Croatian Operational Research Review
CRORR **17:2**(2026), 237–252

Multimethod survival analysis for identifying predictors and forecasting mortality in a heart patient cohort study

Syed Wajahat Ali Bokhari [1,*], **and Nasir Ali**[1]

[1] *Department of Statistics, PMAS-Arid Agriculture University, Rawalpindi 46300, Pakistan*
E-mail: ⟨wajahatbokhari2@gmail.com, nasir_stat@uaar.edu.pk⟩*

Abstract. This study presents a multi-method survival analysis of 125 cardiac patients from IIMCT-Pakistan Railway Hospital in Rawalpindi, Pakistan. Parametric accelerated failure-time modeling identified the Weibull distribution as optimal for describing time-to-event data. Semi-parametric analyses, including Cox proportional hazards and Bayesian Cox regression, consistently identified hypertension, ischemic heart disease, and smoking as significant predictors of elevated mortality risk. Higher systolic blood pressure demonstrated a protective effect. Kaplan-Meier analysis revealed steadily declining survival rates up to 300 days with no significant gender differences. The random survival forest model achieved robust predictive accuracy, identifying ischemic heart disease, smoking, and age as the most influential predictors. Our multi-methodological approach demonstrates the value of integrating parametric, semi-parametric, Bayesian, and machine learning techniques for comprehensive risk assessment in cardiac patient cohorts, offering potential for enhanced clinical risk stratification and personalized prognosis.

Keywords: Survival analysis, heart failure, random survival forest, Bayesian Cox regression, personalized risk prediction.

Received: July 11, 2025; accepted: November 26, 2025; available online: February 23, 2026

DOI: 10.17535/crorr.2026.0018

Original scientific paper.

1. Introduction

Cardiovascular diseases are the highest causes of death worldwide accounting for 17.9 million deaths yearly with HF contributing highly to this burden [28]. In spite of the breakthroughs in pharmaceutical and device oriented management of HF, the outcome of the disease continues to be poor with the five-year survival rate of about 50% even now [30]. Identification of high-risk patients at an early stage and precise estimation of survival outcomes imperatively contribute to treatment strategy optimization and success of patients' outcomes [22, 10].

Survival analysis becomes essential in HF research, particularly if outcomes are time-dependent, and the data are right censored. The Kaplan–Meier (KM) estimator is still a frequently applied non-parametric tool to estimate survival functions, and compare groups with log-rank tests [16]. Its ease of interpretation has made it a standard for the clinical studies even though it does not have any ability to adjust for covariates and assumes homogeneity across groups. The variance of KM estimates is usually computed with the Greenwood's formula [9].

To control for the effects of covariates, the semi-parametric nature of the Cox proportional hazards (PH) model makes it widely used, and would provide interpretable hazard ratios without the need to estimate the baseline hazard function [7]. Nevertheless, it assumes proportional

*Corresponding author.

hazards which can be violated in heterogeneous clinical populations. Schoenfeld residuals among other tools help diagnose such violations [25].

When assumption of Proportional Hazards does not hold, parametric survival models like Accelerated Failure Time models (AFT): Weibull, exponential, log-normal, and log-logistic provide the flexibility as they directly model the distribution of survival time [17]The quality of these models can be enhanced to give more accurate estimates if the underlying distribution was well calibrated. Model selection criteria including Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) help identify the most appropriate model by balancing goodness of fit with complexity [5].

Bayesian survival analysis extends these techniques so as to accommodate prior information and give posterior distributions for the model parameters. This is especially beneficial to studies using small sample size or complex hierarchical structure [14, 2]. Bayesian Cox and AFT models have exhibited more flexibility of quantifying uncertainty, than their frequentist opponents.

Recently, machine learning techniques, including random survival forests (RSF) came out as strong methods for survival analysis. While RSF extends Breiman's Random Forests to censored data it does so automatically allowing nonlinear effects and interactions without stringent parametric assumptions [15]. In HF studies, RSF has been shown to provide excellent predictive capabilities and usefulness for discovering new prognostic factors (concordance indices often greater than 0.70) [21].

Considering the complexity of HF and the relevance of correct risk stratification, this study compares a wide array of survival analysis methods, such as KM estimation, Cox and Bayesian Cox models, AFT models and RSF, on a retrospective sample of 125 cardiac patients collected from IIMCT–Pakistan Railway Hospital in Rawalpindi, Pakistan. Our objectives are to: (i) evaluate distributional fit based on AIC and BIC, (ii) determine significant predictors derived from Cox and Bayesian models, (iii) visualize survival differences using KM curves, and (iv) analyze RSF metrics for risk prediction and feature importance. This combined approach is intended to support the choice of strong modeling strategies in individualized HF prognosis.

In Section 2, we describe our materials and methods for the study. We estimate Kaplan–Meier survival curves and carry out log-rank tests of the difference between groups, train a random survival forest (RSF) for prediction and feature importance ranking, and both, the accelerated failure time (AFT) models to determine the Weibull distribution as the most suitable parametric model and Cox regression under the proportional-hazard assumptions. Section 3 provides a methodological synthesis and analysis, specifically, across all methods: reaffirming the prognostic relevance of hypertension, ischemic heart disease, and smoking, checks and balance on model assumptions, and checking RSF's forecasting performance, lastly Section **??** offers general conclusions on the methods' performance, inference stability, and on the potential of individualized risk assessment for use in clinical practice.

2. Materials and methods

2.1. Data collection and ethical considerations

This study was a retrospective cohort design based on the patients' Electronic Health Record data from the Pakistan Railway Hospital, Islamic International Medical College Trust (IIMCT) with the permission from the hospital management. The participants recruited in the study included all those who were admitted to the hospital for cardiac evaluation from January, 2018 to December, 2023, and out of them, 125 patients only were included and followed up completely.

We enlisted research personnel to abstract all the data from the paper and electronic medical records of the patients and input the information in the database. Some of variables included in the study were age, gender, hypertension, ischemic heart disease, diabetes mellitus, anemia, smoking, water-filter use, education level, systolic arterial blood pressure at admission, event

status and, time to event. Patient identifiers were stripped off in consonance with ethical considerations and the patients were assigned unique identification numbers for analysis purposes. Hypertension was categorized as binary (1 = has high blood pressure, 0 = no). In every regression model, the non-hypertensive patients will be the reference category. Therefore, when the coefficient is positive (or HR > 1), the level of risk is increased with hypertension. Systolic blood pressure (SBP) was measured as a continuous variable in millimeters of mercury (mmHg) and all effects reported were in millimeters of mercury per 10 mmHg (mmHg) so that all models could have consistent results.

2.2. Non-parametric methods

2.2.1. Kaplan–Meier estimator

Kaplan–Meier test was used to evaluate the survival function from a time-to-event data format. This non-parametric method takes into consideration of right-censored data and complements the previous one with displaying probability of survival in time.

Kaplan–Meier survival function is defined as

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right), \tag{1}$$

where t_i is the time of the i-th event, d_i is the number of events at t_i, and n_i is the number at risk just prior to t_i.

The Kaplan–Meier survival curve based on the total number of 125 patients is depicted in Figure 1. This curve offers information in estimating the survival probability of the whole population at a given period.

To determine the gender related differences in terms of survival, Kaplan–Meier curves depicted in Figure 2 were constructed for males and females separately. In order to determine statistical differences between these two groups, the log-rank test was applied. The null hypothesis of the log-rank test is that there is no difference in the survival distribution between the male and female patients.

2.2.2. Random survival forest (RSF) model

The random survival forest (RSF) is an extension of the bagged decision tree model proposed by Breiman for right censoring survival data, in that an ensemble of survival trees is constructed to estimate the survival functions for each subject without giving any specific functional form for the hazard or the survival distribution [15]. Since every of the B trees is built with a bootstrap sample of the initial data set, about one third of observations are rejected from the bootstrap (out-of-bag, OOB) and are used to calculate non-biased internal estimates of prediction error. At each node, a random selection of some features out of candidate is performed, and a split that provides the maximum log-rank test statistic that means that the splitting aims at differences in survival rates of daughters, is selected [19]. Only trees are grown to maximal depth so that no pruning is done to eliminate splitting of a variable and the ensembles are used to reduce variability [1].

At each terminal node of tree b, one estimate $\hat{S}_b(t \mid X)$ is computed with the aid of Kaplan–Meier estimator on the OOB samples. The overall RSF survival function for a new covariate vector X is then the simple average

$$\hat{S}(t \mid X) \;=\; \frac{1}{B} \sum_{b=1}^{B} \hat{S}_b(t \mid X), \tag{2}$$

This results in a fully non-parametric estimator of the survival curve that is built by averaging across all the trees in the forest [11].

Model performance is evaluated on the basis of the test data set which is different from out-of-bag (OOB) data. It enables obtaining the survival curves outside the training data, which provides the estimates of the prediction error, including the concordance index, without the need for the test set [29]. To determine variable importance, all the covariates in the OOB is permuted and the increase in the prediction error measured; predictors are ranked according to the extent of contributing to survival discrimination [23].

Key advantages of RSF include:

- No distributional assumptions – the model does not specifically assume any functional form for the hazard and survival functions [15].
- High-dimensional robustness – it can accommodate many high-dimensional correlated covariates by using the randomized splitting and bagging strategies [3].
- *Built-in imputation* – where the missing covariate values are dealt within by using tree-based imputation techniques [25].

Terminology. The results of cox regression are expressed as hazard ratio (HRs), whereas the results of parametric accelerated-failure-time (AFT) model are expressed as acceleration factors (AFs). A positive AF has been known to correspond to increased survival time (protective) and a negative AF corresponded to reduced survival time (increased risk). This difference introduces uniformity between the time-based and hazard-based model interpretations.

2.3. Parametric methods

In survival analysis the method most commonly applied for assessing the impact of different factors on the time to a specific event is the CPH. This method has no links with any specific survival distribution but with a basic assumption that predictor variables affect survival equally at all times. Nevertheless, it may not give a good result if the distribution of the random variable is based on normal distribution. To overcome this weakness, various forms of parametric survival models including Weibull, exponential, log-normal and log-logistic models are used. These models make certain assumptions about the distribution of survival times and are appropriately suited for a range of applications when chosen. The mathematical formulations of these models are as follows:

- *Weibull model:* The Weibull distribution model can fit different shape of hazard function. Its probability density function (PDF) is given by:

$$f(x;\sigma,p) = p \cdot \frac{x^{p-1}}{\sigma} e^{(-x^p/\sigma)}, \quad x > 0,\ \sigma > 0,\ p > 0. \tag{3}$$

  In model, $\sigma > 0$ is known as the scale parameter, which controls the spread of the distribution and $p > 0$ is the shape parameter that governing behaviour of the hazard function. It is particularly useful in modelling experiments where hazards may change with time such as wear-out failures or improvements with time [17].

- *Exponential model:* The exponential distribution is a special case of the Weibull distribution in which $p = 1$. Its PDF is:

$$f(x;\sigma) = \frac{1}{\sigma} e^{-x/\sigma}, \quad x > 0,\ \sigma > 0. \tag{4}$$

  This distribution involves a constant hazard rate, therefore work well in memory less processes including failure processes in systems which have no aging impact [17].

- *Log-normal model:* The log-normal distribution is derived when distribution of logarithm of the survivable life duration is normal. Its PDF is expressed as:

$$f(x;\mu,\sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, \quad x \in (0,\infty),\ \mu > 0,\ \sigma > 0. \tag{5}$$

  where the parameters which determine the location and spread of the distribution are termed as mean (μ) and standard deviation (σ), respectively. This model is appropriate when the data is positively skewed as are many survival times that are affected by multiply factors [17].

- *Log-logistic model:* Another distribution from the class of parametric models, useful for skewed data on survival, is the log-logistic distribution. Its PDF is given by:

$$f(x;\alpha,\beta) = \frac{(\beta/\alpha)(x/\alpha)^{\beta-1}}{[1+(x/\alpha)^{\beta}]^2}, \quad x > 0,\ \alpha > 0,\ \beta > 0. \tag{6}$$

  whereas α stands for scale parameter and β stands for shape parameter. This model is useful when the hazard function rises and then falls over time, and it is good for time-to-event data with fleshy tails [17].

All the parametric models that had been fitted in accelerated failure time (AFT) framework with the coefficients of fit, denoted by β, which act on $logT$. We therefore present the acceleration factors (AFs) in lieu of hazard ratios, in which the acceleration factor (AF) is given as $\exp(\beta\,\Delta x)$. The value of AF below 1 represents a more harmful and the value of AF above 1 represents a more protective. In continuous predictors the effects are expressed as a clinically significant change (e.g. SBP per 10 mmHg): $\mathrm{AF}_{10} = \exp(10\beta)$, and in binary ones the AF is the level "1 = Yes" and "0 = No." In the parameterization of the covariate effects shown below under the AFT, we have covariate effect by describing it as: $\mathrm{AF} = \exp(\beta)$ (or $\exp(\beta\,\Delta x)$).

The selection of these parametric models depends on characteristics of the data and the assumptions made on the form of the hazard function. They offer a large benefit when CPH's assumption of proportional hazards doesn't apply or if a parametric strategy is likely to provide greater estimation accuracy. The use of these models provides more flexible computation and more accurate estimates when used in survival analysis.

2.4. Semi-parametric methods

2.4.1. Cox proportional hazards model

The Cox Proportional Hazards (PH) model was used in order to evaluate effect of more than one variable in hazard rate. The semi-parametric version of the model can define the hazard ratios without estimating the baseline hazard rate:

$$h(t|X) = h_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p) \tag{7}$$

where $h(t|X)$ is the hazard at time t assuming covariates set to be X and $h_0(t)$ is the unknown baseline hazard. The partial likelihood method was used for fitting the models using the Cox PH function from the `coxph()` in R. To evaluate the PH assumption, the Schoenfeld residuals test was done using the function `cox.zph()`.

2.4.2. Bayesian Cox regression analysis

A Bayesian version of the Cox model was also fitted, with the baseline distribution specified as Weibull. This model takes into account the prior information as well as offers the probability of arriving at the hazard ratios. The model specification was:

$$\log h(t|X) = \log\lambda + \gamma\log t + \beta_1 X_1 + \cdots + \beta_p X_p \tag{8}$$

where λ and γ are scale and shape parameter of the Weibull distribution respectively. The model was fitted using the `survregbayes()` function in R that fits a proportional hazards parameterisation and a Weibull baseline. Various convergence diagnostics and posterior summaries were calculated based on the results of the model.

3. Results and discussion

3.1. Results of non-parametric methods

3.1.1. Results of Kaplan–Meier estimator

Figures 1 and 2 display the Kaplan-Meier (K-M) survival curves. Even the overall survival rate of the cohort of 125 patients (Figure 1) reveals that, over a 300 day interval, there is gradual decline in survival. The 95% confidence interval (shaded region) gives an indication of the uncertainties associated with the survival estimate, and this is a function of the number at risk, which curtails as more members at risk at a given time and expands as the size of the sample shrinks over time. Relations of the survival distributions for gender, obtained from Figure 2, show slightly lower survival probabilities for the male patients in comparison to the female throughout the perceptible period of the survival, while the log-rank test has non-significant p-values ($p > 0.05$) and one could not identify any difference in survival between the genders. The number at the risk table that appears beneath each curve depicts, a diminishing number of participants under observation while time elapses.

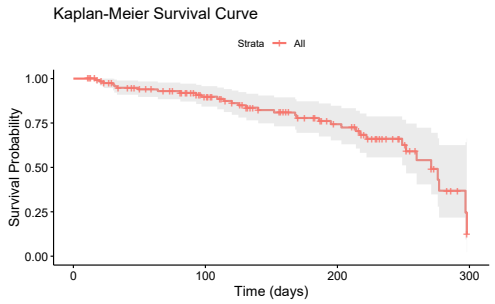


Figure 1: *Kaplan–Meier survival curve for the overall cohort. Time-to-event in days. Y-axis: survival probability; X-axis: follow-up (days).*

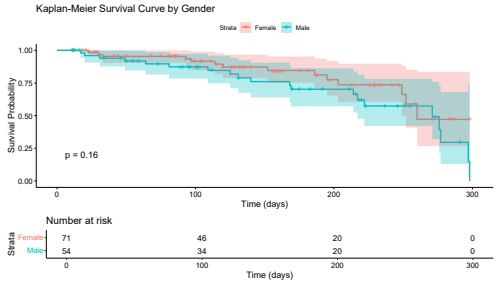


Figure 2: *Kaplan–Meier survival curves by gender. Time-to-event in days. Y-axis: survival probability; X-axis: follow-up (days). Numbers at risk indicated below each curve at each 50-day time point.*

3.1.2. Results of random survival forest (RSF) model

Table 1 shows that RSF was trained on n=125 patients of which 34 had the event of interest. The ensemble contained 1000 trees, each of which grew up to a terminal node size of 15 with the average number of terminal nodes in every tree being 4.691. At each split 4 of the 10 available predictors were chosen at random, and 10 random split points were examined per variable from a log–rank "random" splitting rule. Bootstrap sampling without replacement (SWOR) employed 79 observations per tree, with the OOB error estimated with about one third of the cohort. OOB CRPS (standardized CRPS 0.1315) was 39.1909, corresponding to an OOB prediction error = 0.3312, true OOB prediction error under minimal-depth variable selection 33.1169.

Minimal-depth analysis showed that three top predictors according to minimal-depth values (1.321, 1.728, and 1.752) were, respectively, ischemic heart disease, smoking status, and age, the latter with variable importance (VIMP) reaching –0.009 which in absolute values is low.

Metric	RSF summary	Minimal depth selection
Sample size	125	125
Number of deaths	34	—
Number of trees	1 000	1 000
Forest terminal node size	15	15
Average no. of terminal nodes	4.691	—
No. of variables tried per split	4	4
Total no. of variables	10	10
Resampling used to grow trees	SWOR	—
Resample size used to grow trees	79	—
Analysis	RSF	—
Family	surv	surv
Splitting rule	logrank *random*	logrank
Number of random split points	10	10
(OOB) CRPS	39.1909	—
(OOB) stand. CRPS	0.1315	—
(OOB) Requested performance error	0.3312	33.1169
Conservativeness	—	medium
X-weighting used?	—	TRUE
Dimension (no. of vars)	—	10
Refitted forest	—	FALSE
Model size	—	3
Depth threshold	—	1.7559
Top 3 variables	**Depth / VIMP**	
ischemic_heart_disease	1.321 / 0.244	
smoking	1.728 / 0.102	
age	1.752 / –0.009	

Table 1: *RSF model summary and minimal depth variable selection*

The minimal-depth procedure having the medium conservativeness and X-weighting enabled fitted an unrefitted forest of size 3 with a depth threshold of 1.7559.

The minimal-depth and VIMP measures in the RSF are worth considering, as they are arising at different factors that represent the influence of the predictor. Minimal depth measures the early use of a variable in splitting in the forest which is important to the structure of trees whereas VIMP measures how permuting a variable affects prediction error. The age is very high in our findings in terms of minimal depth, but zero or slightly negative to VIMP. The seeming mismatch suggests that age is a significant contributor to the partitioning of trees (non-linear or interaction effects), but does not significantly enhance the predictive power when other factors (including ischemic heart disease and smoking) are accounted in. The role of age should thus be viewed as being context dependent and secondary to superior clinical predictors.

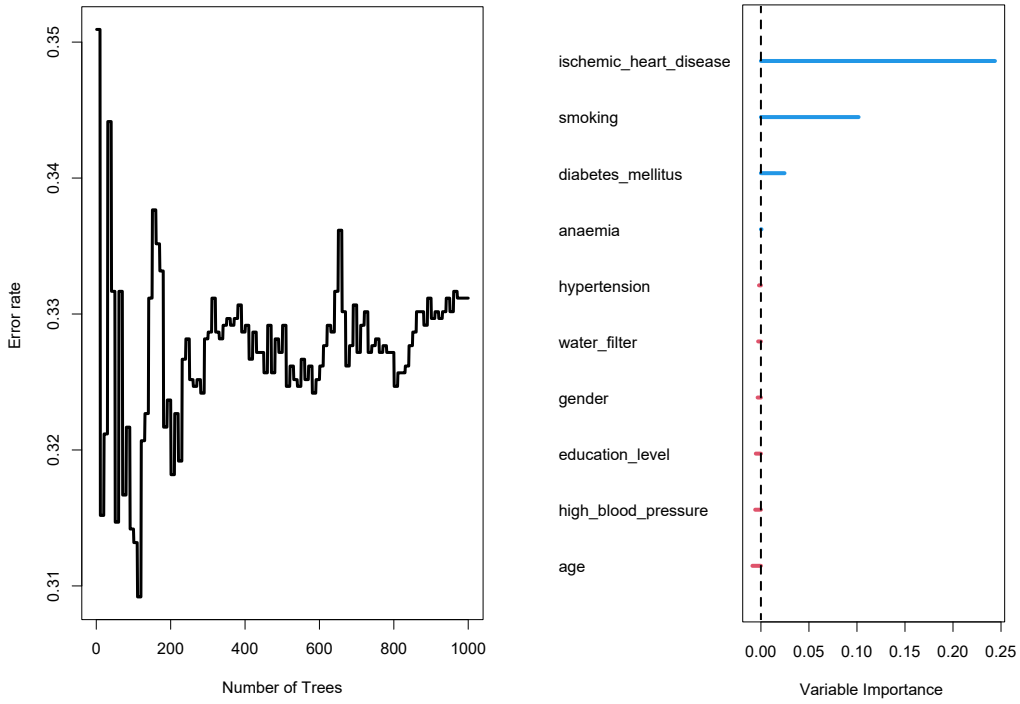
On the whole the RSF had strong performance and interpretability: it used an ensemble of fully mature survival trees for the nonparametric capturing of complex interactions, offered dependable OOB error estimates that do not require separate validation set, and was able to discuss the important covariates through minimal-depth selection.

The variable importance (Table 2) reveals that ischemic heart disease is the best predictor for patient's survival in our random survival forest model (RSF) (VIMP = 0.2437; excluding the station, relative importance = 1.0000), then by smoking status (VIMP = 0.1017; relative importance = 0.4173) and diabetes mellitus (VIMP = 0.0246; relative importance = 0.1008). Other covariates such as anemia have minor positive contributions (VIMP = 0.0006), while hypertension, water filter usage, gender, education level, high blood pressure, and age have

negligible or small negative uses (VIMP range: from –0.0029 to –0.0089), which reflects the weak influence on the model's performance in prediction (Figure 3). The ensemble survival curves (Figure 4) show the risk paths of the entire cohort, as the patient 81 showed the peak value of the predicted survival probability within the follow-up time; other patients' curve is superimposed over all 125 patient estimates and lets one see heterogeneity of the profiles and the ability of the model to generate individual survival estimates.

Variable	Importance	Relative Imp.
ischemic_heart_disease	0.2437	1.0000
smoking	0.1017	0.4173
diabetes_mellitus	0.0246	0.1008
anemia	0.0006	0.0025
hypertension	-0.0020	-0.0083
water_filter	-0.0029	-0.0119
gender	-0.0034	-0.0142
education_level	-0.0053	-0.0218
high_blood_pressure	-0.0060	-0.0245
age	-0.0089	-0.0364

Table 2: *Variable Importance from RSF Model*



Note: Left: OOB error and continuous ranked probability score (CRPS). Right: Minimal depth rankings and variable importance (VIMP). Best predictors: age, ischemic heart disease and smoking.

Figure 3: *Random survival forest (RSF) multimodal results.*

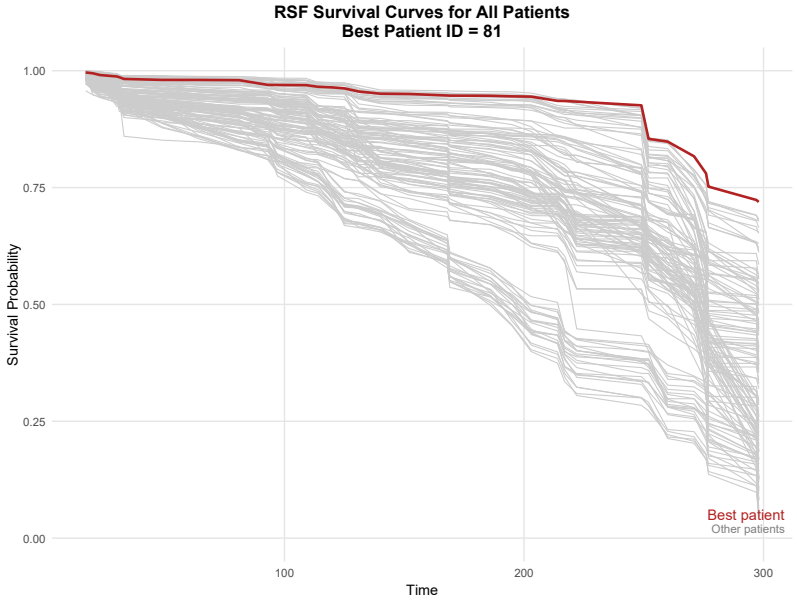


Figure 4: *RSF survival curves for all patients (best patient ID = 81).*

3.2. Results of parametric methods

Table 3 shows maximum-likelihood estimates for four candidate survival distributions: Weibull, Exponential, Log-Normal and Log-Logistic, adjusted for gender, age, hypertension, ischemic heart disease, smoking, diabetes mellitus, anemia, education level, water filter use, and high blood pressure. Across all models, hypertension (Weibull: $\beta = -0.9505$, $p = 0.0152$; Exponential: $\beta = -1.6581$, $p = 0.0122$), ischemic heart disease (Weibull: $\beta = -0.9050$, $p < 0.001$; Exponential: $\beta = -1.5044$, $p < 0.001$), and smoking (Weibull: $\beta = -0.6044$, $p = 0.0054$; Exponential: $\beta = -1.0580$, $p = 0.0067$) emerged as significant predictors of reduced survival time. There is a weak yet noticeable acceleration effect observed with high blood pressure (Weibull: $\beta = 0.0224, p = 0.0193$; Exponential: $\beta = 0.0393, p = 0.0230$) while other covariates such as gender, age, anemia, and education have weaker, non-significant influences in most specifications. It is necessary to note that negative coefficients of hypertension ($\beta < 0$) are associated with a reduced survival time, i.e. an increased hazard of hypertensive patients in the parametric AFT models when transformed into the hazard scale. Therefore, the coefficient sign is negative but the direction of the effect is similar to that of the Cox model where hypertension is a risk factor of mortality. In the case of systolic blood pressure (SBP), the positive coefficient ($\beta = 0.022$ in Weibull AFT model) represents an increment in the predicted survival time by 10 mmHg, and it proves the fact that increasing SBP is related to the decreased risk of mortality. Clinically, the results show that hypertension, ischemic heart disease, and smoking are significant predictors of shorter survival times, and that the patients with higher systolic blood pressure are more likely to have a little longer survival time.

Scale parameters' values differ by distribution (Weibull scale log-estimate -0.6555, $p < 0.0001$; Log-Logistic scale $\hat{\gamma} = 0.463$), whereby different baseline hazard shapes are used. The use of AFT coefficients as acceleration factors (AFs) simplifies the explanation of effects direction and strength of effects on time scale. According to the Weibull AFT model, the hypertension has a $\beta = -0.9505$ meaning the AF $= \exp(-0.9505) =$ **0.387** (the survival time of hypertensive people is about 61% shorter than of non-hypertensive ones).

	Weibull				Exponential				Log Normal				Log Logistic			
Variable	Coef	SE	z	p	Coef	SE	z	p	Coef	SE	z	p	Coef	SE	z	p
(Intercept)	4.136	1.120	3.69	0.0002	3.719	2.114	1.76	0.0785	3.602	1.502	2.40	0.016	3.879	1.394	2.78	0.0054
genderMale	-0.249	0.224	-1.11	0.266	-0.420	0.413	-1.02	0.310	-0.181	0.259	-0.70	0.486	-0.265	0.252	-1.05	0.294
age	0.001	0.006	0.19	0.850	-0.001	0.010	-0.05	0.961	0.004	0.006	0.69	0.490	0.003	0.006	0.51	0.612
hypertension1	-0.950	0.391	-2.43	0.015	-1.658	0.662	-2.51	0.012	-1.018	0.427	-2.38	0.017	-0.910	0.414	-2.20	0.028
ischemic_heart_disease	-0.905	0.250	-3.63	0.0003	-1.504	0.432	-3.48	0.001	-1.093	0.275	-3.98	<0.001	-0.991	0.274	-3.62	0.0003
smoking1	-0.604	0.217	-2.78	0.005	-1.058	0.390	-2.71	0.007	-0.602	0.259	-2.33	0.020	-0.619	0.249	-2.49	0.013
diabetes_mellitus	-0.333	0.230	-1.45	0.147	-0.837	0.405	-2.07	0.039	-0.632	0.263	-2.41	0.016	-0.445	0.251	-1.77	0.077
anemia1	0.124	0.241	0.51	0.609	0.191	0.440	0.43	0.664	0.293	0.275	1.06	0.288	0.225	0.265	0.85	0.394
education_levelPrimary	0.273	0.251	1.09	0.276	0.389	0.476	0.82	0.414	0.519	0.316	1.64	0.100	0.328	0.320	1.03	0.305
education_levelSecondary	-0.103	0.299	-0.34	0.732	-0.210	0.554	-0.38	0.704	0.248	0.371	0.67	0.504	0.106	0.360	0.30	0.767
water_filterNon-filter	0.289	0.211	1.37	0.172	0.309	0.392	0.79	0.430	0.189	0.266	0.71	0.478	0.230	0.250	0.92	0.358
high_blood_pressure	0.022	0.010	2.34	0.019	0.039	0.017	2.27	0.023	0.023	0.013	1.85	0.065	0.022	0.012	1.87	0.062
Log(scale)	-0.656	0.145	-4.53	<0.001	—	—	—	—	-0.123	0.125	-0.98	0.326	-0.770	0.144	-5.35	<0.001
Scale	0.519				1 (fixed)				0.884				0.463			

Note: Binary variables coded as (1 = Yes, 0 = No); reference category = absence of condition. Continuous variables in natural units (e.g., systolic blood pressure in mmHg per 10 mmHg increase). Acceleration factor (AF) = $\exp(\beta\Delta x)$*; AF > 1 indicates longer survival, AF < 1 indicates shorter survival.*

Table 3: *Parameter estimates of parametric AFT models (coefficients on* $\log T$*)*

Model	**Degrees of Freedom (df)**	**AIC**	**BIC**
Weibull	13	**466.5933**	**503.3614**
Exponential	12	480.5786	514.5184
Log-normal	13	476.1975	512.9656
Log-logistic	13	474.4945	511.2626

Table 4: *Model comparison using AIC and BIC for different survival distributions*

Ischemic heart disease has a $\beta = -0.9050$ that gives AF = **0.405** and smoking $\beta = -0.6044$ that gives AF = **0.546**, both reducing survival. In the case of systolic blood pressure (SBP), coefficient is in units per mmHg, the change in survival per unit of 10 mmHg increase in SBP results in the coefficient $AF_{10} = \exp(10 \times 0.0224) = \mathbf{1.251}$, or about 25 % longer survival with a +10 mmHg increase in SBP. These AFs are consistent with the protective direction of SBP in the Cox analysis and also parametric effects on the suitable time scale. Since the AFT model functions on the time scale and the Cox model functions on the hazard scale, the consequences of AFT are interpreted as acceleration factors (AF), whereas the consequences of Cox are interpreted as hazard ratios (HR). The raw coefficients might look inverted across models, but on their scale, both the raw coefficients of the models are the same, i.e. that the interpretation of the raw coefficients of the model using the scale of interpretation are equal. AIC and BIC comparison of models (Table 4) reveals that Weibull model achieves the lowest value of information criteria (AIC = 466.59; BIC = 503.36), is closely followed by the Log-Logistic (AIC = 474.49; BIC = 511.26) suggesting that the Weibull distribution is optimal model in terms of fit and parsimony for these data.

3.3. Results of semi-parametric methods

3.3.1. Results of Cox proportional hazards model

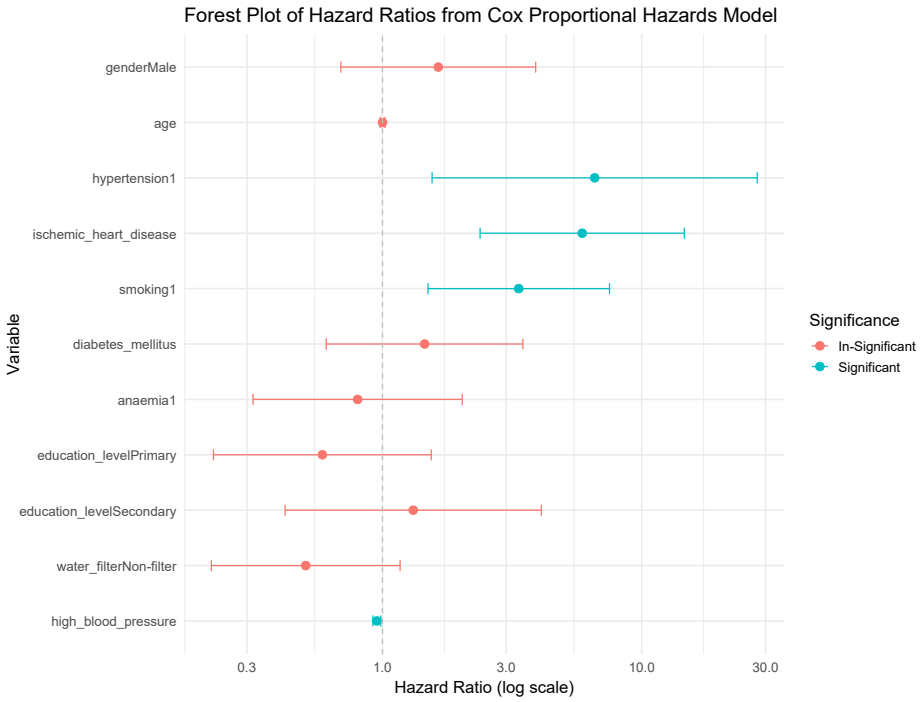
The multivariable Cox PH analysis (Table 5, Figure 5) determined that hypertension (HR = 6.58, 95% CI 1.55–27.87, $p = 0.0105$), ischemic heart disease (HR = 5.89, 95% CI 2.38–14.58, $p < 0.001$), and smoking (HR = 3.35, 95% CI 1.50-7.51, $p = 0.0033$) were significant predictors of mortality. Gender, age, and diabetes mellitus, anemia, education level, and water-filter use did not achieve statistical significance levels since the hazard ratios for these parameters were close to unity with wide confidence intervals crossing 1.0. The concordance, 0.807 (SE = 0.033), indicated excellent discrimination in the model, while the likelihood ratio $p < 0.0001$ reflected a good fit of the model.

Variable	Coef	exp(Coef)	SE(Coef)	z	$Pr(>\|z\|)$	95% CI
genderMale	0.4956	1.6415	0.4413	1.123	0.2614	(0.6912, 3.8983)
age	0.0010	1.0010	0.0108	0.090	0.9283	(0.9800, 1.0223)
hypertension1	1.8842	6.5810	0.7365	2.558	0.0105*	(1.5537, 27.8746)
ischemic_heart_disease	1.7734	5.8907	0.4625	3.834	0.0001***	(2.3795, 14.5829)
smoking1	1.2100	3.3536	0.4112	2.942	0.0033**	(1.4979, 7.5085)
diabetes_mellitus	0.3741	1.4537	0.4453	0.840	0.4009	(0.6073, 3.4798)
anemia1	-0.2203	0.8023	0.4740	-0.465	0.6421	(0.3169, 2.0313)
education_levelPrimary	-0.5335	0.5866	0.4934	-1.081	0.2796	(0.2230, 1.5428)
education_levelSecondary	0.2732	1.3142	0.5804	0.471	0.6379	(0.4213, 4.0993)
water_filterNon-filter	-0.6812	0.5060	0.4275	-1.593	0.1111	(0.2189, 1.1696)
high_blood_pressure	-0.0497	0.9515	0.0180	-2.769	0.0056**	(0.9186, 0.9856)

Note: Binary covariates (1 = Yes, 0 = No); reference category = absence of condition. Continuous variables SBP (mmHg) reported per 10 mmHg increase. Hazard ratio (HR) > 1 indicates higher hazard (shorter survival), HR < 1 indicates lower hazard (longer survival). Concordance = 0.807 (SE = 0.033); Likelihood ratio test = 39.68 on 11 df, p = 4e-05. ***Significance codes:*** ****$p < 0.001$, ** $p < 0.01$, * $p < 0.05$*

Table 5: *Summary of Cox proportional hazards model*

No systematic departures from proportionality were observed for any covariate on Schoenfeld residual plots (Figure 6), thus supporting that the model has valid PH assumption. In all the analyses, hypertensive group (coded as 1) had a consistently high risk over non-hypertensive patients. The differences between the signs of the coefficient in the AFT and Cox frameworks are due to the fact that the parameterizations of the two models are different: the coefficients of AFT are associated with the acceleration of time, whereas Cox coefficients are associated with the increase of hazards. The effect of SBP increase per 10 mmHg was used in the treatment; the negative coefficient (β = -0.0497) and the related HR = 0.95 which show that more SBP



Note: HR > 1 indicates higher hazard (risk); HR < 1 indicates protective effect. Covariates coded as 1 = Yes, 0 = No; SBP expressed per 10 mmHg increase.

Figure 5: *Forest plots of hazard ratios (HRs) and 95% confidence/credible intervals for the Cox model.*

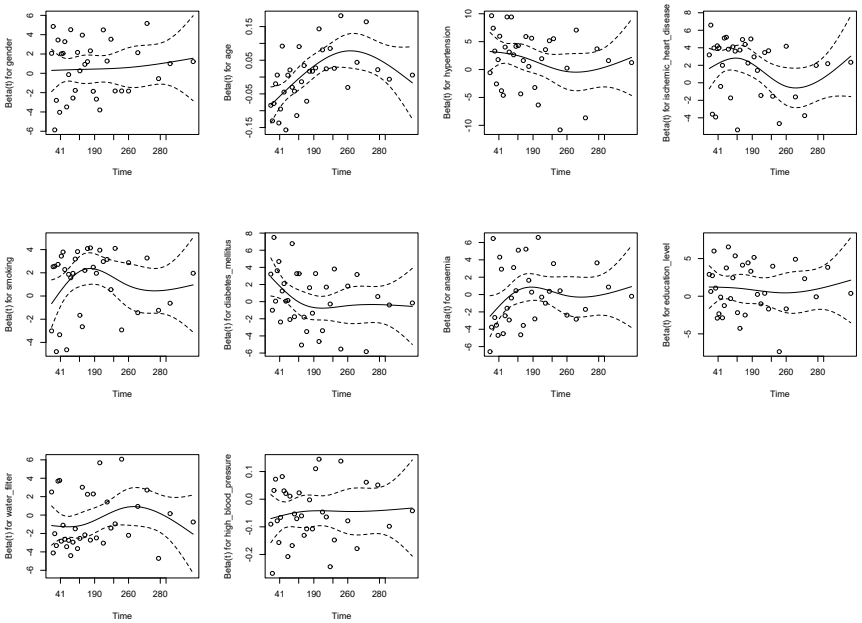


Figure 6: *Schoenfeld residuals for all variables.*

decreased hazard and therefore it was a protective effect. Clinically, these risk ratios indicate that patients with hypertension are at a higher risk of death (more than six times) than patients with normal blood pressure, and that ischemic heart disease and smoking are at a higher risk (three to five times, respectively). Conversely, a 10 mmHg elevation in systolic blood pressure lowers the risk of mortality by approximately 5 % points, which is of a slight protective effect.

It is also important to note that diabetes mellitus exhibited statistically significant effects in certain parametric AFT models (notably the log-normal and log-logistic specifications) but lost significance in the semi-parametric Cox regression framework. These discrepancies stem from the distinct assumptions underlying the two approaches: AFT models directly estimate effects on survival time, while Cox regression models effects on hazard rates and assumes proportional hazards. Because of this, covariates that influence survival time non-proportionally or interact with time may appear significant in one model and not in another. Hence, the diabetes effect should be interpreted cautiously, emphasizing convergence across multiple modeling frameworks rather than any single specification.

3.3.2. Results of Bayesian cox regression model

The Bayesian Cox regression (Table 6, Figure 7) gives estimates of posterior hazard ratio that largely agree with the frequentist Cox results: These include hypertension (posterior mean log-HR = 2.01; 95% CI: 0.62–3.44) and ischemic heart disease (mean = 1.89; 95% CI: 0.94–2.85) as the strongest risk factors, smoking (mean = 1.26; 95% CI: 0.44–2.17) taking second place, and all with credible intervals excluding zero. A consistent protective effect was observed for each 10 mmHg increase in systolic blood pressure (posterior mean log-HR = –0.0476, 95% credible interval: –0.0849 to –0.0097), but other variables, including gender, age, diabetes, anemia, education, and water-filter use have posterior intervals across zero, indicating mixed evidence on the effects of the independent variable based on these parameters. Model fit indices (LPML = –233.46, DIC = 460.94, WAIC = 465.38) indicated adequate predictive performance. The forest plot (Figure 7) displays strong associations for key clinical covariates while appropriately quantifying uncertainty for less influential predictors. Clinically, these Bayesian estimates corroborate the frequentist interpretation: hypertension, ischemic heart disease, and smoking remain key mortality risk factors, whereas higher systolic blood pressure contributes to improved survival probabilities. The case study shows the benefit of integrating classical and machine learning survival models to make risk predictions in underexplored populations. The RSF model enhanced predictive accuracy rates, but it also determined that there were other critical variables like age, which might have a nonlinear influence on mortality risk. Risk factors, as ischemic heart disease and smoking, have even stronger predictive power on mortality in our Pakistani sample, contrary to that typical of Western populations. These results hold significant clinical significance to risk assessment and resource planning in South Asia. Methodologically, the study will reflect the way that RSF can augment customary models through revealing complex associations in the data. The findings of the study have enabled us to make a case in favor of applying machine learning methods to survival analysis to diverse clinical scenarios where data can be scarce or risk characteristics can be dissimilar to historical cohorts.

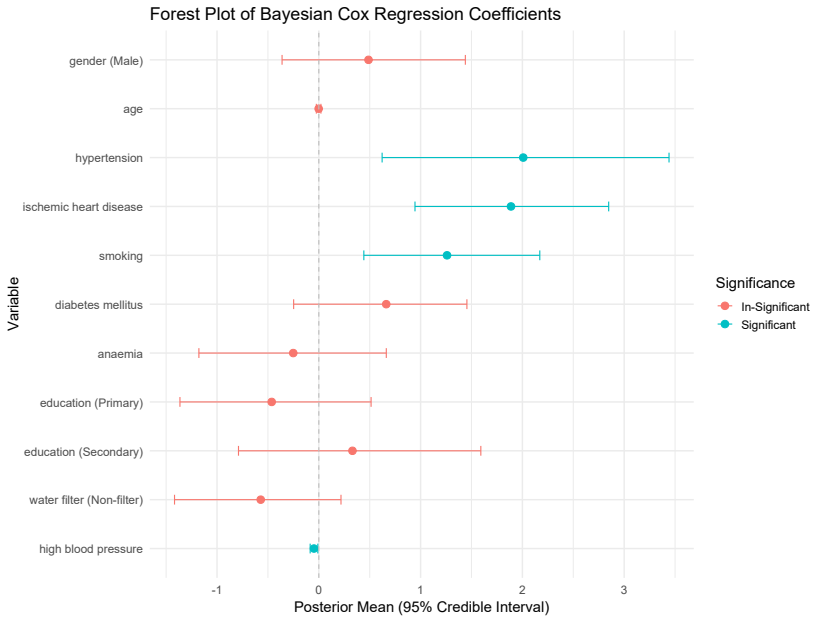
4. Conclusion

This study employed a multi-method survival analysis—integrating parametric, semi-parametric, Bayesian, and machine learning frameworks—to comprehensively identify mortality predictors in a Pakistani cardiac cohort. Our results provide critical insights into risk factors in South Asia and demonstrate the significant potential of combined modeling strategies, particularly random survival forests, for enhancing clinical risk prediction and personalizing patient care in this population. Across all analytical frameworks, hypertension, ischemic heart disease,

Variable	Mean	Median	Std. Dev.	95% CI (Low)	95% CI (Upper)
gender (Male)	0.4881	0.4679	0.4615	-0.3610	1.4397
age	-0.0016	-0.0016	0.0112	-0.0243	0.0209
hypertension	2.0084	2.0253	0.7155	0.6220	3.4407
ischemic heart disease	1.8880	1.8892	0.4831	0.9449	2.8478
smoking	1.2597	1.2606	0.4361	0.4424	2.1710
diabetes mellitus	0.6623	0.6780	0.4331	-0.2473	1.4549
anemia	-0.2513	-0.2576	0.4739	-1.1782	0.6630
education level (Primary)	-0.4629	-0.4591	0.4834	-1.3643	0.5135
education level (Secondary)	0.3305	0.3122	0.5955	-0.7891	1.5917
water filter (Non-filter)	-0.5707	-0.5650	0.4117	-1.4175	0.2179
high blood pressure	-0.0476	-0.0478	0.0192	-0.0849	-0.0097

Note: Binary variables coded as (1 = Yes, 0 = No); reference = absence of condition. Continuous predictors (e.g., SBP) reported per 10 mmHg increase. Posterior means and 95% credible intervals correspond to log-hazard ratios (log-HR). Log Pseudo Marginal Likelihood (LPML) = -233.456, Deviance Information Criterion (DIC) = 460.941, Watanabe-Akaike Information Criterion (WAIC) = 465.378, Number of Subjects = 125

Table 6: *Summary of posterior inference from the Bayesian Cox model*



Note: HR > 1 indicates higher hazard (risk); HR < 1 indicates protective effect. Covariates coded as 1 = Yes, 0 = No; SBP expressed per 10 mmHg increase.

Figure 7: *Forest plots of hazard ratios (HRs) and 95% confidence/credible intervals for the Bayesian Cox model.*

and smoking were consistently significant predictors of mortality. Parametric AFT models confirmed the same direction of effect with acceleration factors indicating shorter survival for these conditions. In contrast, higher systolic blood pressure demonstrated a modest protective effect against mortality.

Semi-parametric Cox proportional hazards and Bayesian Cox regression models identified hypertension, ischemic heart disease, and smoking as the strongest predictors of death, with hazard ratios consistently above unity. An increased systolic blood pressure, reported per 10 mmHg increase, remained a consistent protective factor in both frequentist and Bayesian analyses. Non-parametric Kaplan–Meier analysis revealed steadily declining survival probabilities

over 300 days, with no statistically significant differences between male and female patients. The random survival forest model achieved robust predictive performance and highlighted ischemic heart disease, smoking, and age as the most influential predictors.

Combining these complementary methods provides a comprehensive toolkit for survival analysis: parametric models for direct hazard acceleration inference, Cox and Bayesian Cox regression for robust estimation with minimal assumptions, Kaplan–Meier curves for non-parametric survival estimation, and random survival forests for high-accuracy prediction and variable importance assessment. This multi-methodological approach demonstrates the value of integrating diverse analytical techniques for comprehensive risk assessment in cardiac patient cohorts, offering enhanced potential for clinical risk stratification and personalized prognosis in resource-constrained settings.

Acknowledgments

The authors gratefully acknowledge the institutional support provided by PMAS-Arid Agriculture University, Rawalpindi. We extend our sincere appreciation to **Deputy Superintendent of Police (DSP) Naila Ashraf** of the Pakistan Railways Police for her authoritative support and facilitation during the data collection process. We are particularly grateful to **Dr. Ali Tahir Rana**, Director of Hospitals at Riphah International University and Riphah Healthcare Services, for his significant contributions to data collection. Special thanks are also due to **Ms. Khushboo Khurshid** for her meticulous work in data preparation and cleaning, as well as for her valuable encouragement during this research endeavor. The contributions of all individuals were instrumental in the successful completion of this study.

References

[1] Aalen, O. O., Borgan, Ø., and Gjessing, H. K. (2008). *Survival and Event History Analysis: A Process Point of View*. New York: Springer. doi: 10.1007/978-0-387-68560-1

[2] Ashine, T., Muleta, G., and Tadesse, K. (2021). Assessing survival time of heart failure patients: using Bayesian approach. Journal of Big Data, 8. doi: 10.1186/s40537-021-00537-4

[3] Biau, G., and Scornet, E. (2016). A random forest guided tour. TEST, 25, 197–227. doi: 10.1007/s11749-016-0481-7

[4] Brard, C., Le Teuff, G., Le Deley, M.-C., and Hampson, L. V. (2017). Bayesian survival analysis in clinical trials: What methods are used in practice? Clinical Trials, 14(1), 78–87. doi: 10.1177/1740774516673362

[5] Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* New York: Springer-Verlag. doi: 10.1007/b97636

[6] Collett, D. (2014). *Modelling Survival Data in Medical Research.* New York: Chapman and Hall/CRC. doi: 10.1201/b18041

[7] Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187–220. doi: 10.1111/j.2517-6161.1972.tb00899.x

[8] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. New York: Chapman and Hall/CRC. doi: 10.1201/b16018

[9] Greenwood, M. (1931). On the statistical measure of infectiousness. The Journal of Hygiene, 31(3), 336–351. doi: 10.1017/S002217240001086X

[10] Harrell, F. E., Jr. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis.* Springer Cham. doi: 10.1007/978-3-319-19425-7

[11] Heagerty, P. J., Lumley, T., and Pepe, M. S. (2021). Random survival forests for dynamic predictions of a time-to-event outcome. BMC Medical Research Methodology, 21. doi: 10.1186/s12874-021-01375-x

[12] Hosseinnataj, A., RezaBaneshi, M., and Bahrampour, A. (2020). Mortality risk factors in patients with gastric cancer using Bayesian and ordinary Lasso logistic models: a study in the

Southeast of Iran. Gastroenterology and Hepatology from Bed to Bench, 13(1), 31–36. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7069537/

[13] Hosmer, D. W., Lemeshow, S., and May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. Hoboken, New Jersey: John Wiley & Sons, Inc. doi: 10.1002/9780470258019

[14] Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. New York, NY: Springer. doi: 10.1007/978-1-4757-3447-8

[15] Ishwaran, H., and Kogalur, U. B. (2008). Random survival forests for R. R News, 7(2), 25–31. url: https://journal.r-project.org/articles/RN-2007-015/RN-2007-015.pdf

[16] Kaplan, E. L., and Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282), 457–481. doi: 10.1080/01621459.1958.10501452

[17] Klein, J. P., and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. New York, NY: Springer. doi: 10.1007/b97377

[18] Kleinbaum, D. G. and Klein, M. (2012). *Survival Analysis: A Self-Learning Text, Third Edition* New York, NY: Springer. doi: 10.1007/978-1-4419-6646-9

[19] LeBlanc, M., and Crowley, J. (1992). Relative risk trees for censored survival data. Biometrics, 48(2), 411–425. doi: 10.2307/2532300

[20] Omurlu, I. K., Ozdamar, K., and Ture, M. (2009). Comparison of Bayesian survival analysis and Cox regression analysis in simulated and breast cancer data sets. Expert Systems with Applications, 36(8), 11341–11346. doi: 10.1016/j.eswa.2009.03.058

[21] Penny-Dimri, J. C. , Bergmeir, C., Reid, C. M., Williams-Spence, J., Perry, L. A., and Smith, J. A. (2023). Tree-based survival analysis improves mortality prediction in cardiac surgery. Frontiers in Cardiovascular Medicine, 10, 1211600. doi: 10.3389/fcvm.2023.1211600

[22] Ponikowski, P., Voors, A. A., Anker, S. D., Bueno, H., Cleland, J. G., Coats, A. J., Falk, V., Gonzalez-Juanatey, J. R., Harjola, V. P., Jankowska, E. A., Jessup, M., Linde, C., Nihoyannopoulos, P., Parissis, J. T., Pieske, B., Riley, J. P., Rosano, G. M. C., Ruilope, L. M., Ruschitzka, F., Rutten, F. H., and van der Meer, P. (2016). ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. European Heart Journal, 37(27), 2129–2200. doi: 10.1093/eurheartj/ehw128

[23] Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics, 8. doi: 10.1186/1471-2105-8-25

[24] Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. New York, NY: Springer. doi: 10.1007/978-1-4757-3294-8

[25] Therneau, T. M. (2020). *A Package for Survival Analysis in R (version 3.2-11)*. url: https://cran.r-project.org/package=survival

[26] Thomas, C., Ye, F.Q., Irfanoglu, M.O., Modi, P., Saleem, K.S., Leopold, D.A., and Pierpaoli, C. (2014). Anatomical accuracy of brain connections derived from diffusion MRI tractography is inherently limited. Proc. Natl. Acad. Sci. U.S.A., 111(46), 16574–16579. doi: 10.1073/pnas.1405672111

[27] Van De Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., and Van Loey, N. E. (2015). Analyzing small data sets using Bayesian estimation: the case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. European Journal of Psychotraumatology, 6(1). doi: 10.3402/ejpt.v6.25216

[28] World Health Organization. (2021). Cardiovascular diseases (CVDs) – Fact sheet. World Health Organization. url: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[29] Wright, M. N., and Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, 77(1), 1–17. doi: 10.18637/jss.v077.i01

[30] Yancy, C. W., Jessup, M., Bozkurt, B., et al. (2013). 2013 ACCF/AHA guideline for the management of heart failure. Journal of the American College of Cardiology, 62(16). doi: 10.1016/j.jacc.2013.05.019