Transactions on Engineering and Computing Sciences - Vol. 12, No. 3
Publication Date: June 25, 2024
DOI:10.14738/tecs.123.17112.

Ramkumar, T. B., & Oscar, L. K. (2024). Predictive Accuracy and Covariates Selection Performance of Time-To-Event Data: A Comparison Study of Cox Proportional Hazard and Random Survival Forest Models. Transactions on Engineering and Computing Sciences, 12(3). 104-123.

SERVICES FOR SCIENCE AND EDUCATION

Predictive Accuracy and Covariates Selection Performance of Time-To-Event Data: A Comparison Study of Cox Proportional Hazard and Random Survival Forest Models

Ramkumar T Balan
University of Dodoma, Dodoma, Tanzania

Oscar Leonard. K
University of Dodoma, Dodoma, Tanzania

ABSTRACT

Background: The Cox proportional hazard (Cox PH) model is the most common and widely used method for analyzing time-to-event data. This approach, however, has been criticized for predicting patient outcomes since it assumes that the outcome is a linear combination of covariates and a constant hazard ratio over time, which is invalid in many cases and may not always be fulfilled, and thus Random Survival Forest (RSF) for time-to-event data has been proposed as an alternative to Cox PH. By considering non-linearity and interrelated covariates, the RSF is recognized as a way to correct bias in Cox proportional hazard models. Methods: The Cox PH model was compared to the Random Survival Forest by a relaxed proportional hazard (PH) assumption in this retrospective cohort study, which included 189 patients who underwent a new algorithmic treatment of an advanced human immunodeficiency virus clinical trial proposed by the World Health Organization from January to December 2018 at Kibong'oto infectious diseases hospital. Results: The study results show that the Random Survival Forest outperforms the Cox PH model in time-to-event data analysis by integrating covariates that do not meet proportional hazard assumptions based on predictive error curves, concordance indices, Brier scores, and integrated Brier scores as predictive indicators. Despite the fact that the PH assumption was relaxed for all covariates in the dataset, the Cox PH model selected only a few, and all deselected covariates were highly significant in the RSF, implying that the Cox PH model was inbuilt with the proportional hazard assumption. Conclusions: As a result, only covariates that satisfied the PH inbuilt assumption were chosen, regardless of whether the assumption was tested or not, rendering it ineffective when compared to a random survival forest. Therefore, having a dataset free from non-linearity and interrelated variables is impossible in the real world; hence, the RSF model outperforms the Cox PH model in predictive accuracy and covariate selection performance.

Keywords: Predictive accuracy, Random survival forest, Covariates significance, Cox proportional hazard model, Survival analysis, time-to-event.

Abbreviations

ART, antiretroviral therapy; BS, brier score; CD4, cluster of differentiation 4; CI, confidence interval; C-index, concordance index; CTC, care and treatment clinical; HIV, human immunodeficiency virus; IBS, integrated brier score; LF-LAM, lateral flow-lipoarabinomannan;

MTB mycobacterium tuberculosis; PH, proportional hazard; PLWH, people living with HIV; RSF, random survival forest; TB, tuberculosis; VIMP, variable important; WHO, world health organization

INTRODUCTION

Predicting clinical outcomes and treatment responses in clinical trials is still difficult because of the complexity of the disease, the difference in the lengths of patient follow-up across trials, and the variability of the patients, hence predicting survival time remains a challenge [1,2]. Sun et al [3] make it clear that reliable survival prediction models that can learn complex structures and interactions between explanatory variables are urgently needed. Survival time is a clinical outcome necessary in clinical research to determine a proposed therapy's effectiveness and safety in humans. To do this, researchers need to do clinical studies, collect longitudinal data from participants, and predict people's lives [4]. In medicine, knowing how long patients will live can help design treatment regimens and make decisions about how to share their regimens [5]. In human illness, accurate survival predictions are vital for clinical care and precision medicine decision. A survival model predicts how likely it is that an interesting event will happen in the future based on clinical data, genetic information, and medical imaging [3]. The Cox proportional hazard (PH) model is one of the most commonly used survival models for right-censored time-to-event data [6] and is used to assess the influence of variables on patient survival. The model provides for evaluating the association between a patient's survival time and one or more explanatory variables, as well as estimating the risk of an event happening to a person [7]. According to Austin [8], the proportional hazard assumption states that the effects of a variable on the hazard function stay the same over time. To make sure that your estimates and conclusions are accurate, you first need to figure out if the proportional hazard assumption is attained [9]. However, the model makes assumptions that are difficult to satisfy [6]. On the contrary, the Cox PH model ignores significant issues like missing predictors and nonlinear exponential factors. It also includes assumptions about how observations are related that are not always completely explained [10]. It is shown that till 2014 only 64 published papers verify the proportional hazard assumption among more than 25,000 citations since the first Cox original paper was published [11]. However, this limits the ability to select and predict variables [12]. As a result, the RSF model was proposed as an alternative approach to the Cox PH model in the presence of a non-proportional hazard, nonlinear effects, and complex interactions among multiple variables [6,12]. The RSF model extends the random forest (RF) method for obtaining the right censored data. In non-proportional hazards, RSF can easily leverage high-dimensional variable data [6]. Regardless of the advancement of survival forest, there have only been a few studies conducted to compare the RSF model with the Cox PH model. Du et al.[13] did a study on a real cancer dataset that compared the performance forecasting of RSF in referencing Cox PH without taking into account the nonlinearity and high-dimensional covariate. Miao et al. [14] conducted a study on intensive care unit patients with heart failure to compare variable selection performance between RSF and improved RSF (iRSF). Dietrich et al.[15] studied the type 2 diabetes mellitus dataset by comparing performance between RSF and Cox PH with highly correlated complex survival data. Kim et al [16] compared the survival prediction performance of the oral squamous carcinoma cell dataset between the Deep learning-based-survival model, the RSF, and the Cox PH model using Harrell's c-index. They found that DeepSurv was the best of the three models, followed by RSF and the Cox PH model. Liu et al [6] applied a simulation and

used a real data to compare prediction and variable selection performance among RSF, Conditional inference forest (CIF), and Survival forests with maximally selected rank statistics (MSR-RF) and concluded that CIF and MSR-RF perform better when there are multiple categories of variables, CIF is best when dealing with interaction, and MRS-RF performs better when there are only continuous variables. A simulation study done by Nasejje et al [17] to compare predictive performance between Conditional inference forest (CIF) and RSF concluded that all three-survival forest models have good predictive performance based on the Brier score. With Kenya Demographic Health Survey (KDHS) data, Waititu et al [10] analyzed the performance of a balanced Random Survival Forest (BRSF) using different splitting rules and balanced data about RSF and the Cox PH model and concluded that the log-rank rule is the best splitting rule. Not only that but also Nasejje and Mwambi [18] employed Uganda Demographic Health Survey (UDHS) data on under-five child mortality for 2011 to compare predictive performance between RSF and the Cox PH model in non-PH assumption covariates and PH assumption and concluded that RSF performs better in non-PH assumption and Cox PH perform better in PH assumption. Furthermore, the reviewed studies, which were making the comparison between the Cox PH and RSF, were based on two sides; on one side when the PH assumption is fulfilled, the Cox PH outperformed the RSF and the second side was when the PH assumption is violated the RSF outperformed the Cox PH and failed to come up with one consensus. In the real world, it is inevitable to have datasets that are free of covariates that violate the PH assumption or that contain all covariates that are linear in nature and independent of each other, hence satisfying the PH assumption. In accordance with this, the study examined advanced HIV datasets and compared the Cox PH as a fixed survival model and RSF as a random survival model, taking into account that the PH assumption is invalid in many cases even when the relationship exists.

MATERIAL AND METHODS

In this part, we began by exploring the widely known Cox proportional hazard model before moving on to an alternate approach to modeling time-to-event data: random survival forest. The Cox PH model was compared to the Random Survival Forest by relaxed proportional hazard (PH) assumption in this retrospective cohort study, which included 189 patients who underwent a new algorithm treatment of advanced human immunodeficiency virus clinical trial proposed by the World health organization (WHO) from January to December 2018 at Kibong'oto infectious diseases hospital.

Cox Proportional Hazard Model

The Cox proportional hazards (Cox PH) model is one of the approaches, which is commonly used for determining the combined effects of various covariates on the relative risk (hazard) in survival models. The risks of the several strata generated by the levels of the variables are assumed to be proportionate in this model [19]. In this study, the Cox PH was fitted based on the maximum likelihood estimate (MLE). According to Binder [20], the pseudo-maximum likelihood estimator is obtained by using a maximum likelihood estimate. Let t_i be the failure time for subject i, $i = 1,2, \dots, n$. For a sample size N, the partial likelihood function is given by;

$$L(\beta) = \prod_{i=1}^{N} \left[\frac{h\{t_i | x_i(t_i)\}}{\sum Y_j(t_i) h\{t_i | x_j(t_i)\}} \right]^{\delta_i},$$

where the sum is over the range $j = 1,2, \ldots, N$, and $\delta_i = 1$ if the i^{th} unit is an observed failure, and $\delta_i = 0$ if the i^{th} unit is censored,

$$Y_j(t_i) = \begin{cases} 1\ (t \le t_j), \\ 0\ (t > t_j). \end{cases}$$

To maximize the partial likelihood function we determine β so that

$$\sum_{i=1}^{N} \delta_i \left\{ x_i(t_i) - \frac{S^{(1)}(t_i,\beta)}{S^{(0)}(t_i,\beta)} \right\} = 0, \quad (1)$$

where

$$S^{(0)}(t,\beta) = \frac{1}{N}\sum_{i=1}^{N} Y_i(t) exp\{x_i(t)'\beta\}, \quad (2)$$

$$S^{(1)}(t,\beta) = \frac{1}{N}\sum_{i=1}^{N} Y_i(t) x_i(t) exp\{x_i(t)'\beta\}. \quad (3)$$

Equation 1 is a solution that specifies the finite population parameter of interest β which is determined by all N of the $Y_i(t)$'s and δ_i's, which we want to estimate from an n-sample. The estimating technique substitutes weighted sums for Equation (1, 2, and 3) by using the sample weighted w_i scale so that $\sum w_i = 1$. The weights are chosen in such a way that the weighted sums provide roughly unbiased and consistent estimates of the respective means throughout the limited population. The estimator of $\hat{\beta}$, in specifically, is the estimating equation for β[20].

$$\sum_{i=1}^{n} w_i \delta_i \left\{ x_i(t_i) - \frac{\hat{S}^{(1)}(t_i,\hat{\beta})}{\hat{S}^{(0)}(t_i,\hat{\beta})} \right\} = 0, \quad (4)$$

Where

$$\hat{S}^{(0)}\left(t,\hat{\beta}\right) = \sum_{i=1}^{n} w_i Y_i(t) exp\left\{x_i(t)'\hat{\beta}\right\} \quad (5)$$

$$\hat{S}^{(1)}\left(t,\hat{\beta}\right) = \sum_{i=1}^{N} w_i Y_i(t) x_i(t) exp\left\{x_i(t)'\hat{\beta}\right\}. \quad (6)$$

From Equation (4) above the $\hat{\beta}$ estimator is the solution and it is called the pseudo-maximum likelihood estimator.

Random Survival Forest Model

RSF is a collective method that introduces two types of randomization into the tree-building procedure: bootstrap sampling from the original dataset and the selection of a small number of independent covariates [21]. To compute the RSF model, many random bootstrap samples are drawn [15]. The first step in creating an RSF was to obtain a new dataset of size 189 from an advanced HIV dataset by sampling without replacement [22]. After continuously resampling from the advanced HIV dataset, the RSF outcome was averaged over numerous survival trees. Each

bootstrap sample included an average of 63% of the advanced HIV dataset, with the remaining 37% designated as an out-of-bag (OOB) sample. To develop a decision tree, a random set of candidate explanatory was picked at each node for random node splitting. The square root of the whole number of variables denoted by p correlated to the number of candidate variables denoted by m. This means $mtry = \sqrt{p}$, for this study, 26 variables were used to predict the time to death as an interesting event. Thus, 5 of candidate variables were selected at each split. The above explanation was attained by following the RSF algorithm as outlined by Rono [23]:

1. The method is started by randomly choosing 1000 bootstrap samples from the original dataset, called the Advanced HIV dataset. The bootstrap sample size is 63% of the Advanced HIV dataset, and the remaining 37% is termed the out-of-bag (OOB) dataset.
2. For every bootstrap sample chosen, two full-size survival trees are grown. Randomly chosen $mtry$ candidate predictors from all p predictors at each node. The node is split using candidate predictors that optimize daughter node survival. Log-rank random is the default-splitting rule, and $mtry = \sqrt{p}$ is the candidate variable size.
3. The survival trees are built until the terminal node has $d_0 > 0$ unique deaths.
4. There is a calculation of each tree's cumulative hazard function (CHF). The Nelson-Aalen estimator is estimated the CHF for each terminal node.

$$\widehat{H}_h(t) = \sum_{t_{i,h} \le t} \frac{d_{i,h}}{Y_{i,h}} \tag{7}$$

Where $d_{i,h}$ represents deaths number and $Y_{i,h}$ is individual numbers at risk at time $t_{i,h}$. The CHF for all participants within similar node is equal. There is the calculation of the mean of CHFs for all B trees to obtain the ensemble CHF of the forest.

$$\widehat{H}_E(t|x) = \frac{1}{B}\sum_{i=1}^{B} \widehat{H}_i(t|x) \tag{8}$$

where $\widehat{H}_i(t|x)$ represent the CHF of the tree that is grown in the i^{th} bootstrap sample.

5. By using the OOB dataset, the study is to compute the predictor's error for the ensemble CHF.

In this study, the log-rank random statistic is employed as a splitting rule for node splitting [15]. The log-rank random statistics is the best among the splitting rules as described in the study conducted by Waititu et al [10]. By using the formula below, the log-rank statistic was able to maximize the survival variances between nodes [15].

$$L(x,c) = \frac{\sum_{i=1}^{N}\left(d_{i,1} - Y_{i,1}\frac{d_i}{Y_i}\right)}{\sqrt{\sum_{i=1}^{N}\frac{Y_{i,1}}{Y_i}\left(1-\frac{Y_{i,1}}{Y_i}\right)\left(\frac{Y_i - d_i}{Y_i - 1}\right)d_i}} \tag{9}$$

where d_i is the death number, Y_i is participants at risk at time t and i is the number of participants.

Several random split points were selected at random to compute the log-rank statistics for each candidate variable. This study pre-defined the number of random split points per candidate variable before the computation of an RSF model when using the package random Forest SRC in the R statistical software.

Selecting Significant Covariates: Comparison of the Cox PH Model to Random Survival Forest

In determining the covariates selection performance in both the Cox proportional hazard model and the random survival forest models, the study employed the stepwise and variable important (VIMP) techniques in both Cox PH model and random survival forest model respectively.

Predictive Accuracy: Comparison of the Cox PH Model to Random Survival Forest

This study employed predictive error curves, C-index, BS, and IBS to assess the model's prognostic accuracy. This choice was preferred because these are common and standard ways of checking the accuracy of prediction comparison among the survival models [16–18,24].

Concordance Index (C-index)

Under C-index. The study let $(T_{1,h}, \delta_{1,h}), (T_{2,h}, \delta_{2,h}), \dots, (T_{n,h}, \delta_{n,h})$ be the survival times and the censoring status for n participants in a terminal node h. It also let $t_{1,h} < t_{2,h} < \cdots < t_{m,h}$ be the m different event periods in terminal node h. The study defined $d_{i,h}$ and $Y_{i,h}$ as the death number and participants at risk at time $t_{i,h}$. The Nelson-Aalen estimator for CHF estimate for terminal node h was given in equation (7), hence headed the equation below.

$$H(t|x_i) = \widehat{H}_h(t), \text{If } x_i \in h. \tag{10}$$

In the RSF process, to approximate the CHF of participant i it was described that $I_{i,b} = 1$ if i is a $i, b = 1$, of OOB case for b^{th} bootstrap sample $I_{i,b} = 0$. Letting $H_b(t|x_i)$ represent the CHF for subject i in a tree mature from the b^{th} bootstrap sample, otherwise equal to zero. The collective CHF for i is

$$H^*(t|x_i) = \frac{\sum_{b=1}^{B} I_{i,b} H_b(t|x_i)}{\sum_{b=1}^{B} I_{i,b}} \tag{11}$$

The C-index was calculated using the steps below as given by Yosefian et al [21]

1. All subject-pairs are created in the study
2. Elimination of the permissible pairs whose lesser survival time is censored and pairs (i, j) if $T_i = T_j$, and both were deaths.
3. For each permissible pair where $T_i = T_j$, is counted one, if the shorter survival time has a high risk predicted; and is counted 0.5 if the risk predicted is tied. For each permissible pair, where $T_i = T_j = 1$, and both are deaths, are counted one if the risk predicted is tied; otherwise, is counted 0.5. For each permissible pair where $T_i = T_j$, but at least one is not dead, is counted one if the death had a high risk predicted; otherwise, is counted 0.5. Let concordance denote the sum over all permissible pairs.

4. C-index = Concordance / permissible.

In the survival tree, it is said that, i has a high risk predicted than j if

$$\sum_{i=1}^{m} H(t_i|x_i) > \sum_{i=1}^{m} H\left(t_i\middle|x_j\right) \quad (12)$$

Where $t_1 < t_2 < \cdots < t_m$ are the exceptional event periods in the dataset, a value of 0.5 for C-index is not superior to random guessing and a value of one represents full discriminative capacity. Percentiles 2.5 and 97.5 reflected as lower and upper bounds of CI for final statistics. The method was common and standard in comparing the performance of survival models by calculating the likelihood of concordance between estimated and observed survival. The closer the c-index to 1 indicated the better the fitted model, and the value of 1 indicates the perfect prognostic of the fitted model and a value of 0.5 showed that the fitted model did no better than guessing.

Brier Score (BS)

The Brier score at time t is represented by the following mathematical equation [25]

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\left(0-\hat{s}(t|x_i)\right)^2}{\hat{G}(t_i)} I(T_i \le t, \delta_i = 1) + \frac{\left(1-\hat{s}(t|x_i)\right)^2}{\hat{G}(t)} I(T_i > t)\right\} \quad (13)$$

where $\hat{G}(t) = P(C_i > t)$ represents the Kaplan-Meier, approximation of the censoring survival function. The prediction error curve is obtained by manipulating the Brier score across the times. The Brier score method depends on the selection of a single time point. The smaller the BS, the better the survival model prediction accuracy, and in the rare scenario when the Brier score is equal to zero it indicates the perfect prediction accuracy of the model.

Integrated Brier Score (IBS)

The integrated Brier score (IBS) is the weighted brier score. The IBS comes on hand to suppress the weakness of the BS that is limited to a single time point. The Integrated Brier score provides an overall measure of prediction error by averaging the Brier score over time by considering inverse probability censoring weighting (IPCW). The smaller the value of the IBS, the better the model prediction, that is cumulative prediction error curves over time as specified by Yosefian et al [21] as follows:

$$IBS = \frac{1}{\max(t_i)} \int_0^{\max(t_i)} BS(t)dt. \quad (14)$$

Percentile 2.5 and 97.5 were measured as lower and upper bounds of CI for final statistics.

RESULTS

In this section, the study presented the findings of the fitted the Cox proportional hazard and random survival forest models to time-to-event data. The predictive accuracy and covariates selection performance of the two modeling approaches were compared using the various techniques as discussed above.

Demographic Characteristics of the Advanced HIV Dataset

The demographic characteristics describe the view of patients' distribution to the outcome status, and the p-values of the chi-square and Fisher's exact tests describe the dependence of the variables on the survival outcome status. The response was time to death after being enrolled in the study. Amongst the patients, 12.7% [95% CI: 8.5-18] (24 out of 189 patients) died within 30 days of follow-up as described in Table 1.

The number of survival days ranged from 1 to 30 days. The patients' median and mean survival time follow-up was 30 and 27.2 days, respectively, after enrollment. Out of the 189 patients who participated, 107 were females, 56.6% [95% CI: 49-64], and the rest were males. The patients' age ranges from 18 to 76 years. The median and mean ages were 42 (95% CI: 40-44) and 42.14 (12.17) years, respectively. Most of the patients were aided (88.9%). Ninety-one percent of the patients were hospitalized [95% CI: 86-95] and 90.5% of the patients had more than one danger sign [95% CI: 85-94]. More than three-quarters of the patients experienced classical TB symptoms 96.8% [95% CI: 93-99]. Again, more than three-quarters of 99.5% [95% CI: 97-100] of patients received prior tuberculosis medications during the study. Before enrolling in the study, 95.2% [95% CI: 91-98] had completed the tuberculosis treatment. Most of the patients (94.2%) had absolute CD4 accounts of less than 200 $cells/mm^3$. More than three-quarters 76.2% [95% CI: 69-82] of the patients were in Care and Treatment Clinic (CTC). And 96.8% of the patients had less than 80 Karnofsky scores.

Table 1: The Demographic Characteristics Distribution of Patients' Survivorship outcomes

		Treatment Outcome Status		
Variable	**n**	**Survived, n = 165**	**Died, n = 24**	**p-value**
Participant's Age, Mean (SD)	189	43.2 (12.1)	34.7 (9.6)	<0.001
Gender, n (%)	189			0.11
Female		97 (58.8)	10 (41.7)	
Male		68 (41.2)	14 (58.3)	
Education status, n (%)	189			0.91
no formal education		15 (9.1)	2 (8.3)	
Primary school		123 (74.5)	17 (70.8)	
Secondary school		22 (13.3)	4 (16.7)	
Tertiary		5 (3.0)	1 (4.2)	
Occupation status, n (%)	189			0.25
employed		13 (7.9)	4 (16.7)	
Self-employed		77 (46.7)	12 (50.0)	
Not employed		75 (45.5)	8 (33.3)	
Marital status, n (%)	189			0.59
Married		82 (49.7)	13 (54.2)	
Single		40 (24.2)	5 (20.8)	
Cohabiting		5 (3.0)	2 (8.3)	
Divorced		18 (10.9)	1 (4.2)	
Widowed		20 (12.1)	3 (12.5)	
A patient care, n (%)	189			0.14

Outpatient care		17 (10.3)	0 (0.0)	
Hospitalized care		148 (89.7)	24 (100.0)	
Patient Ability to walk, n (%)	189			0.081
Unaided		21 (12.7)	0 (0.0)	
Aided		144 (87.3)	24 (100.0)	
Danger sings more than one, n (%)	189	148 (89.7)	23 (95.8)	0.48
Classical TB symptoms, n (%)	189	160 (97.0)	23 (95.8)	0.56
Cough, n (%)	189	154 (93.3)	22 (91.7)	0.67
Hemoptysis, n (%)	189	5 (3.0)	1 (4.2)	0.56
Weight loss, n (%)	189	131 (79.4)	18 (75.0)	0.62
Night sweat, n (%)	189	25 (15.2)	6 (25.0)	0.24
Chest pain, n (%)	189	146 (88.5)	19 (79.2)	0.20
Prior to TB diagnosis, n (%)	189			0.088
No		58 (35.2)	14 (58.3)	
Yes		18 (10.9)	1 (4.2)	
Unknown		89 (53.9)	9 (37.5)	
Prior TB Medication, n (%)	189	164 (99.4)	24 (100.0)	>0.99
Prior TB results, n (%)	189			0.69
Cure		7 (4.2)	0 (0.0)	
Completed		156 (94.5)	24 (100.0)	
Default		1 (0.6)	0 (0.0)	
Loss to follow up		1 (0.6)	0 (0.0)	
Patient on ART, n (%)	189			0.86
No		28 (17.0)	3 (12.5)	
Yes		125 (75.8)	19 (79.2)	
Unknown		12 (7.3)	2 (8.3)	
Rifampicin results, n (%)	189			0.20
MTB not detected		113 (68.5)	15 (62.5)	
MTB detected		42 (25.5)	5 (20.8)	
Rif resistance detected		2 (1.2)	0 (0.0)	
No results		8 (4.8)	4 (16.7)	
LF-LAM results, n (%)	189			0.26
Negative		109 (66.1)	13 (54.2)	
Positive		56 (33.9)	11 (45.8)	
Fever, n (%)	189			<0.001
< 39		157 (95.2)	16 (66.7)	
≥ 39		8 (4.8)	8 (33.3)	
Respiratory rate, n (%)	189			<0.001
< 30		159 (96.4)	15 (62.5)	
≥ 30		6 (3.6)	9 (37.5)	
Pulse rate, n (%)	189			0.70
< 120		150 (90.9)	23 (95.8)	
≥ 120		15 (9.1)	1 (4.2)	
Absolute CD4 Counts, n (%)	189			0.094
< 50		47 (28.5)	10 (41.7)	

50-99		56 (33.9)	8 (33.3)	
100-149		45 (27.3)	3 (12.5)	
150-199		6 (3.6)	3 (12.5)	
≥ 200		11 (6.7)	0 (0.0)	
Hemoglobin, n (%)	189			0.008
< 11		63 (38.2)	16 (66.7)	
≥ 11		102 (61.8)	8 (33.3)	
Karnofsky Scores, n (%)	189			0.48
< 50		90 (54.5)	16 (66.7)	
50-70		69 (41.8)	8 (33.3)	
≥ 80		6 (3.6)	0 (0.0)	
[1]Mean (SD) or Frequency (%);[2]Pearson's Chi-squared test; Fisher's exact test				
The dataset originated from Kibong'oto Infectious Diseases Hospital				

Moreover, based on occupation, half of the death 12 (50%) were self-employed, followed by 8 (33.3%) were not employed, and only 4 (16.7%) were employed. Again, centered on marital status, more than half (54%) of the deaths were married patients, followed by 5 (20.8%) of the single patients. Additionally, in terms of patient care, there were only 17 (10.3%) patients who were outpatients and no one died among those patients as they improved in the absolute CD4 account, and they were not aided at all, so they could manage themselves. For hospitalized care, 24 (100%) of the patients died because their situations were critical, so they needed intensive care and management, still likely to die. Nineteen (79.2%) patients who were on ART were more likely to die compared to those 3 (12.5%) who were not in ART. In addition, more than half of 15 (62.5%) of the patients who died, MTB were not detected compared to those 5 (20.8%) who were detected earlier and started treatment. Again, more than half of 14 (58.3%) of the patients who died had no prior TB diagnosis. Furthermore, 10 (41.7%) of the dead patients had a CD4 count of less than 50, and as the number of CD4 increased, the frequency of deaths decreased. As a result, 8 (33.3%) of patients who died had a CD4 count of 50-99, 3 (12.5%) of patients who died had a CD4 count of 100-149 or 150-199, and no one died if their CD4 count was equal to or more than 200. Similarly, more than half of the 16 (66.7%) patients died when their Karnofsky score was less than 50, and mortality decreased as the Karnofsky score increased. Thus, 8 (33.3%) of the patients who died had a Karnofsky score between 50 and 70, but those with values equal to or more than 80 did not die.

Survival Characteristics of Advanced HIV Dataset

Generally, the Kaplan-Meier (KM) method was employed by the study to estimate the overall trend of the survival function of the advanced HIV dataset. Figure 1 shows the KM curve of the general characteristics of the advanced HIV dataset based on the probability of an event at a certain time interval. The results showed that there was a wider interval from some point of event to another at the 95% confidence interval, which implies there were few participants. Similarly, it was portrayed the sex survival characteristics, in which male were at high risk compared to the female patients.

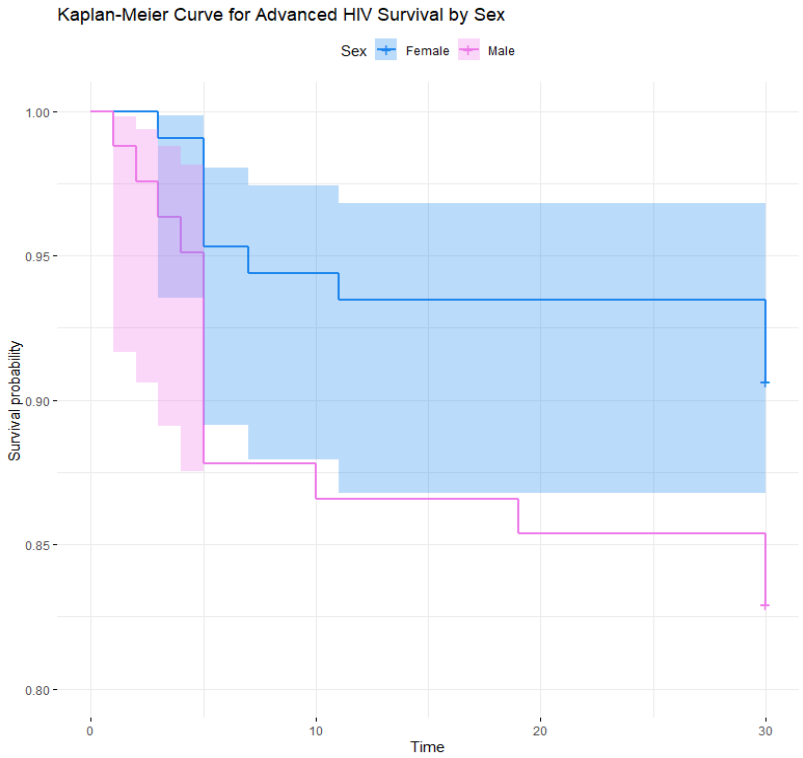


Figure 1: General survival Characteristics of the advanced HIV Patients based on gender

Significance of Selected Covariates in the Cox Proportional Hazard Model

The study performed eighteen iterations by using a stepwise mechanism; which means that both forward and backward selection is made at the same time in the Cox proportional hazard model to classify which predictors are extremely associated with the prognostic of death. Therefore, based on the results from the Cox PH model as in Table 2, the covariate with p values less than 0.05, which are marked with stars, was considered significant variables in predicting the interesting event of the study. The selected covariates have a lower standard error rate and no multicollinearity for all variables as for continuous variables the Variance Inflation Factor (VIF) should not exceed 10 and for categorical variables the VIF should not exceed 2.5 [Table 2]. Therefore, participants' age, Patient care, Classical TB symptoms, Prior TB diagnosis, and TB episode number were highly significantly associated to advance HIV patients' death rate. Other predictors such as Absolute CD4 count, Respiratory rate, Patient ability to walk, Gender, pulse rate, and anti-biotic on day one showed as important in explaining the death rate in advanced HIV patients. Unlikely covariates were Education status, Cough, Hemoptysis, Weight loss, Night sweat, Chest pain, Prior TB medication, Patient on ART, Hemoglobin, Karnofsky score, Xpert-MTB results, LF-LAM results, Marital status, Occupation status, and fever.

Table 2: Significance of selected covariates in the Cox proportional hazard model

Variable	HR12	SE2	95% CI2	p-value	VIF2
Participant's Age	0.85	0.04	0.79, 0.92	<0.001***	1.6
Gender	3.78	0.562	1.26, 11.4	0.005**	1.8
Patient care	2.85	4.04	0.00, 8.25	<0.001***	1
Patient ability to walk?	2.18	5.07	0.00, 8.45	0.002**	1

Pulse rate	0.92	0.014	0.90, 0.95	0.040*	1.7
Classical TB symptoms	0	2.36	0.00, 0.47	<0.001***	2.3
Cough	2.29	0.85	0.74, 1.00	0.063	1.4
Chest pain	1.73	0.573	0.56, 5.30	0.8	2.1
Prior TB diagnosis	0.3	0.416	0.13, 0.69	0.001***	2.5
Respiratory rate	0.7	0.042	0.64, 0.76	0.003**	1.5
TB episode Number	951	2.38	8.98, 10.06	0.001***	2.2
Fever	0.38	0.251	0.23, 0.62	0.066	1.6
Hemoptysis	3.8	1.07	0.47, 30.9	0.5	2.2
Patient on ART?	6.3	0.556	2.12, 18.7	0.2	1.9
Absolute CD4 Counts	0.99	0.003	0.99, 1.00	0.008**	1.4
Hemoglobin	0.96	0.077	0.83, 1.12	0.8	1.9
Weight loss	3.65	0.529	1.29, 10.3	0.3	1.2
Xpert-MTB	0.69	0.544	0.24, 2.01	0.8	2.2
LF-LAM result	0.19	0.465	0.08, 0.47	0.6	2.4
Antibiotic day one	5.91	1.1	4.91, 7.06	0.028*	1.3
Marital status	0.12	0.549	0.04, 0.36	0.075	2.2
Night sweat	0.47	0.534	0.17, 1.34	0.5	2
Education status	0.72	0.501	0.27, 1.92	>0.9	2
Occupation status	0.22	0.46	0.09, 0.54	0.5	1.9
Karnofsky scores	0.94	0.029	0.89, 1.00	0.054	2.3
1*p<0.05; **p<0.01; ***p<0.001, AIC = 219.86, Concordance= 0.87 (se = 0.029), 2HR = Hazard Ratio, SE2 = Standard Error, CI2 = Confidence Interval, VIF = Variance Inflation Factor, Likelihood ratio test= 48.69 on 10 df, p=0.0000005, Wald test = 22.8 on 10 df, p=0.01, Score (logrank) test = 31.82 on 10 df, p=0.0004 The dataset originated from Kibong'oto Infectious Diseases Hospital					

Significance of Selected Covariates in the Random Survival Forest Model

Table 3 shows the first 18 topmost ranked positive VIMP values greater than zero for advanced HIV patients' mortality risk covariates from the RSF fitted model. According to Ishwaran et al. (2008) and Wanyonyi (2019), covariates with VIMP values > 0.002 as a mean for thresh-holding predictors are important in prognosticating the outcome of the interesting event. Hence, the study concluded that patients' age, Karnofsky score, Xpert-MTB/ Rifampicin, Hemoglobin, Pulse rate, Absolute CD4 count, Patient ability to walk and Patient care were highly predictive covariates. Likewise, Gender, Respiratory rate, Fever, Education status, Ant biotic on Day one, Classical TB symptoms, Hemoptysis were moderately predictive and TB episode number, Night sweat, Occupation status, Prior TB medication were slightly predictive. However, LF-LAM results, Patient on ART, Cough, Chest pain, Weight loss, Prior TB diagnosis and marital status were unlikely to be predictive.

Table 3: Significance of selected covariates in the random survival forest model

Variable	VIMP	Significant
Participant's Age	0.0796	**TRUE**
Karnofsky scores	0.0626	**TRUE**
Xpert-MTB/Rifampicin	0.0205	**TRUE**
Hemoglobin	0.0168	**TRUE**

Pulse rate	0.0155	**TRUE**
Absolute CD4 Counts	0.0102	**TRUE**
Patient ability to walk?	0.0069	**TRUE**
Patient care	0.0051	**TRUE**
Gender	0.0044	**TRUE**
Respiratory rate	0.0039	**TRUE**
Fever	0.003	**TRUE**
Education Status	0.0024	**TRUE**
Ant biotic on Day one	0.0023	**TRUE**
Classical TB symptoms	0.0019	**TRUE**
Hemoptysis	0.0019	**TRUE**
TB episode Number	0.0012	**TRUE**
Night sweat	0.0005	**TRUE**
Occupation status	0.0004	**TRUE**
Prior TB Medication	0	FALSE
LF-LAM results	-0.0004	FALSE
Patient on ART?	-0.0016	FALSE
Cough	-0.0018	FALSE
Chest pain	-0.003	FALSE
Weight loss	-0.0031	FALSE
Prior TB diagnosis	-0.0037	FALSE
Marital status	-0.0101	FALSE

The dataset originated from Kibong'oto Infectious Diseases Hospital

Predictive Accuracy Based on the Error Curves

The key criteria for comparing the performance of the Cox PH and RSF models is to check the estimated error fluctuation as time goes on. The study plotted the estimated error fluctuation curve for both Cox PH, RSF, and KM that was used as reference to the both models as portrayed in Figure 2. The results show that at time zero, both models have the same estimated predictive error, and that at 3-time points, the predictive error rate began to increase dramatically up to time point 5, and that at time points 6 through 29, the models differed in error rate increase, with the Cox being higher than the KM and RSF, and that at 30-time points, the trend began to decrease for both models.

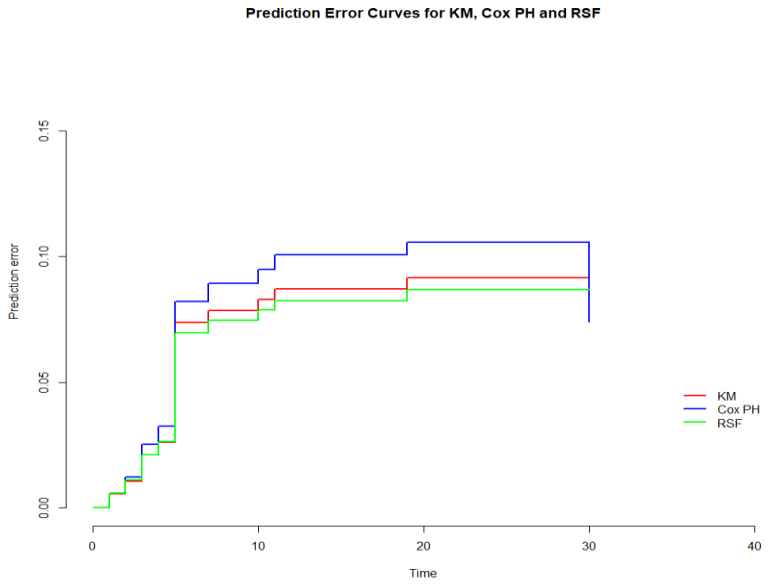


Figure 2: Models Predictive Error Curves for Cox PH, KM and RSF

Predictive Accuracy Based on the Concordance Index

Under the concordance index (C-index) technique the result showed that both the Cox PH and RSF models with cross-validated median C-index values of 0.68 and 0.735, respectively, are closer to 1. Furthermore, the Cox PH has higher variability and is right skewed compared to the RSF as portrayed in Figure 3

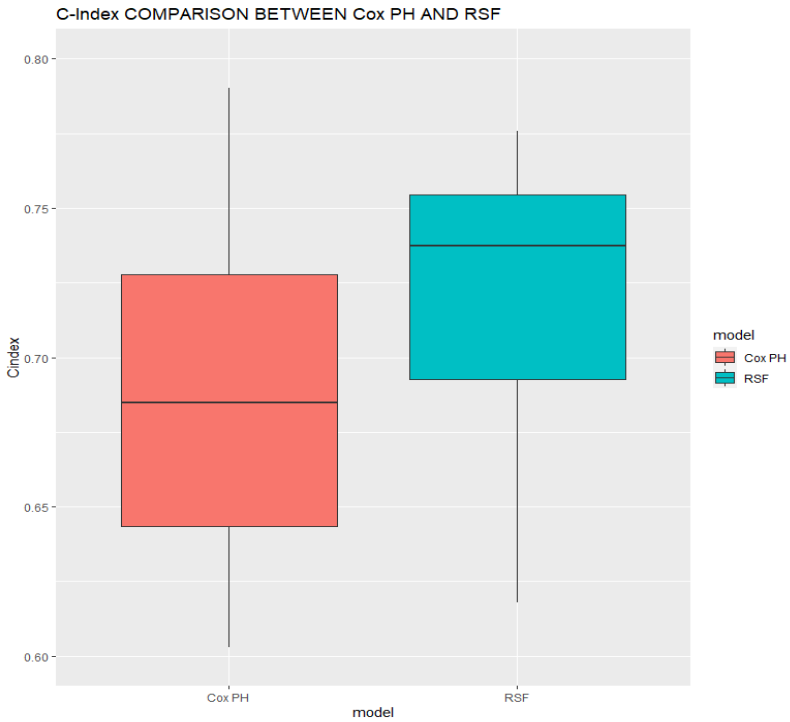


Figure 3: Concordance index of Cox PH and RSF

Predictive Accuracy Based on Brier Score

From Figure 4, the study shows that the Cox PH and RSF have a median predictive error in terms of brier scores of 0.105 and 0.065, respectively, and that Cox has higher variability in the predictive error and left is skewed compared to RSF which has smaller error predictive variability and right skewed.

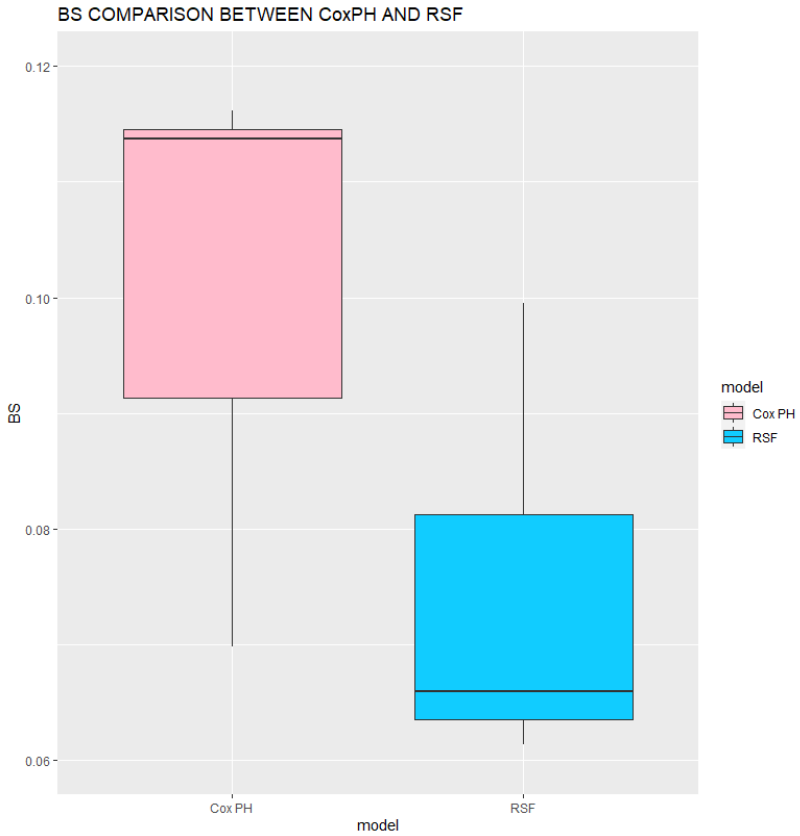


Figure 4: Brier score of Cox PH and RSF

Predictive accuracy based on the integrated Brier score

From Figure 5, the study shows that the Cox PH and RSF models have time prediction errors in terms of Integrated Brier scores of 0.14 and 0.013, respectively, and the Cox PH has higher variability in terms of Integrated Brier score and left skewed compared to the RSF with smaller variability and right skewed.

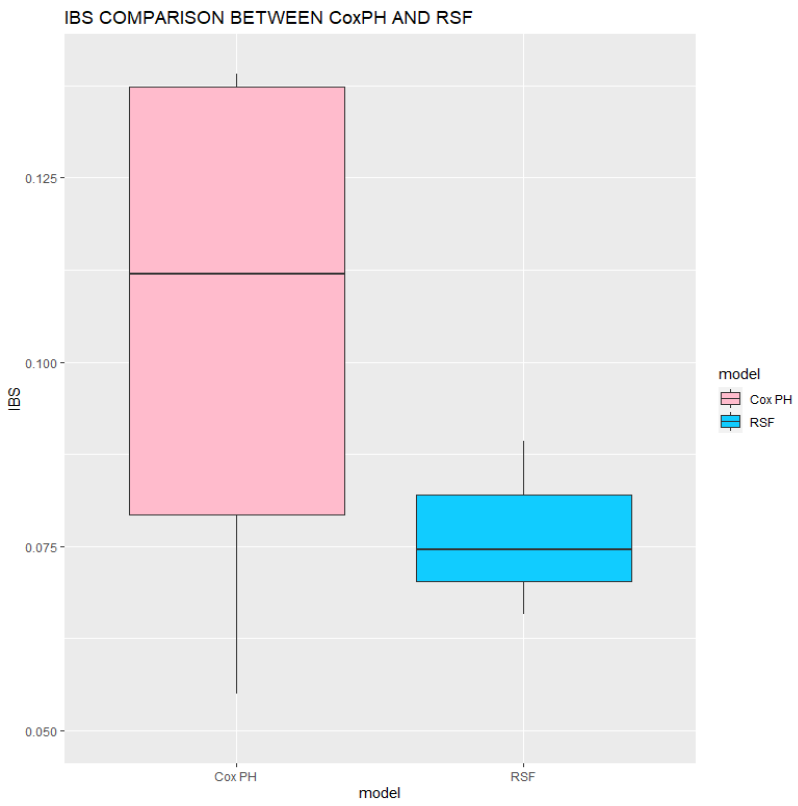


Figure 5: Integrated Brier score for RSF and Cox PH models

DISCUSSION

Random Survival Forest is promising to become an alternative model to the most common and standard Cox Proportional Hazard models for the analysis of survival time to the interesting outcome of an event [13,16,18,24]. Although the Cox model performed better compared to the RSF model for the covariates that satisfy the PH assumption, the RSF outperformed the Cox regression model in the situation where the PH assumption is violated [18]. For that reason, this study compared the predictive accuracy and covariate selection performance of the Cox PH model to the RSF by analyzing the advanced HIV dataset. Furthermore, the study compared the predictive accuracy and predictors' selection performance of the models without testing the proportional hazard assumption as portrayed by Bellera et al [11]. Over 25,000 citation since Cox's original paper was established, did not test the proportional hazard assumption and only 64 tested the assumption.

Under the Cox PH model, results showed that participants' age, Patient care, Classical TB symptoms, Prior TB diagnosis and TB episode number were highly significantly associated to advance HIV patients' death rate. Other predictors such as Absolute CD4 count, Respiratory rate, Patient ability to walk, Gender, pulse rate and anti-biotic on day one showed as important in explaining death rate in advanced HIV patients. Unlikely covariates were Education status, Cough, Hemoptysis, Weight loss, Night sweat, Chest pain, Prior TB medication, Patient on ART, Hemoglobin, Karnofsky score, Xpert-MTB results, LF-LAM results, Marital status and, Occupation status and fever.

Results from RSF model showed that Participants' age, Absolute CD4 count, Hemoglobin, Pulse rate, Karnofsky score, Xpert-MTB results, Patient ability to walk and Patient care were highly significantly associated with death rate to advanced HIV patients. Extra covariates such as Gender, Respiratory rate, Fever, Education status, Antibiotic day one, Classical TB symptoms, Hemoglobin, TB episode number, Night sweat and Occupation status are showed as important in understanding death rate in advanced HIV patients. Unlikely predictors were Prior TB medication, LF-LAM results, Patient on ART, Cough, Chest pain, Weight loss, Prior TB diagnosis and marital status. Hence, to compare the variables selection performance of these two models, the study take into account that covariates Participants' age and Patients' care are found to be highly significantly associated with the death of Advanced HIV patients to both models. Other covariates such as Classical TB symptoms, Prior TB diagnosis and TB episode number are considered to highly significantly by Cox PH but Classical TB symptoms and TB episode number found to be important covariates in explaining the death rate and Prior TB diagnosis found to be unlikely covariate in RSF model. Again, other covariates such as Absolute CD4 count, Hemoglobin, Pulse rate, Karnofsky score, Xpert-MTB results and Patient ability to walk are considered highly significant associated with death by RSF model. But among them, Absolute CD4 count, Pulse rate and Patients' ability to walk are considered to be important covariates in understanding the death and Hemoglobin, Karnofsky score and Xpert-MTB are considered as unlikely covariates in the Cox PH model this implied that it follows the non-linearity and highly dimensional covariates.

This matched with the study conducted by Nasejje and Mwambi [18] who explained that all factors which do not satisfy the proportional hazard assumption seem to be highly significant to the association of under-five children mortality when they used RSF. In line with these results, it was revealed that the Cox proportional hazard model has an inbuilt nature of compatibility with all covariates which satisfy the proportional hazard assumption even if they are not tested to align with the assumption. Likewise, this study differs from the study conducted by Nasejje and Mwambi [18] to test the predictive performance of the model in which they employed only the Integrated Brier score while this study employed the C-index, Brier score, Integrated Brier score, and Predictive error curves. Additionally, both models selected covariates such as Cough, Weight loss, Chest pain, LF-LAM results, marital status, and prior TB medication as unlikely or insignificant covariates. Moreover, covariates such as Education status, Hemoptysis, Night sweat, Occupation status, Patient on ART, and fever were found to be insignificant on the Cox PH model but all of them belong to the important covariates explaining the death in the RSF model. Generally, from the above explanations, the study concluded that RSF model outperformed the Cox PH model in the variables selection which is in line with [16,18]. On other hand, this study differ with the study conducted by Kim et al [16] in employing only Harrell's C-index as a technique of comparing the predictive accuracy performance to Deep learning base, RSF, and Cox proportional hazard models.

The study familiarized with the concordance index, Brier score, predictive error curves and Integrated Brier score as predictive accuracy measure. Aligned with results, as per Figure 2, the Cox PH model had higher prediction error curves than both the RSF model and the referenced model. On the other hand, the RSF model had a smaller predictive error curve when compared to both the Cox PH model and the referenced model curve; hence, the study concluded that the

RSF outperformed the Cox PH model in error fluctuation over time. This is matched with the result of the study conducted by Weathers & Cutler [28] who used predictive error curves and concordance index technique to compare the Cox proportional hazard model and RSF models in the covariates selection performance and they concluded that in both techniques the RSF outperformed the Cox PH model. Again, Figure 3 shows that the cross-validated median C-index values for the Cox PH and RSF models are 0.68 and 0.735, respectively, which are closer to one. This suggests that both models had a better fit, and the RSF model's higher cross-validation median C-index value indicates that it outperformed the cox PH model in terms of survival outcome predictive accuracy. Moreover, Figure 4 reveals that the Cox PH and RSF had median brier score values of 0.105 and 0.065, respectively, with the RSF having a lower median brier score value than the Cox PH, implying that the RSF model outperformed the Cox PH model in prediction accuracy. Furthermore, Figure 5 shows the median IBS values of 0.14 and 0.013, respectively. The RSF's lower median IBS value compared to the Cox PH model suggests that the RSF outperformed the Cox PH model in the advanced HIV dataset survival model predictive of the outcome of the interesting event. This is aligned with the study conducted by Dietrich et al [24], Du et al [13], Kim et al [16], Nasejje et al [17], Nasejje & Mwambi [18], Saha [29], Tilman [5] and Wubie [30].

The differences is revealed by Tilman [5] who employed C-index, Integrated Brier score and Calibration plots techniques in comparing Partial Logistic regression model with Artificial Neural Network (PLANN), RSF and Cox model. Its results showed that in terms of IBS and Calibration plots, the RSF and Cox models have similar predictive performance and in terms of C-index, the Cox model had better predictive accuracy performance compared to RSF. and similarly Wubie [30] differ with this study by conducting the comparison in the covariates selection performance using data clustering and concluded that both Cox proportional hazard and RSF yield the same results when it comes to selecting the most relevant covariates for predicting survival status.

CONCLUSION

All previous studies that were conducted to compare the predictive accuracy and covariates selection performance of the Cox proportional hazard models and random survival models were based on two sides. One side tested the proportional hazard assumption and concluded that the Cox PH outperformed the RSF and the second side when the proportional hazard assumption was violated by covariates and concluded that the RSF outperformed the Cox PH model, which make them fail to come up with one consensus. Therefore, even though the PH assumption was relaxed on all covariates of the dataset, still, Cox PH model selected only a few covariates, and all deselected covariates were highly significant in the RSF, which implies Cox PH model was inbuilt with the proportional hazard assumption. Aligned with this, only covariates that satisfied the PH inbuilt assumption were selected regardless assumption was or was not tested and making it ineffective compared to a random survival forest. Consequently, having a dataset free from non-linearity and interrelated variables is impossible in the real world; hence, the RSF model outperforms the Cox PH model in predictive accuracy and covariates selection performance. In addition, the study proposed the Random Survival Forest model as an alternative to the Cox PH model. However, few reviews supported RSF as the promised alternative in analyzing survival datasets [6,12,18,23]. As a result, the study recommended

clinicians and biostatisticians apply the RSF model when analyzing survival or time-to-event datasets in order to pinpoint any covariates that violate the PH assumption and are interrelated.

Conflicts of interest

The authors declared no potential conflicts of interest with respect to the research, authorship, or publication of this article.

Reference

1. Wang W, Liu W. Integration of gene interaction information into a reweighted random survival forest approach for accurate survival prediction and survival biomarker discovery. *Sci Rep* 2018; 8: 1–14.

2. Freeman SC, Cooper NJ, Sutton AJ, et al. Challenges of modelling approaches for network meta-analysis of time-to-event outcomes in the presence of non-proportional hazards to aid decision making: Application to a melanoma network. *Stat Methods Med Res* 2022; 31: 839–861.

3. Sun T, Wei Y, Chen W, et al. Genome-wide association study-based deep learning for survival prediction. *Stat Med* 2020; 39: 4605–4620.

4. Huang Z, Zhang H, Boss J, et al. Complete hazard ranking to analyze right-censored data: An ALS survival study. *PLoS Comput Biol* 2017; 13: 1–21.

5. Tilman N. *Prediction models for survival data with machine learning: an application to soft tissue sarcoma cohort statistical science for the life and behavioural science*. Universiteit Leiden, 2020.

6. Liu Y, Zhou S, Wei H, et al. A comparative study of forest methods for time-to-event data: variable selection and predictive performance. *BMC Med Res Methodol* 2021; 21: 193.

7. Babińska M, Chudek J, Chełmecka E, et al. Limitations of Cox proportional hazards analysis in mortality prediction of patients with acute coronary syndrome. *Stud Logic, Gramm Rhetor* 2015; 43: 33–48.

8. Austin PC. Statistical power to detect violation of the proportional hazards assumption when using the Cox regression model. *J Stat Comput Simul* 2018; 88: 533–552.

9. Xue Y, Schifano ED. Diagnostics for the Cox model. *Commun Stat Appl Methods* 2017; 24: 583–604.

10. Waititu HW, Koske JKA, Onyango NO. Analysis of Balanced Random Survival Forest Using Different Splitting Rules : Application on Child Mortality. 2021; 11: 37–49.

11. Bellera CA, MacGrogan G, Debled M, et al. Variables with time-varying effects and the Cox model: Some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC Med Res Methodol* 2010; 10: 20.

12. Wang H, Zhou L. Random survival forest with space extensions for censored data. *Artif Intell Med* 2017; 79: 52–61.

13. Du M, Haag DG, Lynch JW, et al. Comparison of the tree-based machine learning algorithms to cox regression in predicting the survival of oral and pharyngeal cancers: Analyses based on seer database. *Cancers (Basel)* 2020; 12: 1–16.

14. Miao F, Cai YP, Zhang YX, et al. Predictive modeling of hospital mortality for patients with heart failure by using an improved random survival forest. *IEEE Access* 2018; 6: 7244–7253.

15. 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 .* Technischen Universität Berlin, 2016.

16. Kim DW, Lee S, Kwon S, et al. Deep learning-based survival prediction of oral cancer patients. *Sci Rep* 2019; 9: 1–10.

17. Nasejje JB, Mwambi H, Dheda K, et al. A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC Med Res Methodol* 2017; 17: 115.

18. Nasejje JB, Mwambi H. Application of random survival forests in understanding the determinants of under-five child mortality in Uganda in the presence of covariates that satisfy the proportional and non-proportional hazards assumption. *BMC Res Notes* 2017; 10: 1–18.

19. Abeysekera W, Sooriyarachchi M. Use of Schoenfeld's global test to test the proportional hazards assumption in the Cox proportional hazards model: an application to a clinical study. *J Natl Sci Found Sri Lanka* 2009; 37: 41.

20. Binder DA. Fitting Cox's proportional hazards models from survey data. *Biometrika* 1992; 79: 139–147.

21. Yosefian I, Mosa Farkhani E, Baneshi MR. Application of Random Forest Survival Models to Increase Generalizability of Decision Trees: A Case Study in Acute Myocardial Infarction. *Comput Math Methods Med* 2015; 2015: 1–6.

22. Lipson R. *Predicting Ovarian Cancer Survival Times: Performance of Parametric Methods and Random Survival Forests*. Simon Frasel University, http://summit.sfu.ca/item/13829 (2014).

23. Rono CM. *Feature selection methods and Resampling techniques in Survival data: Determination of risk factors of under-five child mortality Research*. University of Nairobi, 2021.

24. Dietrich S, Floegel A, Troll M, et al. Random Survival Forest in practice: a method for modelling complex metabolomics data in time to event analysis. *Int J Epidemiol* 2016; 45: 1406–1420.

25. Haider H, Hoehn B, Davis S, et al. Effective ways to build and evaluate individual survival distributions. *J Mach Learn Res* 2020; 21: 1–63.

26. Ishwaran H, Kogalur UB, Blackstone EH, et al. Random survival forests. *Ann Appl Stat* 2008; 2: 841–860.

27. Wanyonyi KS. *Comparison of Random Survival Forests Split rules in selecting the determinants of under-five mortality in Kenya using 2014 DHS data.* University of Nairobi, https://math.umn.edu/sites/math.umn.edu/files/math.umn_.edu-newsletter_2009_0.pdf (2019).

28. Weathers B, Cutler R. Comparision of Survival Curves Between Cox Proportional Hazards, Random Forests, and Conditional Inference Forests in Survival Analysis. *Utah State Univ Thesis*; 2, https://digitalcommons.usu.edu/gradreports/927 (2017).

29. Saha S, Ryu D, Ebrahimi N. Variable Selection with Random Survival Forest and Bayesian Additive Regression Tree for Survival Data. 2019; 48824: 1–27.

30. Wubie BA. *Clustering Survival Data using Random Forest and Persistent Master of Science*. University of Alberta, 2016.