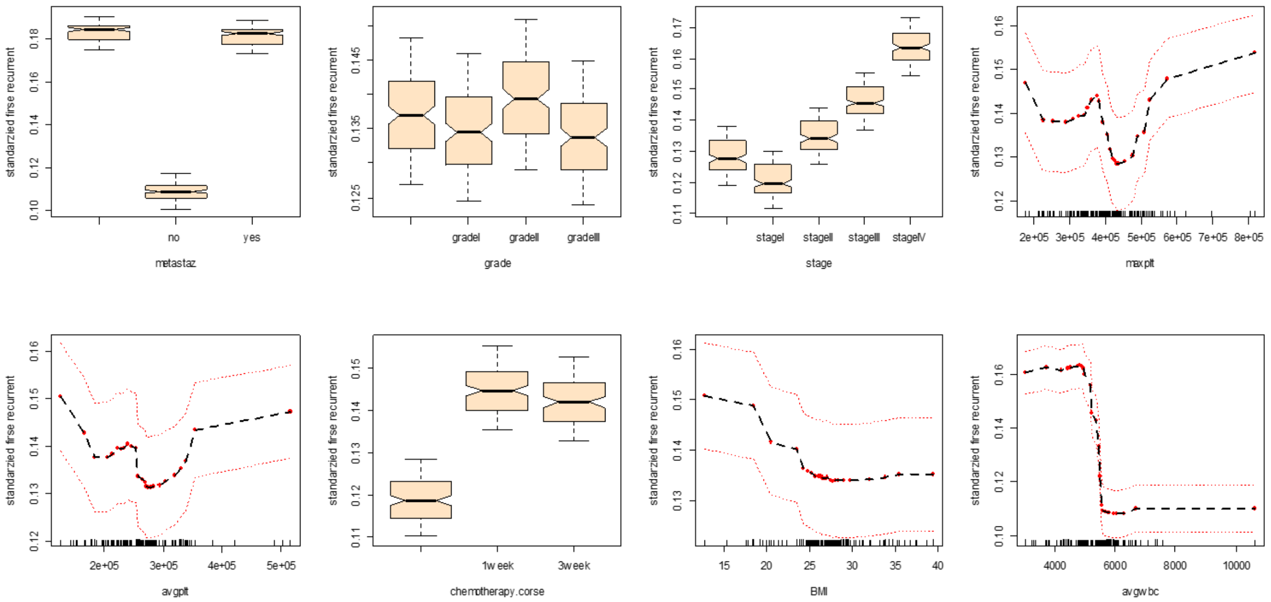

Figure 3. Partial plots of the selected variables most affecting regarding the first recurrence of epithelial ovarian cancer. The plots including the partial values (red points)±2 SE (dashed red lines)

Discussion

In our study, important variables were identified in the first recurrence of ovarian cancer as metastatic tumor, tumor stage, and maximal platelet count that these variables were identified as important variables in the study (10) by Cox and Weibull methods Therefore, random survival forests with more accuracy than Cox and Weibull models without the need for any specific assumptions can better determine the factors affecting our event.

Okunade, in his 2020 study, showed that the recurrence rate of EOC was 76.4%. Suboptimal debulking surgery is the only independent predictor of premature tumor recurrence (5). Xiaoyan, in a 2016 study evaluating independent risk factors for the risk of recurrence of epithelial ovarian cancer, showed that more than 4 (*P*=0.03) cycles of chemotherapy were associated with recurrence of ovarian epithelial cancer. Logistic regression analysis showed that patients with end-stage (Ⅲ/Ⅳ) and positive tumor cells in ascites had a higher risk of recurrence for epithelial ovarian cancer (6). Ditto, in his study in 2019, showed that in multivariable analysis, (FIGO) stage greater than 1 was the only prognostic factor for lower disease-free survival. Also, multivariate analysis showed that age at diagnosis and stage were independent predictive factors for poorer overall survival (OS) (7).

The RSF process is largely automatic, and only a few parameters such as the number of Bootstrap samples or the number of node divisions need to be specified (4). Therefore, the RSF method is a good tool for analyzing data that our prior information is limited about. One of the main strengths of this paper is the use of Random survival forest machine learning, which is suitable for the survival analysis of many variables and correlated variables. For tree growth, RSF uses random subsets of variables in each node. As a result, the correlated variables are chosen independently to divide the nodes and cause a break in the correlation structure of the variables. As a result, there is less competition between highly correlated variables, and the choice of a reliable variable is possible (11).

In addition, the problem of over-fitting when multivariate regression models are performed on a large number of variables is greatly reduced due to randomization through bootstrap samples (12). This feature makes RSF very attractive for the analysis of survival in high-dimensional data, where false positives are considered a major problem due to over-fitting (13).

Limitation

A disadvantage of the RSF model is that the relative risk, which is a meaningful measure of relevance in epidemiological studies, cannot be immediately calculated for the variables considered in an RSF model.

Instead, the role of each variable in its relative relation to the endpoint must be evaluated by ranking the minimum depth and orientation of the partial plots. However, the variables selected by the RSF can be analyzed in subsequent regression models to estimate relative risks. In addition, regression models involving all selected variables may not be appropriate, given the fact that RSF can independently select highly correlated variables (8).

Another disadvantage of tree-based methods is the priority of continuous variables for node division if the data is a combination of continuous and class variables (14). However, as shown, this bias can be eliminated by using the least number of divisions per variable selected for node division. In addition, each RSF is performed based on computing a large number of decision trees, so it is difficult to examine the node splitting process for the relevant variables.

Regarding high recurrence rate of epithelial ovarian cancer and our results that tumor stage and metastatic tumor were identified as two important factors in the early recurrence of this cancer, one of the causes of more recurrence can be identified as late diagnosis. Therefore, further planning is needed to detect earlier this cancer.

When the number of predictor variables is high due to the interaction effects of the variables, regression methods are not suitable; the nonparametric model of classification and regression tree without any specific assumptions can estimate the likelihood of recurrence under different subgroups. Nonparametric random forest survival by averaging over the results of regression and classification trees yields far more accurate results than a single tree. These models can be used by physicians and paramedics because of their lower prediction error than parametric and semi-parametric models and because of the easy interpretation.

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 Chemother Pharmacol 2019;83:97-105.
2. Paik ES, Lee J-W, Park J-Y, Kim J-H, Kim M, Kim T-J, et al. Prediction of survival outcomes in patients with epithelial ovarian cancer using machine learning methods. J Gynecol Oncol 2019; 30: e65.
3. Lipson R. Predicting Ovarian Cancer Survival Times: Performance of Parametric Methods and Random Survival Forests [dissertation]. Canada: Simon Fraser University., 2014.
4. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat 2008;2:841-60.
5. Okunade KS, Adetuyi IE, Adenekan M, Ohazurike E, Anorlu RI. Risk predictors of early recurrence in women with epithelial ovarian cancer in Lagos, Nigeria. Pan Afr Med J 2020;36:272.
6. Xiaoyan L. Risk factors of epithelial ovarian cancer recurrence. Modern Oncol 2016;24:2283-5.
7. Ditto A, Leone Roberti Maggiore U, Bogani G, Martinelli F, Chiappa V, Evangelista MT, et al. Predictive factors of recurrence in patients with early-stage epithelial ovarian cancer. Int J Obstet Gynecol 2019;145:28-33.
8. 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 [dissertation]. Berlin: University of Berlin., 2016.
9. Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. High-dimensional variable selection for survival data. J Am Stat Assoc 2010;105:205-17.
10. Deldar M, Anbiaee R, Jalilian A, Sayehmiri K, Azimi S. Comparison of Cox's Regression Model and Weibull' Parametric Model in Evaluating Factors Affecting in First Recurrence of Epithelial Ovarian Cancer. Acta Med Iran 2020;58:445-51
11. Siroky DS. Navigating random forests and related advances in algorithmic modeling. Stat Surv 2009;3:147-63.
12. van der Schaaf A, Xu C-J, van Luijk P, van't Veld AA, Langendijk JA, Schilstra C. Multivariate modeling of complications with data driven variable selection: guarding against overfitting and effects of data set size. Radiother Oncol 2012;105:115-21.
13. Broadhurst DI, Kell DB. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics 2006;2:171-96.
14. Loh WY, Shih YS. Split selection methods for classification trees. Stat Sin 1997;7:815-40.