International Journal of Statistics and Applications 2021, 11(2): 37-49
DOI: 10.5923/j.statistics.20211102.03

Analysis of Balanced Random Survival Forest Using Different Splitting Rules: Application on Child Mortality

Hellen Wanjiru Waititu[1,*], Joseph K. Arap Koske[1], Nelson Owuor Onyango[2]

[1]School of Physical and Biological Sciences, Moi University, Eldoret, Kenya
[2]School of Mathematics, University of Nairobi, Nairobi, Kenya

Abstract Under Five Child Mortality (U5CM) remains a major health problem in the developing world. The Sustainable Development Goals target of 25 deaths per 1000 live births has not yet been achieved in many Low and Middle Income Countries (LMIC). This study used the Kenya Demographic and Health Survey (KDHS) data (2014) to understand the determinants of U5CM. KDHS (2014) data is characterized by high dimensionality, high imbalance and violation of Proportional Hazard (PH) assumptions among other statistical challenges. This study aimed at handling the problem of non proportional hazard assumptions that characterize covariates of survival regression models. To achieve this we used various split rules, namely: log-rank, log-rank score and Bs.gradient splitting rules. The data used was balanced using Random Under-sampling method. The balanced data was integrated in RSF for variable selection while applying the three specified splitting rules. Respective selected variables were fitted in the Cox Aalen's model for prediction while model selection was carried out using concordance index. The model with log-rank splitting rule recorded the highest concordance of 0.916 followed by Bs.gradient with a concordance of 0.864 while log-rank score resulted in a concordance of 0.799. In conclusion, the results from the analysis presented in this paper show the superiority of log-rank splitting rule. However, optimality of log-rank is achieved when the hazard is proportional over time. Some of the variables in the data were found to violate the PH assumption making the use of log-rank splitting rule not optimal. According to our analysis, we settle on Bs.gradient splitting method which still has a high concordance index of 0.86 and smaller error rate of 0.028. Using Balanced Random Survival Forests (BRSF) with Bs.gradient splitting rule, the identified determinants of U5CM are; V207 (sum of deceased daughters), V219 (sum total of living children) and B8 (age of the child). Hence, the age of the child and the siblings' information are identified as some of the key determinants of U5CM.

Keywords Splitting rules, Balanced Random Survival Forests, Under Five Child Mortality, Cox Aalen's model

1. Introduction

In survival analysis, censored survival data are frequently predicted using semi-parametric methods. One of the most commonly used semi-parametric method in analysis of time to event data is the Cox Proportional Hazard (PH) method [1]. The Cox model estimates survival by evaluating various explanatory variables all at once. Other than being semi-parametric, Cox PH model is also popular due to its ability to produce adequately good regression coefficient estimates, survival curves and hazard ratios of interest for a wide variety of data occurrence [2]. However, the method makes assumptions which are not easily satisfied. One such assumption is that the hazard ratio between any two observations is proportional over time. In addition, Cox PH model does not take in to account the missing predictors, non-linearity of exponential factors and interdependence among observations.

To deal with these limitations, nonparametric survival trees and forests have recently become useful alternatives. In 2008, [3] introduced Random Survival Forests (RSF), a fully non parametric ensemble tree based method for analysis of right censored survival data. The method can determine survival risk factors without assuming parametric associations. RSF generates a forest by randomly selecting a given number of bootstrap samples from the data in use. Each of the bootstrap samples develops into a tree through recursive partitioning of the covariate space. The trees then grow to full size until the end most node has no less than a predetermined number of exclusive events. RSF has the ability to discover non-linear effects, impute missing data and discover interactions beforehand.

However, characteristic of survival data pose significant challenges to RSF. In this study, we worked with the 2014 Kenya Demographic and Health Survey (KDHS) data in trying to understand the determinants of Under Five Child Mortality. Some of the statistical challenges with the 2014

* Corresponding author:
hwaititu@cuea.edu (Hellen Wanjiru Waititu)
Received: Jul. 20, 2021; Accepted: Aug. 6, 2021; Published: Aug. 15, 2021
Published online at http://journal.sapub.org/statistics

KDHS data include high dimensionality, high imbalance between mortality and non mortality classes and violation of Proportional Hazard (PH) assumptions among others. This study aimed at handling the problem of non proportional hazard assumptions that characterize covariates of survival regression models as well as dealing with high imbalance between mortality and non mortality classes. The aspect of high imbalance and violation of PH assumptions are often ignored by many researchers which can result to biased and inaccurate results.

When working with highly imbalanced datasets, machine learning algorithms like RSF leads to biased results supporting the majority class. Additionally, highly imbalanced datasets poses significant challenges in RSF due to the stopping criterion which states that the terminal node should have no less than a predetermined number of unique events. This may result in premature termination of the tree especially when the mortality class is too small compared to the non mortality class.

In 2020, [4] compared the performance of RSF in data balanced using different methods where random under-sampling method performed best in model selection using concordance index. In this paper, we used data balanced using random under-sampling method where the mortality and non mortality classes have equal number of observations.

The problem of variables violating PH assumptions has often been ignored by many researchers. Some survival analysis methods such as the Cox PH models apply the restriction of satisfaction of PH assumption. During node splitting process in RSF, the most commonly used splitting rule is log-rank test. The optimality of log-rank test is achieved when the hazards are proportional over time. Proportional hazard (PH) assumption indicates that the effect of covariate is the same at all points. However this is usually not the case since in many instances, variables violate the PH assumption making the use of log-rank test and Cox PH model among others not optimal.

In some cases, researchers ignore this assumption leading to inaccurate conclusions. Others delete variables that violate the PH assumption in order to work with the restricting methods. However, the deleted variables could be important predictors of mortality. There is need to therefore work with methods that take into consideration these statistical challenges.

Quite a number of studies have researched on splitting rules used in RSF. These include [5] who proposed an improved RSF by using weighted log-rank test in splitting the node while using the model of [6]. [7] In 2018 used R-squared splitting rule in survival forests, [8] compared RSF using different splitting rules among others. [9], [10], [11], [12] and [13] among others used Log-rank test for survival splitting as a measure for maximizing survival difference between nodes.

In this research, we analyze the performance of BRSF using different splitting rules while using data balanced using under-sampling method. This was done using 2014 KDHS data in identification of determinants of Under Five Child Mortality (U5CM).

The rest of the paper is laid out as follows: Section 2 discusses the methodology employed in this study, from description of the data, exploration of PH assumption, the theory behind RSF and node splitting, the different splitting methods used, survival tree estimators, effect of predictors using Cox Aalen's model and finally model selection criteria using concordance statistic. Section 3 summarizes the results of the study with respect application of RSF, variable selection using RSF with the different splitting methods and model selection using concordance statistic. Finally, section 4 offers a discussion of our results against other ongoing research.

2. Methodology

2.1. Data Description and Management

The data used in this research was obtained from the 2014 Kenya Demographic and Health Survey (KDHS) data [14]. KDHS is a national proceeding which is carried out every five years in the country all the way back from 1989. Thus, 2014 KDHS is the 6th Demographic and Health Survey (DHS) operation in Kenya. The survey is headed up by the Kenya National Bureau of Statistics (KNBS) which is the leading Government agency for official statistics in collaboration with the National Council for Population Development, Ministry of Health, and a number of development partners. Their aim is to collect data required for health, nutrition, planning, monitoring and evaluation of population among others. 2014 KDHS is a sample survey data in which households are randomly chosen from the KNBS sampling frame. The units of analysis in this data are household, individual, children aged o to 5 years, woman aged 15 to 49 years and man aged 15 to 54 years.

After successful application and acquisition of the 2014 KDHS data, it was downloaded and viewed using Statistical Package for Social Sciences (SPSS). The overall data had 1099 variables and 20994 observations. From the overall data, variable with 100% missing observations and others which were highly correlated were deleted. At the same time, status and time variables which are of great importance in survival data were calculated and included in the data. Time from birth to date of interview was taken as the follow-up period. This resulted to a total of 786 variables while the observations remained at 20964. The data was found to be highly imbalanced with the mortality class taking 4% of the data as seen in [4]. Various covariates were similarly found to be highly imbalanced resulting to a range of 3-6% of the mortality class.

The data was categorized into regions which are equivalent to the former provinces in Kenya. Since our interest was to deal with a relatively smaller sized data, we analyzed data from Nairobi region which is a unique urban setting with diverse levels of social economic status among populations.

The Nairobi region data consisted of 532 observations and 757 variables. This was after removal of variables with 100% missing observations in the region. Other variables like region, residence, county which had similar response in the region were also deleted from the data. Nairobi region dataset was similarly highly imbalance. The mortality class representation in this region was 6.4% while the range of mortality representation in the covariates was 0-7% as seen in our previous work, [4]. Data balancing was performed using 4 methods. Model selection using concordance index showed good performance in random under-sampling method followed by synthetic minority oversampling technique (SMOTE).

In this paper, we analyze the performance of the under-sampled dataset using RSF with different splitting rules. The dataset is fully balanced after random under-sampling with each class taking half of the sample size. The balanced dataset had a total of 68 observations with the mortality and censored classes each having 34 observations which represent 50% of the sample. The number of variables in the dataset was 757.

2.2. Proportional Hazard (PH) Assumption

Hazard is the likelihood of an event happening at any given time point given that the event had not occurred. The PH assumption indicates that the ratio of the hazard comparing any two specifications of covariates is unchanging or proportional over time. It is essential to verify whether the predictor variables in the model satisfy the PH assumptions. We used Kaplan Meier curves to explore the PH situation.

The Kaplan Meier curves plots the estimated proportion at risk (survival probability) against time giving the estimated survival functions [15]. The curves are in form of step functions with each vertical drop pointing out one or more deaths happening [16].

If the variables satisfy PH assumption, the survival curves should be parallel. If for two or more categories of a variable of interest do not result to parallel curves or the curves cross, then it is an indication that the PH assumption is violated.

Before the exploration of PH assumption, we got the general view of the survival data by demonstrating the occurrence of event during the follow-up period as shown in table 1 and the survival curve in figure 1.

Table 1. Survival table for the under-sampled dataset

Time	No. at risk	No. of events	Survival	Std.Error
0	68	18	0.735	0.0535
1	50	1	0.721	0.0544
2	49	1	0.706	0.0553
3	48	3	0.662	0.0574
5	44	1	0.647	0.0580
6	43	1	0.632	0.0586
7	42	1	0.617	0.0591
9	39	2	0.585	0.0601
11	37	1	0.569	0.0606
12	36	2	0.538	0.0612
17	32	1	0.521	0.0615
19	30	1	0.503	0.0619
24	28	1	0.485	0.0622

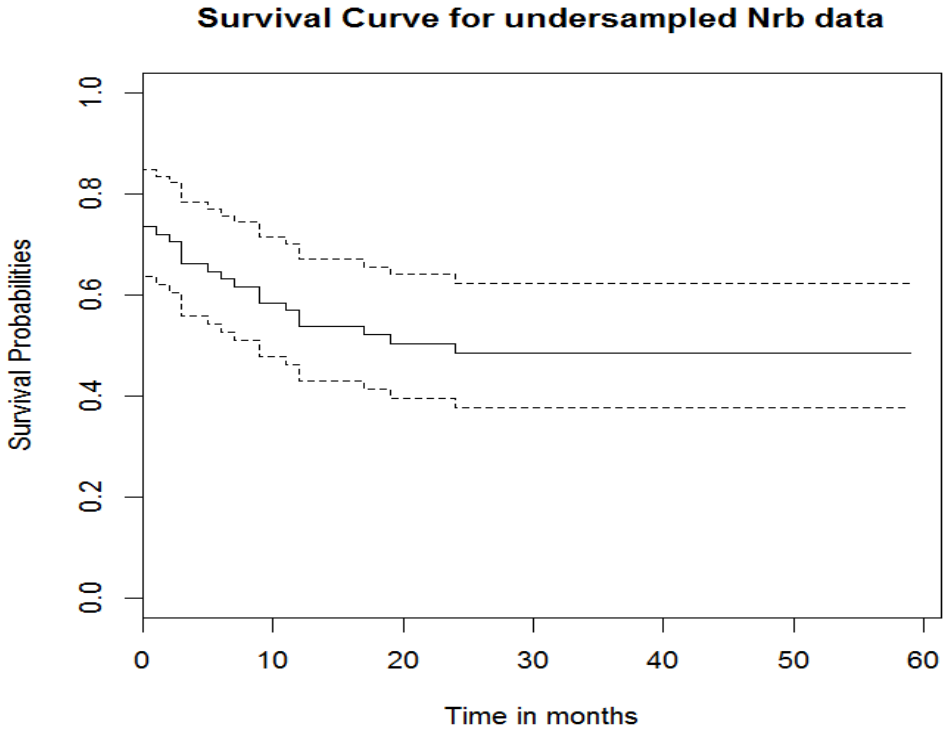


Figure 1. General survival curves for the data used

From the survival curve, the horizontal axis indicates time in months starting from 0 to 60 months while the y axis indicates the survival probabilities or the proportion of individuals surviving (at risk). The curve is a step function with each step downwards representing one or more deaths occurring. At time 0, the proportion of individuals surviving is approximately 0.7. This is because 18 individuals got an event before the first month was over. The curve then drops gradually as few Individuals continue to die up to 24 months after which the survival probability is constant at about 0.048. No more events were experienced after 24 months. The survival probability gives the probability of an individual surviving past the specified time.

2.3. Exploration of Proportional Hazard

The aspect of hazard being constant over time was explored using Kaplan Meir curves. Figure 2 shows Kaplan Meir curves for some of the covariates.

From the survival curves in figure 2, there is evidence of violation of PH assumption by some of the covariates as demonstrated by the presence of crossing curves. Survival curves by child sex are almost parallel with female children having a higher survival probability throughout the period than the male children. The two curves do not cross but they are not perfectly parallel. From survival curve by level of education, there exist individuals with no education but none of them got an event. Hence the horizontal blue line with survival probability of 1. Crossing curves are observed between secondary level and higher education level showing violation of PH assumption. The survival probability for the secondary education category is higher than that of higher level from 0 to about 12 months when the curves cross. From 12 to about 25 months, individuals in higher level show a higher survival probability than those in secondary level. Individuals with primary level of education are observed to have a higher survival probability than those in secondary and higher education level. Similarly, survival curve by wealth index indicate violation of PH assumption since the curves for poorer and middle level categories are crossing.

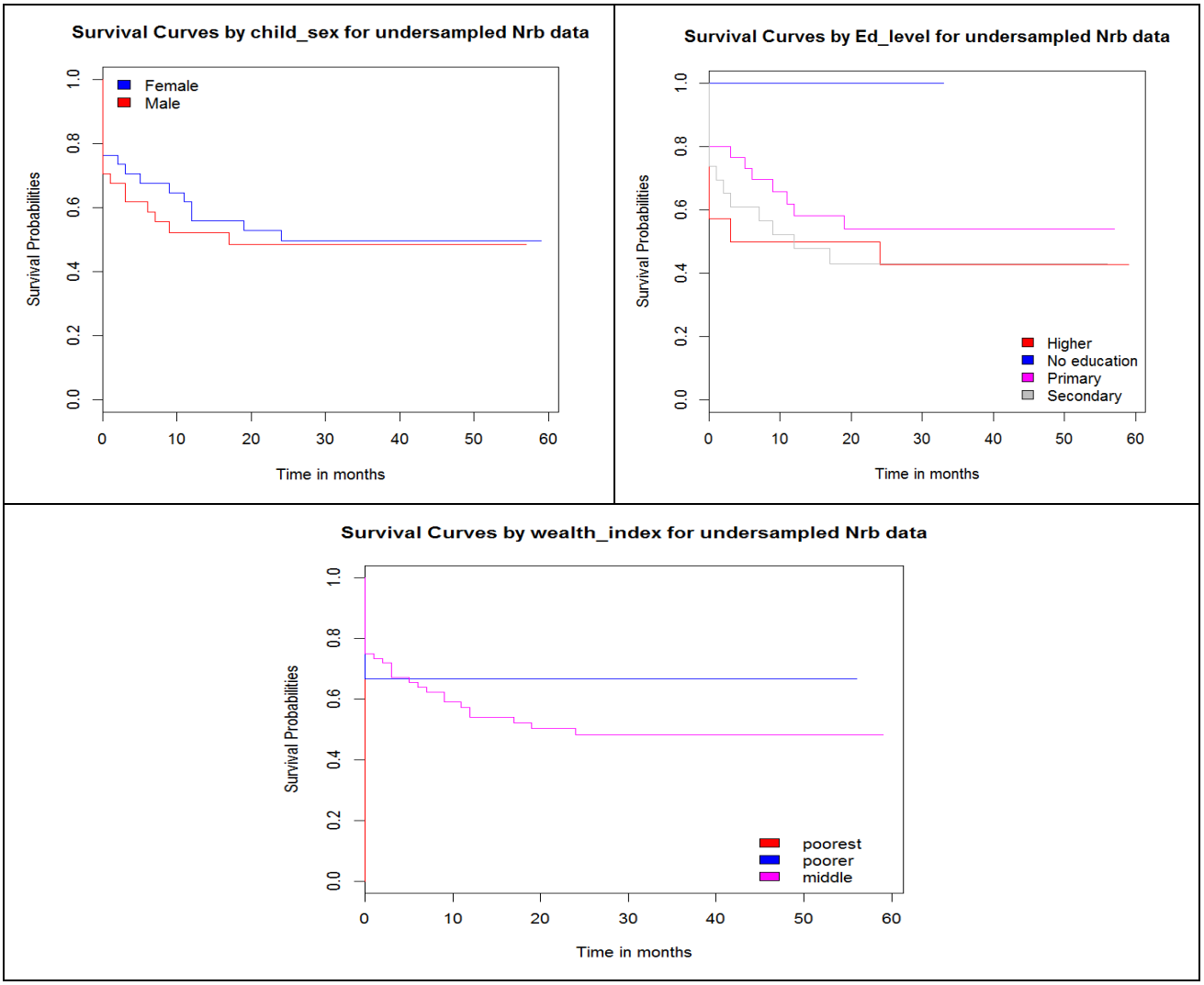


Figure 2. Survival curves by covariates

2.4. RSF Algorithm for Variable Selection

The balanced data was integrated with RSF algorithm which is described by [3] as follows:

a) The process starts by drawing at random $ntree$ bootstrap samples from the original data having G samples. On average, each bootstrap sample sets aside 37% of the data named as Out of Bag (OOB) data with respect to the bootstrap sample and each sample has R predictors.
b) For each bootstrapped sample, a survival tree is developed. This is done by randomly choosing $mtry$ out of R variables in x for splitting on. The value of $mtry$ depends on the number of available predictors and is data specific. All the $ntree$ bootstrap samples are designated to the top most node of the tree which is also referred to as the root node. This root node is then separated into two daughter nodes each of which is recursively split progressively maximizing survival difference between daughter nodes.
c) The trees are grown to full size where the end is indicated by the restriction that the endmost node should have larger than or equal to $nodesize$ unique events.
d) After the tree is fully grown, the in-bag and out of bag (OOB) ensemble estimators are computed by taking the mean value over all the trees predictors.
e) The ensemble OOB error is calculated using the first b trees, where $b = 1, \dots, ntree$.
f) OOB estimation is used to calculate the Variable Importance (VIMP) [17].

By averaging over all trees, a reliable measure of importance of a variable regarding time to event can be obtained [18].

RSF gives a measure of VIMP which is totally nonparametric. VIMP has been found to be effective in many applied settings for selecting variables [19], [20], [21], [22], [23]. In this study, using the RSF model, the highly predictive risk factors using three different splitting rules were extracted.

2.4.1. Splitting the Node in RSF

Time and status variables are of great importance in survival data. The time variable indicates the survival duration while status variable indicates whether the observation experienced an event or was censored. The actual survival time of censored observation cannot be calculated since censored observations do not terminate. The only indication of known survival duration is the occurrence of an event. The presence of censoring in survival data complicates certain aspects of implementing RSF [17]. While taking into account right censoring, the observed data is given in the form (X, δ) where X is defined as the minimum of the event and censoring time. Hence $X = \min(T, C)$ where T is the event time and C the censoring time. δ is the censoring indicator defined as

$$\delta = \begin{cases} 1 \; if \; T < C. ie. \text{event occured} \\ 0 \; if \; T > C. ie. \text{observation is censored} \end{cases}.$$

While growing a tree in RSF, node splitting must take censoring into consideration.

With reference to the RSF algorithm, a forest develops from randomly drawn $ntree$ bootstrap samples each of which becomes the root of each tree in the forest. There being R predictors in each bootstrap sample, $mtry$ predictors are randomly chosen for splitting on. Suppose we take h to be the top most node of the tree which is to be split into two daughter nodes. Within the node, there exist $mtry$ predictors and n observations. The splitting process is as follows [24].

- Take any predictor X from the $mtry$ predictors.
- Find the splitting value c such that the survival difference between $x \leq c$ and $x > c$ for predictor x is maximum. In this case $x \leq c$ splits to the left daughter node while $x > c$ splits to the right daughter node.
- Calculate the survival difference between the two daughter nodes using a pre-determined splitting method.
- Take another split value c in predictor x until we get a split value which results in maximum survival difference for predictor x.
- From the remaining $mtry - 1$ predictors in the node the process is repeated until we get predictor x^* and split value c^* which give maximum survival difference between the two daughter nodes.
- Applying the node splitting process in each of the new daughter nodes and recursively partitioning the nodes leads to the growth of the tree.
- The process is applied to all the $ntree$ root nodes leading to the growth the forest.

When survival difference is maximum, unlike cases with respect to survival are pushed apart by the tree. Increase in the number of nodes causes dissimilar cases to separate more. This results in homogeneous nodes in the tree consisting of cases with similar survival.

In this research, the following splitting methods were used to calculate the survival difference between any two daughter nodes.

2.4.2. Log-Rank Splitting Rule

Log-rank splitting rule separates the nodes by selecting the split that yields the largest log-rank test. The log-rank test is the most frequently used statistical test to compare two or more samples non-parametrically in censored data. PH assumption is the key requirement for the optimality of log rank test.

Suppose we want to split node h of a tree using log-rank splitting rule. The data at the node is presented as $(X_1, T_1, \delta_1), \dots, (X_n, T_n, \delta_n)$ where X_i is the i^{th} predictor, T_i and δ_i represent the i^{th} survival duration and censoring status respectively.

The information at time t_i can be summarized as in the table below.

Time t_i	Event set	Survivors	Risk Set
Node 1	$d_{i,1}$	$Y_{i,1} - d_{i,1}$	$Y_{i,1}$
Node 2	$d_{i,2}$	$Y_{i,2} - d_{i,2}$	$Y_{i,2}$
Total	d_i	$Y_i - d_i$	Y_i

Where, $d_{i,j}$ stands for the number of events in daughter node $j = 1,2$ at time t_i, $d_i = d_{i,1} + d_{i,2}$.

$Y_{i,j}$ represent individuals who are alive in daughter node j, $j = 1,2$ at time t_i. $Y_{i,1}$ is the number of $T_i \geq t_i$, $x_i \leq C$, where T_i is the duration of survival for the i^{th} individual and t_i the distinct event time in node h.

$Y_{i,2}$ is the number of $T_i \geq t_i$, $x_i > C$.

$Y_i = Y_{i,1} + Y_{i,2}$

For a split using covariate x and its splitting value c, The survival difference between any two daughter nodes is calculated using log-rank test given as;

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

This equation measures the magnitude of separation between two daughter nodes. The best split is given by the greatest difference between the two daughter nodes which is given by the largest value of $|L(x,c)|$ [25].

2.4.3. Log-Rank Score Splitting Rule

Log-rank score splitting rule [26] was developed from log-rank split rule. The ranks for each survival time T_l are computed given an ordered predictor x such that $x_1 \leq x_2 \leq \cdots \leq x_n$. The rank for each time T_l is calculated as

$a_l = \delta_l - \sum_{k=1}^{\Gamma_l}\frac{\delta_k}{n-\Gamma_l+1}$ where $\Gamma_l =$ the number of $(t: T_t \leq T_k)$. Let $\bar{a}$ and s_a^2 be the sample mean and sample variance for a_l for $l \in 1,2,\ldots,n$. The formula for log-rank score test is given by

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

This split rule gives the magnitude of node separation by $|S(x,c)|$ where the best split is given by the maximum value over x and c.

2.4.4. Gradient-Based Brier Score (Bs. Gradient) Splitting Rule

Brier Score (BS) is the most frequently used scalar summary of correctness for probability predictions for binary events. Let $y_i, i = 1,2,\ldots,n$ be the i^{th} likelihood prediction in a series of n such predictions. The paired observation is given as $x_i = 1$ if the event of interest occurs on the i^{th} occasion, and $x_i = 0$ otherwise. The BS is the mean-squared error over the n pairs of prediction observations,

$$BS = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2$$

In this case, the time horizon used for the Brier score is set to the 90^{th} percentile of the observed event times which must be a value between 0 and 1. Suppose we have a pair of predictor-response, say (X_i, Y_i) for $i = 1,2,\ldots,n$. The usual regression procedure attaches the conditional average of the response variable Y to a specified set of predictors X. [27] introduced Quantile Regression Forests (QRF) which connects between an empirical cumulative distribution function and the outputs of a tree.

Let D_0 be a group of randomly selected variables to be split into two daughter nodes D_1 and D_2. Suppose the homogeneity of each group is defined by

$v(D_j) = \sum_{Y \in D_j}\left[Y - \bar{Y}(D_j)\right]^2$ where $\bar{Y}(D_j)$ is the sample mean in D_j. For an optimal splitting selection, comparison is done between the homogeneities of $v(D_1)$ and $v(D_2)$ with that of $v(D_0)$. The splitting value s is the one that maximizes

$$H(D_1, D_2) = \max_{s \in \varepsilon^*}[v(D_0) - v(D_1) - v(D_2)]$$

Where ε^* is a randomly selected sample of predictors from the predictor space ε. The resulting nodes are recursively split until the stopping criterion is reached. The terminal node gives the predicted value. [28] Suggested that instead of maximizing variance heterogeneity of the daughter nodes, one maximizes the criterion

$\Delta(D_1, D_2) = \sum_{j=1}^{2}\frac{-1}{i:Y_i \in D_j}\left(\sum_{\{i:Y_i \in D_j\}}\rho_i\right)^2$ where $\rho_i = 1\left(\{Y_i \geq \hat{\theta}_q, D_0\}\right)$ is an indicator function which takes a value of 1 when Y_i is more than the q^{th} quantile $\hat{\theta}_q, D_0$ of the observations of node D_0. The selection of ρ_i is connected with a gradient based approximation of the quantile function $\Psi_{\hat{\theta}_q, D_0}(Y_i) = q1\{(Y_i > q\}) + (1-q)1(\{Y_i \leq q\})$, hence the term gradient forest. The order for each split is chosen among given orders $(0.1, 0.5, 0.9)$.

2.5. Survival Tree Estimators

In RSF, the tree growing process begins by randomly selecting $ntree$ bootstrap samples from the original data. Each bootsrap sample sets aside on average 37% of the data called OOB data while the remaining 63% is called the in-bag data. The in-bag data is used to grow the tree and gives estimators which are used for prediction. On the other hand, the OOB data is not involved in the growth of the tree but used for cross-validation purposes. RSF estimates cumulative hazard function (CHF) and survival function based on the terminal nodes using the in-bag and out-of-bag estimators.

2.5.1. In-Bag Estimators

Let: h denote the terminal node of a tree.

$t_{1,h} < t_{2,h} < \cdots < t_{m(h),h}$ Indicate the distinct event

times within node h,

$d_{j,h}^*$ Indicate the number of deaths at time $t_{j,h}$ and

$Y_{j,h}^*$ Indicate the number of individuals at risk at time $t_{j,h}$.

The CHF for node h is approximated using the bootstrapped Nelson–Aalen estimators;

$H_h^*(t) = \sum_{t_{j,h} \le t} \frac{d_{j,h}^*}{Y_{j,h}^*}$. This implies that for a given tree, the hazard estimate for node h is the ratio of events to individuals at risk summed across all unique event times. Each terminal node of a tree provides a sequence of such estimates and each individual in node h has the same CHF.

The survival function for node h is estimated using bootstrapped Kaplan Meier estimator;

$S_h^*(t) = \prod_{t_{j,h} \le t} \left(1 - \frac{d_{j,h}^*}{Y_{j,h}^*}\right)$. This gives the estimates for the individuals in node h at a given time t.

To estimate the CHF for a given predictor X, $H(t/X)$ and the survival function of a given predictor X, $S(t/X)$, X is dropped down the tree and ends up in a distinct endmost node as a result of the binary nature of the tree. This implies that $H^*(t/X) = H_h^*(t)$ and $S^*(t/X) = S_h^*(t)$.

This defines the CHF and survival function for all individuals in the data and the estimates for the tree. Due to bootstrapping (sampling with replacement) an observation can be found in various bootstrap samples and hence in various trees.

The in-bag ensemble estimators are computed by averaging the trees estimators. Hence the in-bag ensemble CHF and survival estimators are respectively given as

$$\bar{H}_e^*(t/X) = \frac{1}{ntree}\sum_{b=1}^{ntree} H_b^*(t/X) \text{ and}$$
$$\bar{S}_e^*(t/X) = \frac{1}{ntree}\sum_{b=1}^{ntree} S_b^*(t/X)$$

2.5.2. Out-Of-Bag (OOB) Estimators

Let I_i be an indicator pointing to whether case i is in-bag or out of bag such that

$$I_i = \begin{cases} 0 \; if \; i \text{ is in} - \text{bag} \\ 1 \; if \; i \text{ is out} - \text{of} - \text{bag} \end{cases}$$

To determine the CHF and survival estimators for an OOB case i, the case is dropped down the tree to a endmost node h. The OOB CHF and survival estimators for i becomes

$H^{**}(t/X) = H_h^*(t)$ and $S^{**}(t/X) = S_h^*(t)$ respectively.

The OOB ensemble estimators are calculated by getting the mean of the OOB tree estimators. Hence the OOB ensemble estimators are given as

$$\overline{H}_e^*(t/X_i) = \frac{1}{ntree}\sum_{b=1}^{ntree} H_b^*(t/X_i) \text{ and}$$
$$\bar{S}_e^*(t/X_i) = \frac{1}{ntree}\sum_{b=1}^{ntree} S_b^*(t/X_i)$$

2.6. Determining the Variable Effects

RSF gives a measure of Variable Importance (VIMP) which is totally nonparametric. In this study, the highly predictive risk factors from the balanced dataset were extracted using the RSF model. The extracted important predictors were then examined for PH assumption satisfaction. This was done by examining scaled Schoenfeld residuals and use of statistical tests shown in table 4. Since the PH assumption did not hold for all the variables, we took into consideration analyses that take into account time-varying effects.

Cox-Aalen's model proposed by [29] is one of the tools for handling the problem of non-proportional effects in the variables. The model provides a way of including time-varying covariate effects. In a datasets where some of the covariates effects are constant while others are not constant, Cox-Aalen's model is a better alternative since it combines the two types of covariates. In the model, the covariates are subdivided into two parts in which one part act additively on the intensity while the other work multiplicatively. The Cox Aalen's model is defined by,

$$h(t\backslash x) = Y(t)[X(t)^T \alpha(t)]\exp(Z(t)^T)\beta \text{ where,}$$

$Y(t)$ is the indicator of the risk. $(X(t), Z(t))$ is a $p+q \times 1$ vector of covariates where $X(t)$ is the additive non parametric time varying covariate and $Z(t)$ are the covariates with constant multiplicative effects. $\alpha(t)$ Is a $p \times 1$ vector of time varying regression coefficients and β is a $q \times 1$ vector of relative risk regression coefficients.

Comparison of prediction accuracy of the different models was done based on concordance index.

3. Results

3.1. Variable Selection Using BRSF with Different Splitting Rules

Table 2. Application of BRSF using different splitting methods

Description	Log-rank	Logrank score	Bs.gradient
Sample size	68	68	68
N0. of deaths	34	34	34
N0. of trees	1000	1000	1000
Forest terminal node size	15	15	15
Average no. of terminal nodes	2.523	3.367	2.2469
No. of variables tried at each split	28	28	28
Total no. of variables	757	757	757
Resample size used to grow trees	43	43	43
Number of random split points	10	10	10
Error rate	12.73	26.47	14.16

From the BRSF output in table 2, a forest of 1000 trees was developed from a dataset consisting of 757 variables. This was done by randomly selecting 1000 bootstrap samples from the initial dataset consisting of 757 variables and 68 observations. The dataset was balanced in that out of the 68 observations, 34 were censored and 34 had acquired an event. Trees become fully grown when the most extreme

node has 15 events. The same balanced dataset was used to generate trees using the specified splitting rules. Hence, a lot of similarity in the output of results in table 2 is observed. The only difference observed is in the error rate and average number of terminal nodes. This difference is brought about by the different splitting rules used. Log-rank splitting rule splits the nodes with greater accuracy returning an OOB error rate of 12.73% followed by Bs.gradient rule with 14.16% while Log-rank score yielded the highest error rate of 26.47%. From the average number of terminal nodes which were generated, Bs.gradient had the least followed by log-rank test while log-rank score had the highest number of terminal nodes. The identified predictors based on BRSF using different splitting rules are presented in table 3.

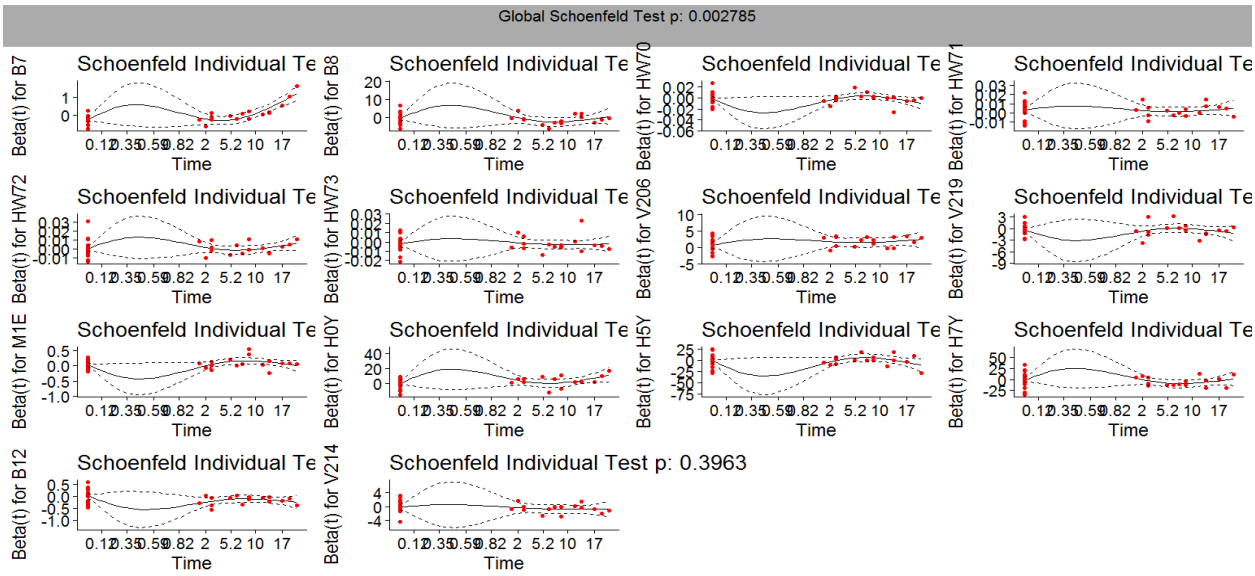
Table 3. Important variables extracted from BRSF using different splitting rules. The (selected variables had a variable importance > 0.02. For variable names, refer to Appendix)

logrank		logrankscore		Bs.gradient	
variable	importance	Variable	importance	variable	importance
B7	0.0263	HW71	0.0038	B7	0.0180
B12	0.0141	B16	0.0033	HW73	0.0072
HW70	0.0075	H5Y	0.0031	HW72	0.0069
HW72	0.0062	HW70	0.0030	B12	0.0061
HW73	0.0055	B12	0.0026	HW70	0.0060
B8	0.0053	H6Y	0.0025	B8	0.0055
B16	0.0047	H2Y	0.0024	HW71	0.0050
V219	0.0045	M6	0.0023	V219	0.0047
HW71	0.0044	H4Y	0.0023	B16	0.0046
V206	0.0041	H3Y	0.0023	V506	0.0039
V220	0.0033	H8Y	0.0020	H9Y	0.0032
M1E	0.0031	ML1	0.0020	H7Y	0.0029
H3Y	0.0027	V626	0.0020	V321	0.0026
H9Y	0.0025	V218	0.0020	H5Y	0.0025
V207	0.0020			V220	0.0024
				V3A07	0.0023
				V214	0.0023
				V207	0.0022

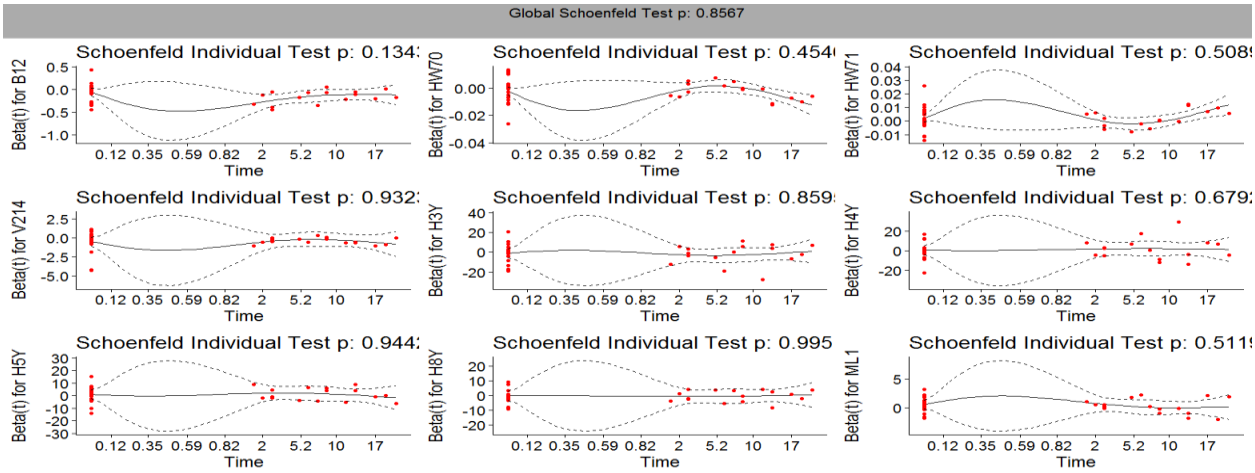
Table 4. Statistical Tests (Test for PH assumption). PH assumption is supported by non significant P-values

Log-rank				Log-rank score				Bs.gradient			
var	rho	chisq	p	var	Rho	chisq	P	var	rho	chisq	p
B7	0.5807	25.03	5.65e-07	B12	-0.2683	2.72	0.0991	B7	0.6730	28.78	8.12e-8
B8	-0.1492	0.2627	0.608	HW70	0.0587	0.297	0.5856	B8	0.0378	0.0182	0.893
B12	-0.3266	6.3360	0.0118	HW71	-0.0643	0.352	0.5529	B12	-0.3653	7.0356	0.0079
HW70	-0.0795	0.4568	0.499	V218	-0.2177	3.19	0.0743	HW70	-0.0074	0.0037	0.952
HW71	-0.1491	1.2483	0.264	H2Y	0.1192	0.292	0.589	HW71	-0.0531	0.2170	0.641
HW72	0.2499	3.4646	0.0627	H3Y	0.0841	0.329	0.5660	HW72	0.0512	0.1794	0.672
HW73	0.0462	0.1168	0.733	H4Y	-0.0339	0.043	0.8358	HW73	0.0108	0.0064	0.936
V206	0.1953	1.0421	0.307	H5Y	0.1198	0.518	0.4717	V207	0.0959	0.3536	0.552
V207	-0.1144	0.2972	0.586	H6Y	-0.2206	1.76	0.1841	V214	-0.0982	0.3273	0.567
V219	-0.1862	1.0546	0.304	H8Y	0.1929	0.581	0.4460	V219	-0.1198	0.5868	0.444
M1E	0.2448	3.8905	0.0486	ML1	-0.0004	2.96e-6	0.9986	V321	-0.059	0.1487	0.700
H3Y	0.0791	0.1105	0.740	Global	NA	11.4	0.4130	H5Y	0.0198	0.0136	0.907
H9Y	0.0106	0.0019	0.965					H7Y	0.0530	0.0989	0.753
Global	NA	43.900	3.19e-05					H9Y	-0.2158	0.7076	0.400
								Global	NA	42.52	0.0001

Log-rank residuals



Log-rankscore residuals



Bs.gradient residuals

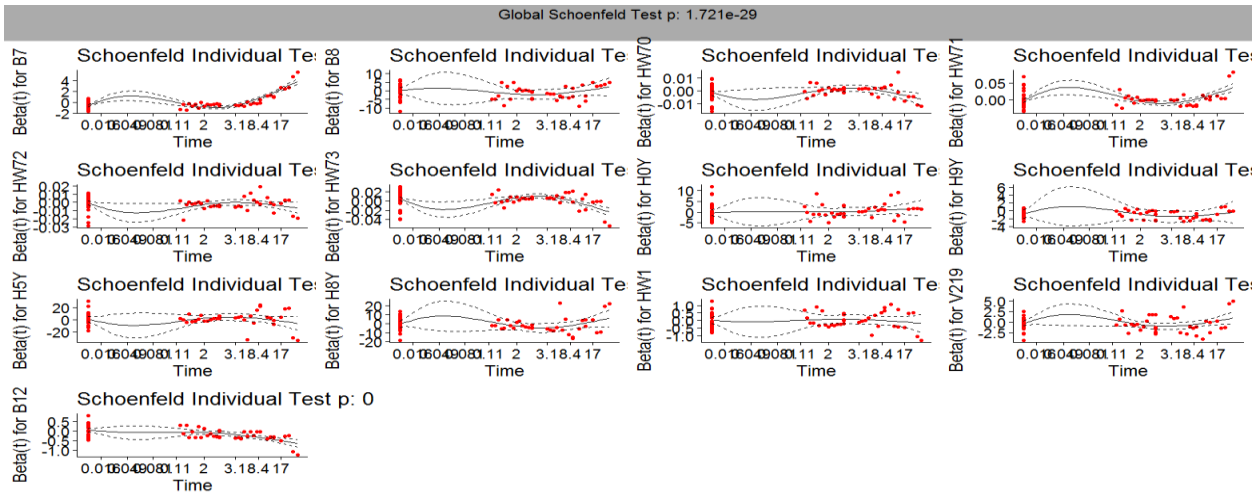


Figure 3. Schoenfeld residuals

Variables with importance value above 0.002 are considered predictive according to [3]. From our results in table 3, 15 variables were selected as predictive for under five child mortality based on log-rank splitting rule, 14 based on log-rank score model while Bs.gradient resulted to 18 predictive variables. It is also observed from table 3 that most of the extracted predictors using log-rank and Bs.gradient methods are similar. However, the extracted important variables and their importance measure are slightly different in the three models.

3.2. Results of Test for PH Assumption

PH assumptions that the relative risks are constant over time were tested by examining scaled Schoenfeld residuals as shown in figure 3 and use of statistical tests shown in table 4.

PH assumption is supported by a non significant (p-values >0.05) test of hypothesis result. According to the statistical table, there exist some variables which are statistically significant hence violating the PH assumption. These are variables B7, B12 and M1E in log-rank model and variables B7 and B12 in Bs.gradient model. The global test gives a general picture of PH violation among the variables in the model. P value less than 0.05 suggests one or more violations. Violation of PH assumption is therefore evident in log-rank and Bs.gradient models according to the global test making log-rank test and Cox PH model not optimal in this dataset.

When the PH assumption is violated, the Cox model can lead to biased results. We therefore used the Cox-Aalen's model which takes into account time-varying effects. The extracted important variables were fitted to a Cox-Aalen's model and results presented in table 5.

3.3. Results of Prediction Using Cox Aalen's Model

Table 5. Results of fitting the selected variables in Cox-Aalen's model

Log-rank model			
Variable	coefficient	Standard error	p-value
V206	1.8400	0.2990	3.47e-10
V207	1.6000	0.3020	3.81e-07
V219	-0.3400	0.1890	0.0578
B7	-0.1490	0.0884	0.0482
B8	-0.7300	0.2320	0.00221
B12	-0.0567	0.0393	0.0725
M1E	0.0145	0.0313	0.503
HW70	-0.000355	0.00159	0.771
HW71	0.001880	0.00171	0.180
HW72	-0.000293	0.00280	0.902
HW73	-0.001160	0.00146	0.350
Logrankscore model			
Variable	coefficient	Standard error	p-value
V218	-0.7100	0.2890	0.00401
B12	-0.0874	0.0252	2.37e-05
ML1	0.5330	0.1640	0.00821
HW70	-0.00126	0.00138	0.231
HW71	0.00134	0.00138	0.200
Bs.gradient model			
Variable	coefficient	Standard error	p-value
V207	0.6750	0.2820	0.0129
V214	-0.2250	0.1680	0.2840
V219	-0.5000	0.2360	0.0208
B7	-0.1170	0.0973	0.1410
B8	-0.4480	0.1340	0.0170
B12	-0.0519	0.0404	0.1150
HW70	0.000068	0.00152	0.954
HW71	0.001860	0.00154	0.129
HW72	-0.001100	0.00247	0.589
HW73	-0.000752	0.00142	0.535

From table 5, the covariates which turned to be statistically significant (have p value ≤ 0.05) are V206, V207, B7, and B8 from the log-rank model, V207, V219, and B8 from the Bs.gradient model and V218, B12 and ML1 in the log-rank score model.

To compare the different models, concordance index was used in order to determine the effect of the various splitting methods and results shown in table 6.

3.4. Results of Model Selection

Table 6. Concordance measure of model fit statistics

Description/Method	bs.gradient	Log-rank	Log-rank score
Sample size	68	68	68
Concordance	0.8641	0.916	0.7991
Standard error	0.02797	0.0196	0.0426
Discordant	1316	1395	1217
Concordant	207	128	306
Tied.x	0	0	0
Tied.y	158	158	158
Tied.xy	0	0	0

According to the results in table 6, the three models resulted in good fit with reference to the concordance index whereby all had concordance values above 0.79. The log-rank test resulted to the best performance with concordance of 0.92 followed by Bs.gradient with a concordance of 0.86. Log-rank score had the lowest concordance among the three models.

As indicated earlier, the optimality of log-rank is achieved when the model variables satisfy the PH assumptions. Since some variables were found to violate PH assumptions as in table 4, Bs.gradient splitting rule becomes the most suitable when PH assumption is violated.

4. Discussion

In this study, we analyzed the performance of data

balanced using random under-sampling method under different splitting rules. The study addressed the challenges of data balancing and violation of PH assumptions. The splitting rules used are the log-rank, log-rank score and BS.gradient splitting rules. The performance was based on 757 variables in prediction of Under Five Child Mortality from the 2014 KDHS data.

From our findings, the model that used log-rank splitting rule performed best with the largest concordance index of 0.916 and smallest error rate of 0.019. This was followed by the model with Bs.gradient splitting rule with a concordance index of 0.864 and error rate of 0.028. Log-rank score had the smallest concordance index of 0.799 and highest error rate of 0.043.

Other researchers have explored the issue of splitting rules in RSF. [31] Used four splitting rules (log-rank, log-rank approximation, log-rank score and conservation of events) in RSF where log-rank splitting rule performed best with the smallest error rate. Log-rank test has been used for survival splitting as a means of maximizing survival difference between nodes [9], [10], [11], [12], and [13] among others.

Despite its good performance, use of log-rank splitting rule is not appropriate when the PH assumption is violated. From the 2014 KDHS data, some of the variables were found to violate PH assumption making the use of log-rank method not appropriate.

A number of studies have explored on the different splitting rules in RSF. [5] Proposed an improved RSF by using weighted log-rank test in splitting the node while using the model of [6]. [7] Used R-squared splitting rule in survival forests while [8] compared RSF using different splitting rules. [32] Proposed Harrell's Concordance index (C index) split criterion in RSF. In their research, they did a comparison between C index splitting rule with the log-rank splitting rule where C index splitting rule was found to perform better than log-rank when censoring rate is high and in smaller scale clinical studies.

In this study, we have addressed the challenges of imbalance in the datasets and selection of the most optimal splitting method when PH assumption is violated. The challenge of high imbalance between mortality and non mortality class was addressed by working with data balanced using random under-sampling method before integrating the data with RSF. Studies that researched on use of balanced data in RSF includes [30] who developed a BRSF by integrating synthetic minority over-sampling technique in RSF and [4] who analyzed the performance of RSF using different balancing techniques (random under-sampling, random over-sampling, both-sampling and synthetic minority over-sampling technique), where random under-sampling emerged the best.

In this study, we considered a balanced dataset for maximum growth of the trees. The BRSF was then analyzed using log-rank, log-rank score and Bs. Gradient splitting rules when the PH assumptions are violated where Bs.gradient splitting rules was found to be the most suitable splitting rule.

5. Conclusions

In this paper, we have analyzed the performance of BRSF using different splitting rules. This addressed the challenges of imbalance which is commonly occurring in survival data and also selection of the most optimal splitting rule when PH assumption is violated. The results from the analysis presented in this paper show the superiority of log-rank splitting rule. However, optimality of log-rank is achieved when the hazard is proportional over time. It is clear from the survival curves by covariates shown in this paper as well as the statistical tests that most variables in the dataset used violate the PH assumption making log-rank not appropriate for node splitting. According to our analysis, we settle on Bs.gradient splitting method which still has a high concordance index of 0.86 and smaller error rate of 0.028. Use of balanced data and good choice of an optimal splitting rule that can separate high risk variables and low risk variables leads to further improvements in RSF. This brings about high levels of accuracy during prediction process. Using BRSF with Bs.gradient splitting rule, the identified determinants of U5CM are; V207 (sum of deceased daughters), V219 (sum total of living children) and B8 (age of the child). Hence, the age of the child and the siblings' information are identified as key determinants of U5CM.

Appendix

Table 7. Description of Important variables

Category	Variable	Description
Childs characteristics at birth	B7	Age at death of the child in completed months
	B8	Current age of the child in single years for all living children
	B12	Succeeding birth interval
	B16	Line number of the child in the household.
Height and Weight and Hemoglobin	HW70	Height for age standard deviation (according to WHO)
	HW71	Weight for age standard deviation (according to WHO)
	HW72	Weight for height standard deviations (according to WHO)
	HW73	BMI standard deviations (according to WHO)

Category	Variable	Description
Reproduction (siblings information)	V206	Total number of sons who have died
	V207	Total number of daughters who have died
	V214	Imputed duration of the current pregnancy.
	V218	Total number of living children
	V219	Total number of living children including current pregnancy
	V220	Total number of living children including current pregnancy
Contraceptive use	V321	Marital duration at sterilization
	V3A07	First source for current contraceptive method.
Marriage	V506	The rank of the respondent among the partner's wives.
Fertility Preferences	V626	Unmet need for family planning
Vaccination History	H2Y	BCG vaccination date - year
	H3Y	DPT-HEP.B-HIB 1 year
	H4Y	Oral Polio 1 year
	H5Y	DPT-HEP.B-HIB 2 years
	H6Y	Oral Polio 2 year
	H7Y	Pneumococcal 1 date
	H8Y	Oral Polio 3 year
	H9Y	Measles 1 year
maternity	M1E	Last tetanus injection before last pregnancy
Malaria	ML1	Times took SP/Fansidar during pregnancy

REFERENCES

[1] Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological), 34(2), 187-220.

[2] Kleinbaum, D. G. and Klein, M. (2005). Survival Analysis: A Self-Learning Text. Springer, New York.

[3] Ishwaran, H., Kogalurt, U. B., Blackstone, E. H., and Lauer, M.S. (2008). Random Survival Forests. The Annals of Applied Statistics, 2(3), 841-860.

[4] Waititu, H.W., Koskei, J. K., and Onyango, N. O. (2020). Determinants of Under Five Child Mortality from KDHS Data: A Balanced Random Survival Forests (BRSF) Technique, International Journal of Statistics and Applications, 10(5), 118-130.

[5] Miao, F., Cai, Y.P., Zhang, Y.X., Fan, X.M., and Li, Y. (2018). Predictive Modeling of Hospital Mortality for Patients with Heart Failure by Using an Improved Random Survival Forest. Department of Industrial and Systems Engineering, 6.

[6] Yang, S. and Prentice, R. (2005). Semi parametric analysis of short-term and long-term hazard ratios with two-sample survival data. Biometrika, 92, 1-17.

[7] Hong Wang, Xiaolin Chen, and Gang Li (2018) Survival Forests with R-Squared Splitting Rules. Journal Of Computational Biology 25(4), 388–395.

[8] Wanyonyi, K.S., Owuor,N. and Sarguta R. (2019). Comparison of Random survival forests split rules in selecting the determinants of under five mortality in Kenya using 2014 DHS data. Research report in Mathematics, Number 15.

[9] Ciampi, A., Hogg, S.A., McKinney, S., and Thiffault, J. (1988). RECPAM: A computer program for recursive partition and amalgamation for censored survival data. Comp.Methods Programs Biomed., 26 (3), 239–256.

[10] Segal, M.R. (1988). Regression trees for censored data. Biometrics, 35–47.

[11] Segal, M.R. (1995). Extending the elements of tree-structured regression. Stat. MethodsMed. Res., 4 (3), 219–236.

[12] LeBlanc, M. and Crowley, J. (1992) Relative risk trees for censored survival data. Biometrics, 411–425.

[13] LeBlanc, M. and Crowley, J. (1993) Survival trees by goodness of split. J. Am. Stat. Assoc., 88 (422), 457–467.

[14] Kenya National Bureau of Statistics, Ministry of Health [Kenya], National AID Control Council [Kenya], Kenya Medical Research Institute, National Council for Population and Development [Kenya], ICF International. Kenya demographic and health survey 2014. Nairobi, Kenya, 2015.

[15] Clark, T., Bradburn, M., Love, S., and Altman, D. (2003). Survival analysis part I: Basic concepts and first analyses. Br J Cancer., 89:232238.

[16] Bewick, V., Cheek, L., and Ball, J. (2004). Statistics review 12: survival analysis. Crit Care., 8:389394.

[17] Ishwaran, H. and Lu, M. (2019). Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. Stat Med. PMC.

[18] Ishwaran, H. and Kogalur, U. (2015). RandomForestSRC: Random Forests for Survival, Regression and Classification

(RF-SRC). https://cran.rproject.org/web/packages/randomForestSRC/randomForestSRC.

[19] Breiman, L. (2001b). Statistical modeling: The two cultures. Statistical Science, 6(199-231): 206.

[20] Lunetta, K. L., Hayward, L. B., Segal, J., and Eerdewegh, P. V. (2004). Screening large- scale association study data: Exploiting interactions using random forests. BMC Genetics, 5(32): 206.

[21] Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P., and Eerdewegh, P. V. (2005). Identifying snps predictive of phenotype using random forests. Genetic Epidemiology, 28(171-182): 206.

[22] Diaz-Uriarte, R. and Alvarez de Andres, S. (2006). Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7(3): 205, 206.

[23] Ishwaran, H., Blackstone, E. H., Hansen, C. A., and Rice, T. W. (2009). A novel approach to cancer staging: Application to esophageal cancer. Biostatistics, 10(206): 603-620.

[24] Weathers, W. and Cutler, R. (2017). Comparison of Survival Curves between Cox Proportional Hazards, Random Forests, and Conditional Inference Forests in Survival Analysis. All Graduate Plan B and other reports, 927. https://digitalcommons.usu.edu/gradreports/927.

[25] Ciampi, A., Chang, C.H., Hogg, S. McKinney, S. (1987). Recursive partition: A versatile method for exploratory-data analysis in biostatistics. In: Biostatistics. New York: Springer, 23–50.

[26] Hothorn, T., Lausen, B. On the exact distribution of maximally selected rankstatistics. (2003). Comput Stat Data Anal. 43(2): 121–37.

[27] Meinshausen, N. Quantile regression forests. (2006). The Journal of Machine Learning Research, 7: 983-999.

[28] Athey, S., Tibshirani, J., and Wager, S. (2006). Solving heterogeneous estimating equations with gradient forests. arXiv preprint arXiv.

[29] Scheike, T. and Zhang, M. (2002). An additive-multiplicative cox-aalen model. Scand. J. Statist, 28:75-88.

[30] Afrin, K., Illangovan, G., Srivatsa, S. S., and Bukkapatnam, S. T. S. (2018). Balanced random survival forests for extremely unbalanced, right censored data. Department of Industrial and Systems Engineering, 108: 246-257.

[31] Frank R. Datema, Ana Moya, Peter Krause , Thomas Bäck , Lars Willmes, Ton Langeveld , Robert J. Baatenburg de Jong, and Henk M. Blom. (2011). Novel head and neck cancer survival analysis approach: Random survival forests versus cox proportional hazards regression. Journal of the sciences and specialties of the head and neck, 34 (1): 50-58.

[32] Matthias Schmid, Marvin N. Wright and Andreas Ziegler. (2016). On the use of Harrell's C for clinical risk prediction via random survival forests. Expert Systems with Applications. 63: 450-459.