# Random Survival Forests Models: Review and Illustration

A project submitted in partial fulfillment of the requirements for the degree of BSc in Statistics and Statistics & Computer Science at the Faculty of Mathematical Sciences and Informatics, University of Khartoum


Mutaz Abdelrahman Osman Hassan

Najwa Ahmed Elrayah Hussein

Taif Manoful Abd Albagi Ahmed


**Supervisor:**

Ms. Sanaa Ahmed Hussein

January, 2025 - Shaban, 1446

## Abstract

The Random Survival Forest (RSF) model represents a transformative approach to survival analysis, addressing limitations in traditional methods like the Cox Proportional Hazards (CPH) model. This research investigates the RSF model's application, focusing on its advantages in handling complex, high-dimensional datasets and its ability to uncover non-linear relationships among variables. Through a structured study across four chapters, this project compares RSF with CPH, evaluates their performance using metrics like Brier scores and C-index, and highlights RSF's strengths in managing censored data and reducing prediction errors. The literature review showcased RSF's effectiveness in diverse medical applications, such as predicting the onset of type 2 diabetes and analyzing mortality risks in liver failure. A methodological analysis emphasized RSF's independence from hypothesis testing, its flexibility in variable selection, and its ability to produce more accurate and interpretable predictions compared to CPH. Additionally, this project incorporated a detailed comparison table summarizing findings from multiple studies, reinforcing RSF's superiority in handling variable interactions and enhancing predictive accuracy. While the study demonstrates RSF's potential, it acknowledges the need for future work to explore its application on real-world medical datasets, investigate non-linear variable effects further, and integrate explainable AI tools like SHAP for greater interpretability. This research concludes that RSF is a powerful tool for survival analysis, with promising implications for improving clinical decision-making and advancing predictive modeling in medical research.

## Acknowledgment

First and foremost, we express our deepest gratitude to *Allah Almighty* for His countless blessings, His guidance, and the strength He granted us to complete this research.

We would like to convey our sincere appreciation to our project supervisor, *Ms. Sanaa Ahmed*, whose invaluable guidance, continuous support, and insightful feedback have been instrumental in every stage of this work.

We are profoundly thankful to our parents, whose unwavering support, sincere prayers, and encouragement have been the cornerstone of our success throughout this journey.

We are grateful to our friends who stood by us and offered their support at every stage of this project. A special thanks goes to our colleagues *Yosra Adil* and *Ekram Elrayah*, whose constant encouragement, unwavering support, and motivating words have been a source of strength and inspiration for us.

Finally, we honor the memory of the martyrs we lost in the war, asking Allah to accept them and to encompass them with His boundless mercy.

We are truly grateful for every contribution that made this work possible.

# Contents

## List of Figures

## List of Tables

# CHAPTER 1

# INTRODUCTION

## 1.1 Background

Survival analysis is a statistical approach used to predict the time until an event of interest, such as failure or death occurs. It is widely applied in various fields such as medicine, engineering, and finance. Traditional survival models, like the Cox proportional hazards model, assume a linear relationship between covariates and the hazard function, which may not always hold. To address non-linearity and interactions between covariates, Random Survival Forests (RSF) have emerged as a robust and flexible alternative. RSF models are an extension of traditional Random Forests designed to handle survival data. These models provide flexibility in dealing with non-linear and complex data by aggregating many random survival trees, each constructed using a sample of the data and a random subset of covariates. This enhances the model's ability to handle variability in the data.

## 1.2 Problem Statement

The aim of this project is to highlight the importance of Random Survival Forests (RSF) in the field of survival data analysis. Traditional survival models often struggle with non-linearity and interactions between covariates, limiting their predictive accuracy. RSF, an ensemble learning technique, addresses these challenges effectively by leveraging multiple decision trees to model complex relationships in survival data. This project seeks to demonstrate the superiority of RSF in providing accurate predictions and valuable insights into variable importance for survival outcomes. By applying RSF to real-world survival data, we aim to showcase its potential and practical applications in medicine field.

## 1.3 Objective

This project aims to utilize the Random Survival Forest (RSF) algorithm to accurately predict survival times for patients diagnosed with specific diseases, RSF, known for its ability to handle non-linear relationships and interactions between variables, will be employed to enhance survival analysis in complex datasets with censored data. Additionally, this study intends to compare RSF

with the Cox proportional hazards (CPH) model to assess their respective strengths and weaknesses in survival analysis applications.

### 1.4 Research Questions

1. How can the Random Survival Forest (RSF) algorithm be used to predict the survival times of patients diagnosed with specific diseases?
2. How accurate is the RSF algorithm in handling non-linear relationships and interactions between variables in survival analysis?
3. How does the RSF algorithm contribute to the analysis of complex survival datasets with censored data?
4. What are the practical benefits of applying the RSF algorithm in the medical field, particularly in the development of personalized treatment plans for patients?
5. Compare the RSF algorithm with the CPH model in terms of their effectiveness and differences in survival analysis.

### 1.5 Research Significance

Random Survival Forest (RSF) represents a significant advancement in the field of survival analysis, especially in handling complex datasets with censoring. RSF excels in effectively handling incomplete data by integrating these incomplete observations into the analysis to provide robust predictions. Moreover, RSF accurately captures non-linear relationships and interactions between variables without the strict assumptions imposed by traditional models like the Cox proportional hazards (CPH) model. RSF is also efficient in managing high-dimensional data, making it suitable for modern applications involving large sets of predictors such as genetic information and clinical measurements. Additionally, RSF provides measures of variable importance, aiding in identifying the most influential factors in the data and improving medical treatment strategies and interventions.

### 1.6 Research Structure

This research consists of the following chapters:

- Chapter 1: Introduction includes research background on survival models, problem statement, objective, research questions, research significance, research structure and terminologies.

- Chapter 2: Literature Review - Analysis of current research on RSF and Random Forests.
- Chapter 3: Methodology - Comparison between CPH Models and RSF - Detailed study of differences and similarities.
- Chapter 4: Summary and Future Work.

### 1.7 Terminologies

Random Survival Forests (RSF), Cox Proportional Hazards (CPH), Survival Analysis, Predictive Modeling, Non-linear Relationships, High-dimensional Data, C-index, Brier Score, Variable Importance (VIMP), Censored Data, Medical Applications, Explainable AI (SHAP), Clinical Decision-making, Risk Prediction.

# CHAPTER 2

# LITERATURE REVIEW

Random Forest Survival models are an extension of the Random Forest algorithm, designed specifically for survival data, which deals with time-to-event data. The basic concept revolves around constructing multiple decision trees based on bootstrapped samples and combining their predictions to improve accuracy and robustness. Key Studies and Contributions: The initial development of Random Survival Forests (RSF) was introduced by Breiman (2001) and extended by Ishwaran et al. (2008), who formalized the approach for survival analysis. Ishwaran's work provided a solid foundation for using Random Forests in survival analysis, showcasing their effectiveness in handling high-dimensional data, censoring, and variable selection.

Numerous studies have applied RSF across various fields, from medical research to reliability engineering. For instance, RSF has been compared with traditional Cox proportional hazards models, often showing superior performance, especially in non-linear or complex datasets. Advantages' handles censoring effectively, accommodates high- dimensional data, and does not require proportional hazards assumptions. Interpretation can be challenging due to the complexity of the model. Additionally, it may require substantial computational resources compared to simpler models like Cox regression. Recent research has focused on improving the interpretability and computational efficiency of RSF, exploring hybrid models, and integrating deep learning techniques for survival analysis.

In the field of healthcare, predictive models have traditionally relied on statistical methods like logistic regression and the Cox proportional hazards model. These approaches have been instrumental in providing valuable insights, but they come with significant limitations. One key drawback is their struggle to effectively capture non-linear relationships and complex interactions between variables. Additionally, these models often rely on assumptions about the distribution of survival times, which may not be valid in the diverse and complex datasets encountered in real-world clinical settings. As a result, the accuracy and relevance of these models can be compromised when applied to intricate patient data, highlighting the need for more advanced methods that can

better handle such complexities. In this paper [1], the author applied RSF, and this application results in a significant advantage of RSF models lies in their use of ensemble learning techniques such as bagging and boosting. These methods enhance the stability and accuracy of predictions by combining the outputs of multiple models, thereby reducing variance and improving robustness. The literature also highlights the importance of hyperparameter tuning in maximizing the performance of RSF models. Techniques such as grid search enable the finetuning of model parameters, leading to significant improvements in predictive accuracy and model reliability. The integration of RSF models into clinical practice offers substantial potential for improving decision support systems. By providing more accurate and individualized predictions, RSF models can help clinicians make better-informed decisions, ultimately leading to improved patient outcomes. in liver disease, RSF models can guide therapeutic strategies by predicting survival probabilities and identifying critical prognostic factors [1].

Another application was done in the field of predictive modeling in healthcare, such as predicting time to diabetes diagnosis, focusing on the application of Random Survival Forest (RSF) models, particularly in the context of type 2 diabetes. Type 2 diabetes is a rapidly escalating public health concern, with increasing incidence rates worldwide. This condition, often associated with severe complications such as cardiovascular disease, kidney failure, and neuropathy, underscores the urgent need for timely and precise diagnostic tools. Early detection and management are crucial in mitigating the adverse outcomes associated with diabetes, necessitating models that can accurately predict the onset and progression of the disease. In this paper [Predicting time to diabetes diagnosis using random survival forests], the author highlighted the RSF model importance to be, an extension of the Random Forest algorithm, has emerged as a powerful alternative for survival analysis. RSF models excel in handling high-dimensional and complex datasets, offering flexibility in capturing non-linear relationships and interactions between variables without the rigid assumptions imposed by classical models. RSF's ability to aggregate multiple decision trees, each constructed using a random subset of data and covariates, allows it to manage variability effectively and improve predictive accuracy. This ensemble learning technique not only enhances model robustness but also provides valuable insights into the relative importance of different variables in predicting outcomes, which is particularly beneficial in a clinical context. The result of the study showed that in the context of type 2 diabetes, the inclusion of comorbid conditions such as hypertension, depression, and obesity in predictive models is

critical. These factors significantly influence the onset and progression of diabetes, and their interaction with other variables can be complex. RSF models are well-suited to incorporate these comorbidities, offering a more holistic and individualized prediction of disease progression. By leveraging electronic medical records (EMR) and integrating a wide array of biomarkers, RSF models can provide clinicians with tailored predictions that support more informed decision-making and personalized treatment plans [2].

On the other hand, oncologist used to address the problem of Nasopharyngeal carcinoma (NPC), which is a distinct type of squamous cell carcinoma originating from the nasopharynx. It is characterized by its unique epidemiology, particularly its prevalence in Southeast Asia, and by the challenges associated with its diagnosis and treatment. Due to its anatomical location, NPC often presents at an advanced stage, leading to poor prognosis for many patients. Early detection is challenging, as the cancer may not cause noticeable symptoms until it has progressed significantly. The integration of machine learning techniques such as RSF and Survival-SVM into survival analysis represents a significant advancement in the prognostic modeling of NPC. These models offer greater accuracy than traditional Cox regression, providing clinicians with more reliable tools for predicting patient outcomes and tailoring treatment strategies. As machine learning continues to evolve, its application in oncology promises to enhance patient care through more precise and personalized prognostic assessments. The RSF and Survival Support Vector Machine (Survival-SVM) are two such methods that have shown promise in improving survival predictions. RSF is an ensemble method that extends the traditional random forest approach to survival analysis. It can accommodate complex interactions between variables and does not require the proportional hazards assumption. This model generates multiple survival trees and aggregates their predictions, thus offering a robust approach to survival prediction in NPC. The RSF also allows for the ranking of variable importance using the Variable Importance (VIMP) method, providing insights into which factors most significantly impact patient outcomes. Survival-SVM is a machine learning method adapted for survival analysis. It is particularly advantageous in handling small sample sizes and complex, high-dimensional data. Survival-SVM models the survival time directly, offering a more flexible and potentially more accurate approach than the Cox model. Despite the promising results, the study acknowledges several limitations. The data used were primarily from the SEER database, which mainly includes patients from Western populations. This raises concerns about the generalizability of the findings to other populations,

particularly in Asia, where NPC is more prevalent. The authors suggest that future research should focus on validating these models using more diverse datasets, including those from Chinese populations, to ensure their applicability across different demographic groups. Furthermore, while machine learning models like RSF and Survival SVM offer significant improvements over traditional methods, their complexity and the computational resources required for their implementation may limit their widespread use in clinical settings. Future work should aim to streamline these models and make them more accessible for routine clinical use [3].

Random Survival Forests (RSF) has recently showed better performance than statistical survival methods as Cox proportional hazard (CPH) in predicting conversion risk from mild cognitive impairment (MCI) to Alzheimer's disease (AD). However, RSF application in real world clinical setting is still limited due to its black-box nature. For this reason, Sarica et al. aimed to providing a comprehensive study of RSF explainability with Shapley Additive Explanations (SHAP) on biomarkers of stable and progressive patients (sMCI and pMCI) from Alzheimer's Disease Neuroimaging Initiative. They evaluated three global explanations—RSF feature importance, permutation importance and SHAP importance—and they quantitatively compared them with Rank-Biased Overlap (RBO). Moreover, they assessed whether multicollinearity among variables may perturb SHAP outcome. Lastly, stratified pMCI test patients in high, medium and low risk grade, to investigate individual SHAP explanation of one pMCI patient per risk group. They confirmed that RSF had higher accuracy (0.890) than CPH (0.819), and its stability and robustness was demonstrated by high overlap (RBO>90%) between feature rankings within first eight features. SHAP local explanations with and without correlated variables had no substantial difference, showing that multicollinearity did not alter the model. FDG, ABETA42 and HCI were the first important features in global explanations, with the highest contribution also in local explanation. FAQ, mPACCdigit, mPACCtrailsB and RAVLT immediate had the highest influence among all clinical and neuropsychological assessments in increasing progression risk, as particularly evident in pMCI patients' individual explanation. In conclusion, our findings suggest that RSF represents a useful tool to support clinicians in estimating conversion-to-AD risk and that SHAP explainer boosts its clinical utility with intelligible and interpretable individual outcomes that highlights key features associated with AD prognosis [4].

Compared to squamous cell carcinoma, head and neck non-squamous cell carcinoma (HNnSCC) is rarer. Integrated survival prediction tools are lacking. Methods: 4458 patients of HNnSCC were collected from the SEER database. The endpoints were overall survivals (OSs) and disease-specific survivals (DSSs) of 3 and 5 years. Cases were stratified–randomly divided into the train & validation (70%) and test cohorts (30%). Tenfold cross validation was used in establishment of the model. The performance was evaluated with the test cohort by the receiver operating characteristic, calibration, and decision curves. Results: The prognostic factors found with multivariate analyses were used to establish the prediction model. The area under the curve (AUC) is 0.866 (95%CI: 0.844–0.888) for 3-year OS, 0.862 (95%CI: 0.842–0.882) for 5-year OS, 0.902 (95%CI: 0.888–0.916) for 3-year DSS, and 0.903 (95%CI: 0.881–0.925) for 5-year DSS. The net benefit of this model is greater than that of the traditional prediction methods. Among predictors, pathology, involved cervical nodes level, and tumor size are found contributing the most variance to the prediction. The model was then deployed online for easy use. Conclusions: The present study incorporated the clinical, pathological, and therapeutic features comprehensively and established a clinically effective survival prediction model for post-treatment HNnSCC patients [5].

Xu and Wang, in their paper addressed that it's essential to predict the survival status of patients based on their prognosis. This can assist physicians in evaluating treatment decisions. Random forest is an excellent machine learning algorithm even without any modification. They propose a new random forest weighting method and apply it to the gastric cancer patient data from the Surveillance, Epidemiology, and End Results (SEER) program. They evaluated the generalization ability of this weighted random forest algorithm on 10 public medical datasets. Furthermore, for the same weighting mode, the difference between using out-of-bag (OOB) data and all training sets as the weighting basis is explored. 110 697 cases of gastric cancer patients diagnosed between 1975 and 2016 obtained from the SEER database were included in the experiment. In addition, 10 public medical datasets were used for the generalization ability evaluation of this weighted random forest algorithm. Through experimental verification, on the SEER gastric cancer patient data, the weighted random forest algorithm improves the accuracy by 0.79% compared with the original random forest. In AUC, macro-averaging increased by 2.32% and micro-averaging increased by 0.51% on average. Among the 10 public datasets, the random forest weighted in accuracy has the best performance on 6 datasets, with an average increase of

1.44% in accuracy and an average increase of 1.2% in AUC. Compared with the original random forest, the weighted random forest model shows a significant improvement in performance, and the effect of using all training data as the weighting basis is better than using OOB data [6].

In a study by Zhang et al. published in Clinical Epidemiology, the objective was to improve the prediction of dyslipidemia. Data were collected from 6,328 participants aged between 19 and 90 years, who were followed for up to 5 years. The researchers used the Random Survival Forest (RSF) technique to develop predictive models and compared its results with the Cox Regression model. The study found that the incidence rate of dyslipidemia was 101 cases per 1,000 individuals, with 121 cases per 1,000 in men and 69 cases per 1,000 in women. RSF model results showed that Harrell's Concordance C-statistics were 0.731 for males and 0.801 for females, indicating that the model provides more accurate predictions compared to the Cox model, which had values of 0.719 for males and 0.787 for females. The RSF model predicted that key influencing variables include LDL-C, TC, TG, HDL-C, and BMI for men, and LDL-C, TC, TG, HDL-C, age, and FBG for women. RSF outperformed the Cox model, highlighting its effectiveness in predicting dyslipidemia. Despite limitations related to the socioeconomic representation of the participants, the study confirmed that RSF is a powerful tool that can be used in routine health screenings to identify individuals at risk and guide strategies for dyslipidemia prevention [7].

Another important application was done in a research related to chronic hepatitis B poses a significant threat to liver health globally. Acute-on-chronic liver failure (ACLF) is a critical condition characterized by the sudden deterioration of liver function in patients with pre-existing chronic liver disease, often triggered by various hepatic insults. Due to compromised liver function and the likelihood of multiple organ failure, 60% to 70% of ACLF patients face rapid disease progression and a high risk of death within three months. This alarming statistic underscores the urgent need for a reliable prognostic model to identify ACLF patients at high risk of mortality who may require immediate liver transplantation. In response to this need, the present study aimed to develop an online tool that can predict individual mortality risk for ACLF patients. This tool was built using the Random Survival Forest (RSF) algorithm, which offers the ability to predict mortality risk on a per-patient basis. The study involved 391 ACLF patients from three hospitals in China, with the final follow-up conducted on September 10, 2018. After applying specific inclusion criteria, 276 patients were selected for the final survival analysis and model development.

The RSF model used in this study demonstrated exceptional predictive accuracy. The Brier score, which measures the accuracy of probabilistic predictions, indicated the RSF model's superior performance compared to other models. Specifically, the RSF model achieved an area under the time-dependent ROC curve (AUROC) of 0.916 for both 3-month and 6-month mortality predictions, and 0.905 for 12-month mortality predictions. This high level of accuracy was statistically significant, with a P value of less than 0.05. Further analysis showed that patients with higher RSF scores had significantly poorer survival outcomes than those with lower scores. The predicted mortality rates closely mirrored the actual outcomes, confirming the model's reliability. Additionally, when compared to other models such as the Cox proportional hazards model, the Model for End Stage Liver Disease (MELD), the Age-Bilirubin-INR-Creatinine (ABIC) score, and the integrated MELD (iMELD), the RSF model consistently outperformed them across all evaluated time points. In terms of Brier scores, the RSF model achieved results of 0.119 for both 3-month and 6 month mortality predictions, and 0.128 for 12-month predictions, demonstrating its accuracy over other models. Decision tree analysis also supported the RSF model's superiority in forecasting patient outcomes at various time points. In summary, the RSF model excelled in predicting mortality at 3, 6, and 12 months, showing not only high accuracy but also consistency with actual patient outcomes. The model's effectiveness was highlighted through time-dependent ROC curve analysis, decision tree evaluation, and Brier score comparisons, all of which confirmed the RSF model's edge over the Cox model in prognosticating ACLF patient outcomes. Moreover, the study identified several independent risk factors for ACLF patients through multivariate Cox regression analysis, including hepatic encephalopathy (HE), age, serum sodium levels, acute kidney injury (AKI), red cell distribution width (RDW), and international normalized ratio (INR). The RSF algorithm validated these findings, underscoring the importance of these variables in predicting patient prognosis. The RSF model also proved adept at identifying variables with nonlinear effects on patient outcomes, further affirming its superiority over the Cox model. The AUROC scores for the RSF model were 0.916, 0.916, and 0.905, compared to 0.872, 0.866, and 0.848 for the Cox model. Similarly, the RSF model achieved better Brier scores of 0.119, 0.119, and 0.128, compared to 0.138, 0.146, and 0.156 for the Cox model, demonstrating its superior performance in mortality prediction for ACLF patients [8].

In a study conducted on patients diagnosed with pulmonary hypertension (PHTN) between 2000 and 2017 in Hong Kong, clinical and pharmaceutical data from 2,560 patients were analyzed

using both CPH models and the RSF model to predict mortality and complications. The RSF model achieved an Area Under the Curve (AUC) of 0.79, which was higher than the 0.71 AUC achieved by the Cox model, highlighting its superior predictive accuracy and ability to capture complex interactions between variables. The importance of variables was evaluated using two methods: the variable importance value, which compared prediction errors before and after randomizing the variables, and the minimum depth approach, which ranked variables based on their proximity to the root node of the decision tree. During the follow-up period, 38% of the patients died, with the deceased having a median age of 75.2 years, compared to 44.6 years for the survivors. Multivariate Cox analysis identified key factors predicting mortality, such as age at diagnosis (hazard ratio 1.822), cumulative hospital stays (1.0007), cardiovascular diseases (1.266), kidney diseases (1.279), diabetes (1.208), and hypertension (1.549). Pulmonary hypertension is defined as a mean pulmonary arterial pressure greater than 25 mmHg at rest, measured using right heart catheterization. The condition is traditionally classified into primary and secondary types and is progressive, potentially leading to severe complications and death if untreated. Guidelines from the European Society of Cardiology and the European Respiratory Society emphasize the importance of dynamic risk stratification to predict outcomes and guide treatment strategies. Risk factors such as acute right heart failure, chronic kidney disease, diabetes, and metabolic syndrome further exacerbate the disease. This study was the first to apply the Electronic Frailty Index (eFI) to PHTN, demonstrating the superiority of the RSF model over traditional methods. However, the study had limitations, including being confined to Hong Kong, with clinical data constraints due to the limitations of the CDARS database, and the inability to classify PHTN according to WHO criteria [9].

# CHAPTER 3

# METHODOLOGY

This chapter outlines the methodology adopted in this study, which aims to explore and illustrate the application of Random Survival Forests (RSF) models in survival analysis. We begin with a comprehensive review of the theoretical foundations of these models, focusing on their key concepts and how they function, before moving to a practical application to evaluate their effectiveness in predicting survival outcomes. We also include a detailed comparison between the RSF model and the Cox Proportional Hazards (CPH) model, based on an analysis of selected research studies. These studies have been summarized, and the results are presented in a table to highlight the key points of comparison, including the performance of the models in different contexts. This comparison emphasizes the practical differences between the two models and their roles in handling complex, high-dimensional data, providing deeper insight into the applications of each model.

## 3.1 Data Sources and Selection Criteria

In this project, research papers were selected based on specific criteria to ensure the quality and reliability of the analysis. The focus was on studies that utilize the RSF model and the CPH model in medical applications. These studies were chosen for their ability to provide accurate and comprehensive analyses of survival and prognosis in medical contexts such as cancer and diabetes. When selecting the studies, we assessed the quality of the data presented by reviewing the accuracy of the analytical methods used and the reliability of the results. We also ensured the inclusion of studies that cover diverse medical populations to guarantee that the models applied are relevant to a wide range of medical cases. This diversity helps achieve a comprehensive understanding of the models' effectiveness in various medical contexts. The analytical methods used in the studies were reviewed to ensure their accuracy.

## 3.2 Cox Proportional Hazards Model

### 3.2.1 Overview

The Cox Proportional Hazards (CPH) model, developed by David Cox in 1972, is a statistical tool commonly used in medical research and survival analysis. It assumes that the relative hazard rates remain constant over time, making it suitable for analyzing the timing of events. The model does not require the specification of a baseline hazard function, enhancing its flexibility across various scenarios. It can accommodate both continuous and categorical predictor variables, enabling researchers to evaluate the impact of multiple factors on survival outcomes. The model provides hazard ratios that help interpret the effects of different variables on survival, making it a popular choice in medical research [10].

### 3.2.2 Definition

The Cox model is a regression technique for performing survival analyses in epidemiological and clinical research. This model estimates the hazard ratio (HR) of a given endpoint associated with a specific risk factor, which can be either a continuous variable like age and C-reactive protein level or a categorical variable like gender and diabetes mellitus. When the risk factor is a continuous variable, the Cox model provides the HR of the study endpoint associated with a predefined unit of increase in the independent variable (e.g., for every 1-year increase in age, 2 mg/L increase in C-reactive protein). A fundamental assumption underlying the application of the Cox model is proportional hazards; in other words, the effects of different variables on survival are constant over time and additive over a particular scale. The Cox regression model, when applied to etiological studies, also allows an adjustment for potential confounders; in an exposure-outcome pathway, a confounder is a variable which is associated with the exposure, is not an effect of the exposure, does not lie in the causal pathway between the exposure and the outcome, and represents a risk factor for the outcome [11].

### 3.2.3 Formula

$$H(t) = H_0(t) \times exp[b_1x_1 + b_2x_2 + \cdots + b_kx_k]$$

The CPH model expresses the hazard at a given time for an individual with a set of explanatory variables (X) through a specific formula. This formula represents the hazard as the product of two quantities: the first is the baseline hazard function, and the second is the exponential

expression. An important feature of this formula is the Proportional Hazards Assumption, where the baseline hazard depends on time but does not include the explanatory variables, while the exponential expression depends on the explanatory variables but does not involve time. This allows us to understand the impact of the explanatory variables on the hazard without a direct influence from time. If the hazard ratios vary over time, this assumption cannot be met, making the CPH model inappropriate [12].

### 3.2.4 Assumptions

1) Proportional Hazards Assumption: The hazard ratios between individuals should remain constant over time.
2) Independence Assumption: Events should be independent, meaning one event does not affect another.
3) Non-informative Censoring Assumption: Censoring should not be related to the time of the event or the covariates.
4) Linearity Assumption: The relationship between covariates and the hazard function should be linear.
5) Absence of Multicollinearity: Covariates should not be highly correlated to avoid difficulties in interpreting effects [10].

### 3.2.5 Advantages

The following are some advantages of the CPH model [16]:

1) Flexibility: The model handles both continuous and categorical variables, which allows it to be applied to a broad range of risk analysis scenarios.
2) Interpretability: The model offers estimated coefficients that help explain how covariates impact risk.
3) Handling Censored Data: The model can handle right-censored data, providing more accurate analysis when events have not occurred for all subjects.
4) Wide Applicability: The model can be used across various fields, such as medicine, biology, and social sciences, making it a versatile tool for survival data analysis.

### 3.2.6 Example

For example, if you want to analyze the survival time after the detection of a disease, you are often not interested in the survival time itself, but in what influences the survival time. So, we want to know if the survival time depends on one or more factors, called "predictors" or "independent variables" For simple situations with a single factor with only two values, the Log Rank Test is used. For example, if you want to test whether there is a difference in survival time when two different drugs are given [13].

If you want to include the age of the subjects, a special type of regression is needed. This is the Proportional Hazards Survival Regression. This regression is then used to evaluate the effect of each predictor on the shape of the survival curve [13].

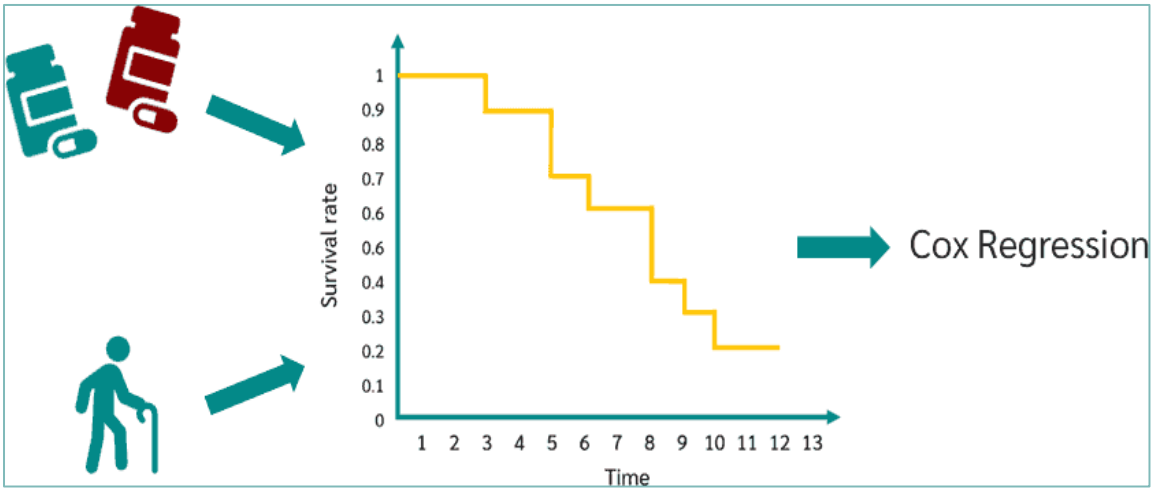


Figure 3.2.6.1: Survival curve with cox regression Model [13]

In our example, we have as predictors, on the one hand, the drug used and, on the other hand, the age of the subjects. We want to know what effect these variables have on the survival time curve. To do this, we use Cox regression. We will now look at the steps of Cox regression using an example. Let's assume we have the following data and want to analyze it [13].

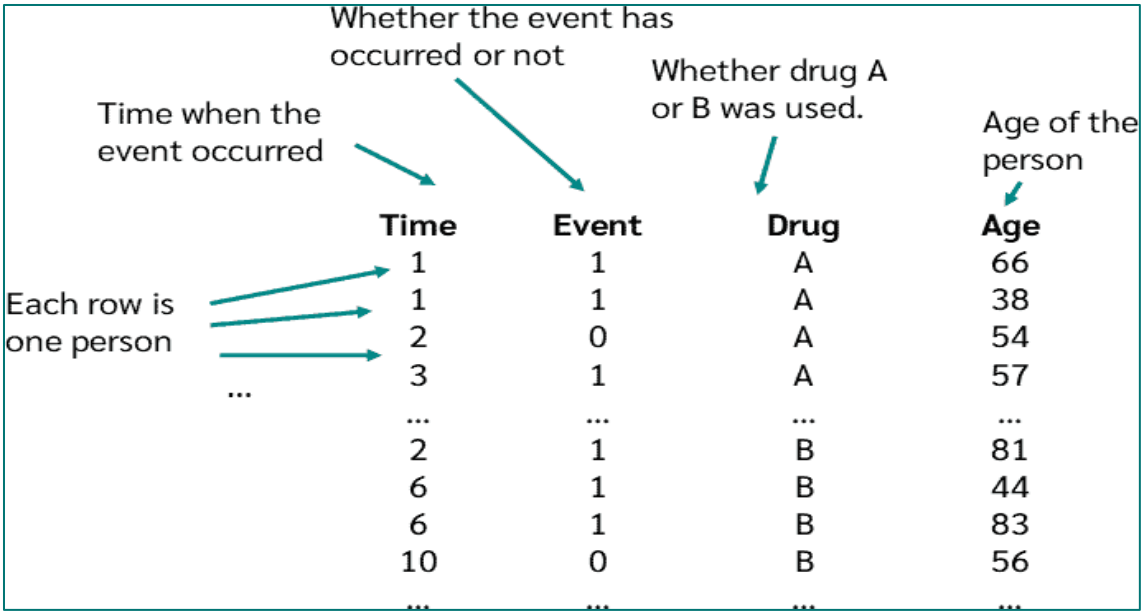


Figure 3.2.6.2: Cox regression data table [13]

Each row describes a patient with the corresponding disease. The time indicates when the event or death occurred. Of course, we also have information about which drug was used and the age of the subjects [13].

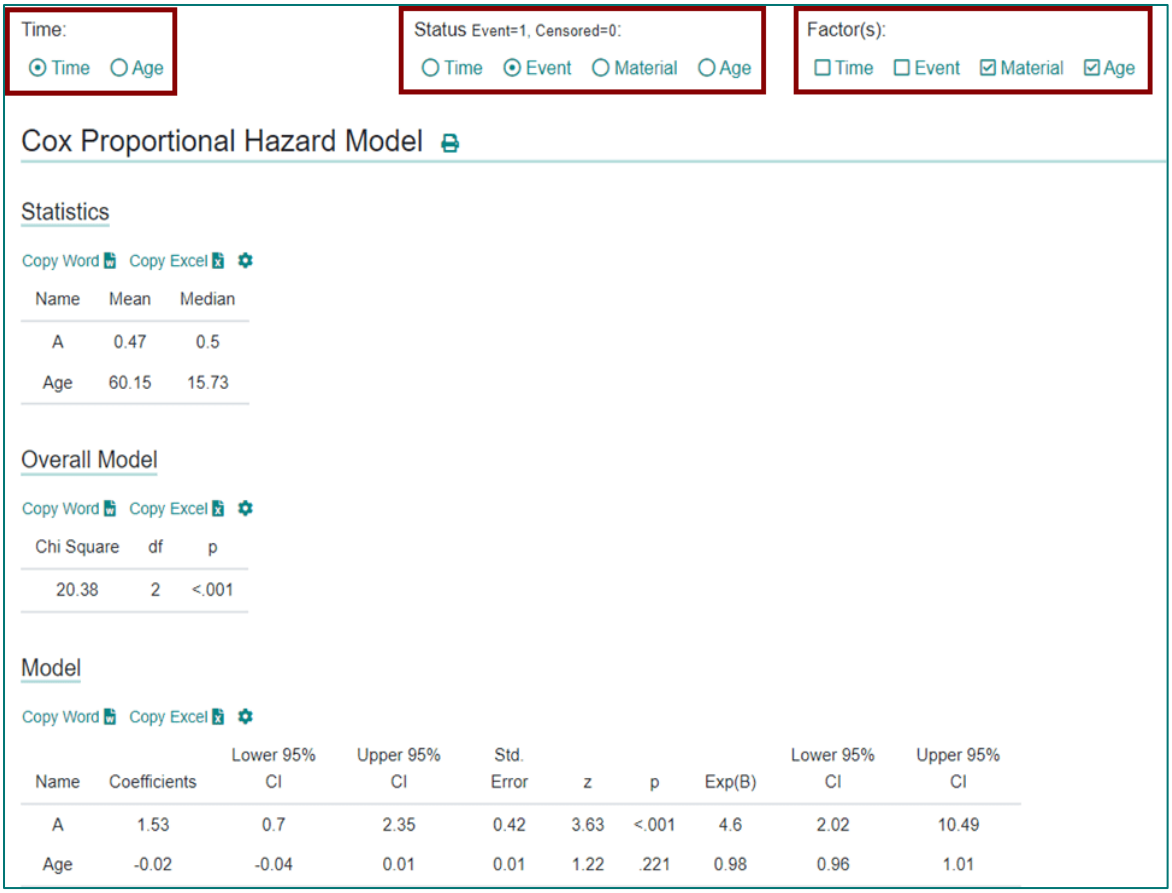


Figure 3.2.6.3: Cox model data setup [13]

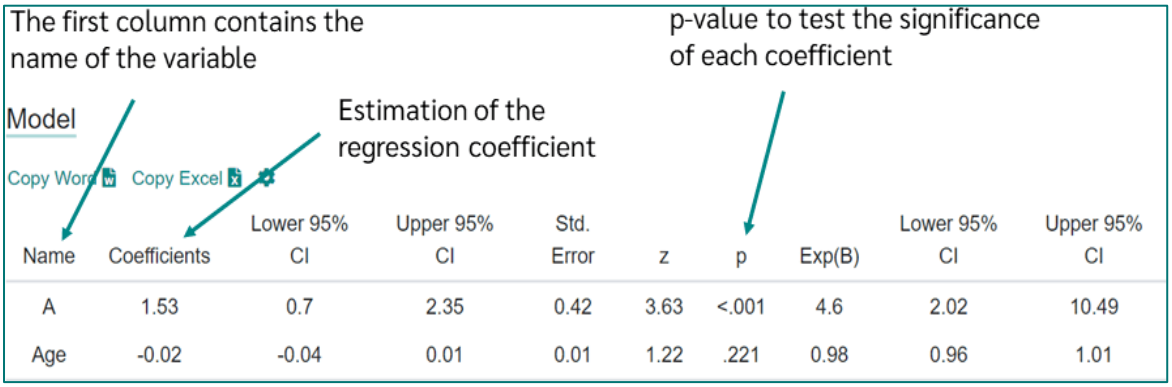


Figure 3.2.6.4: Cox model analysis results [13]

The first column contains the names of the variables. The first row shows the variable drug and the second row shows the age of the persons.

The most important values in this table are the estimated regression coefficient and the p-value. The p-value tells you whether the regression coefficient is significantly different from zero.

So, the null hypothesis is that the coefficient is zero in the population. Assuming, as usual, that the significance level is set at 5%, the null hypothesis is rejected for p-values less than 5% or 0.05. This means that the coefficient is significantly different from zero.

In the case of drug, the p-value is less than 0.05 and therefore there is a significant difference from zero.

In the case of age, we obtain a p-value of 0.221, which is greater than 0.05. Therefore, in this case, the null hypothesis is neither rejected nor accepted and we assume, based on these data, that age does not have a significant effect on the survival curve [13].

## 3.3 Random Survival Forest Model

In recent years, there has been a growing interest in applying statistical machine learning techniques in survival analysis. Ensemble methods, such as Random Survival Forests (RSF), have been developed across various fields, including medical sciences, due to their high accuracy, non-parametric nature, and ability to handle high-dimensional datasets [14].

### 3.3.1 Concept of RSF Method

The RSF method is an extension of the Random Forest (RF) method for survival data analysis. This method relies on randomness in two phases: first, a bootstrap sample is taken for each tree, and second, a subset of explanatory variables is randomly selected at each node. This randomness helps reduce variance, thereby improving model performance [14].

### 3.3.2 Steps to Implement RSF

The steps to implement RSF include:

1) Bootstrap Sample Extraction: B bootstrap samples are extracted from the dataset, with each sample containing about two-thirds of the data. Each bootstrap sample is used to calculate a separate decision tree, reducing the problem of overfitting with other data.
2) Variable Selection: A random number of explanatory variables is selected when splitting the nodes in the decision tree. In each node, a predetermined number of variables are considered, and the variable that causes the greatest difference between the resulting daughter nodes is selected to split the node.

3) Decision Tree Creation: A decision tree is created for each bootstrap sample, where a random set of explanatory variables is selected at each node. The goal of splitting the nodes is to maximize the survival difference using log-rank statistics.
4) Calculating Ensemble Survival Forest: The survival forest is computed by calculating the cumulative hazard function (CHF) for each tree using the Nelson-Aalen estimator and then averaging these functions for each tree to obtain the ensemble CHF.
5) Model Performance: Model performance is evaluated using the out-of-bag (OOB) error rate and Brier score. Additionally, the C-index can be calculated to assess prediction accuracy [14].

**The following figure illustrates these steps:**

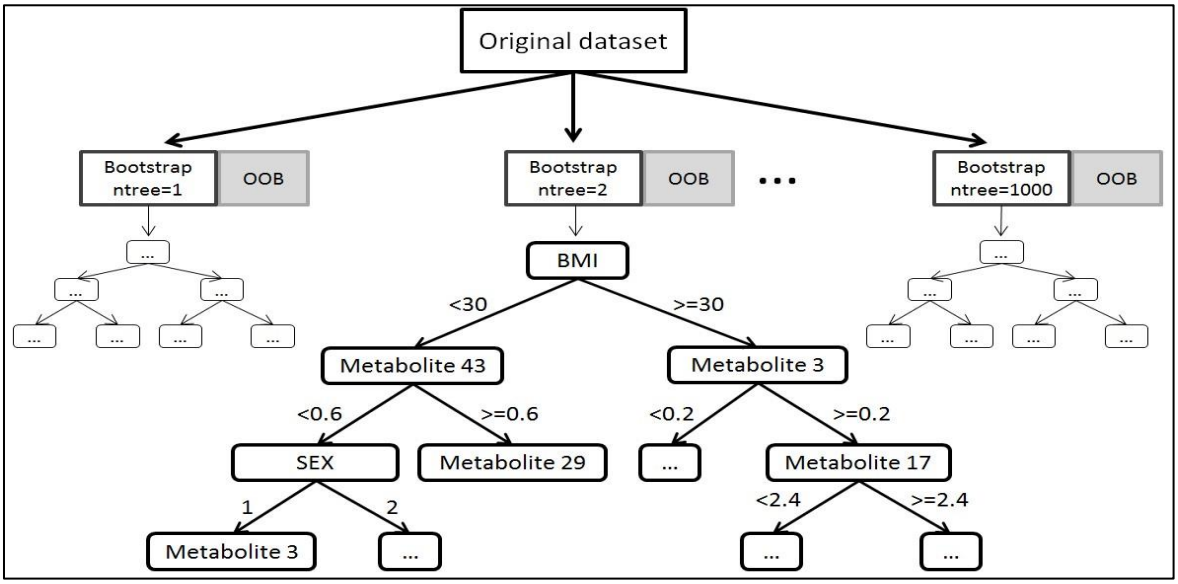


Figure 3.3.2.1: Steps to implement RSF [14]

### 3.3.3 Machine Learning Techniques

Machine learning (ML) techniques have recently been applied to analyze large amounts of data. ML relies on learning algorithms from repeated input data, discovering complex patterns and relationships. Tree-based methods are part of ML, where variables are split in each tree using specified splitting rules [14].

### 3.3.4 Reducing Variance and Improving Predictions

Each tree in this framework is non-deterministic, leading to significant variance in results. However, by aggregating trees, this variance can be reduced, enhancing prediction

accuracy. Key ensemble techniques include bagging, boosting, and random forests, each relying on multiple bootstrap samples to improve model performance [14].

### 3.3.5 Advantages of RSF Method

The RSF method has several advantages compared to regression methods, including [14]:

- Independence from Hypothesis Testing: RSF relies entirely on data, making it independent of hypothesis testing and not requiring goodness-of-fit testing or variable distribution.
- Reduced Competition Among Variables: RSF uses a random subset of variables when growing the tree, reducing competition among interrelated variables.
- Model Evaluation: The OOB error rate and C-index can be calculated without requiring a separate validation dataset.

### 3.3.6 Disadvantages of RSF Method

A drawback of the RSF method is its inability to compute hazard ratios and odds ratios. However, risk factor importance can be measured using minimal and VIMP indicators. Additionally, tree-based methods tend to prefer continuous variable selection when splitting nodes, requiring fewer cut points in the presence of mixed variables [14].

### 3.3.7 Review of RSF Method

The RSF technique is used to build a model based on aggregating multiple decision trees. A decision tree is created for each bootstrap sample, where a random set of explanatory variables is selected at each node. The number of random split points for each variable can be determined before fitting the model using the Random Forest SRC package in R [14].

### 3.3.8 RSF Algorithm Steps

1) Extract B bootstrap samples from the dataset.
2) Grow a survival tree for each bootstrap sample, where p random variables are chosen at each node.
3) Each terminal node should contain fewer than d unique outcomes.
4) Calculate the cumulative hazard function (CHF) for each tree and use the average to obtain the aggregated CHF.

5) Use OOB data to calculate the expected CHF prediction error.
6) Estimate variable importance (VIMP) using OOB data [14].

### 3.3.9 C-index Calculation

1) Obtain all possible pairs of cases.
2) Remove pairs with tied times.
3) Evaluate allowed pairs based on survival time.
4) Calculate the C-index based on the number of concordant pairs.

Prediction Error Rate: The prediction error rate is calculated using the C-index as follows [14]:

$$\{\boldsymbol{Prediction\ Error\ Rate}\} = \mathbf{1} - \mathbf{C\text{-}index}$$

This reflects the accuracy of the model's predictions.

### 3.4 Comparison of Cox Proportional Hazards and Random Survival Forests

| Aspect | Cox Proportional Hazards (CPH) | Random Survival Forests (RSF) |
|---|---|---|
| **Model Type** | Semi-parametric | Fully non-parametric |
| **Hazard Assumption** | Assumes proportional hazards over time | Does not assume proportional hazards |
| **Explanatory Variables** | Linear relationship between covariates and hazards | No assumptions on linearity or relationships |
| **Baseline Hazard** | Non-parametric | Based on aggregating multiple decision trees |
| **Predictive Accuracy** | Performs well when proportional hazards assumption holds | High predictive accuracy, especially with complex, multi-dimensional data |
| **Computational Efficiency** | Less computational time needed | Requires more time to train, especially with large datasets |
| **Concordance Index (c-index)** | Consistent performance | Often outperforms CPH in long-term predictions |
| **Prediction Error Curves** | Shows small differences in predictive errors | More accurate predictions compared to other models |

Table 3.4.1: Comparison of cox proportional hazards and random survival forests [15]

| Factors / Papers | Accuracy | Brier Score* | handling of variables | Prediction error- Error rates and other evaluation metrics | Variable Importance (VIMP) | Non-Linear Relationship ** | conclusion |
|---|---|---|---|---|---|---|---|
| **1.Risk Prediction of Dyslipidemia for Chinese Han Adults using Random Forest Survival Model [1]** | RSF recorded an AUC of 0.731 for males and 0.801 for females, indicating a higher accuracy in predicting mortality and better differentiation between survivors and non-survivors.<br><br>On the other hand, the Cox model achieved a slightly lower AUC, with 0.719 for males and 0.787 for females, showing that RSF is more precise in predicting risk for dyslipidemia | For the Random Survival Forest (RSF) model, the Brier Scores were as follows: 0.119 for 3 months, 0.119 for 6 months, and 0.128 for 12 months. These lower scores indicate that the RSF model provides highly accurate predictions of mortality, closely aligning with actual outcomes.<br><br>In contrast, the Cox Proportional Hazards (Cox PH) model exhibited higher Brier Scores: 0.138 for 3 months, 0.146 for 6 months, and 0.156 for 12 months. These higher values suggest that the predictions made by the Cox model are less accurate compared to those of the RSF model. | In dyslipidemia studies, the Random Survival Forest (RSF) model outperforms the Cox Proportional Hazards (Cox PH) model in handling various factors. RSF effectively manages nonlinear relationships among variables such as age, gender, and cholesterol levels, revealing complex interactions that influence mortality risk, while accommodating a large number of variables simultaneously. In contrast, Cox relies on linear assumptions, which limits its ability to analyze complex data and capture multiple interactions between variables. Additionally, Cox requires careful management of missing data, which can impact its predictive accuracy. Thus, RSF provides a more comprehensive analysis, making it a superior choice for assessing risks in dyslipidemia. | The RSF model demonstrates stable prediction error rates when using 1000 trees, indicating its efficiency. Additionally, the error rate is lower in females than in males, suggesting that the model performs better in predicting health outcomes for women. In contrast, the CPH model lacks direct information on prediction error but measures performance through C-statistics. The C-statistics for the CPH model are lower than those for the RSF model, implying that it may be less accurate in its predictions | The Cox Proportional Hazards Model (CPH) does not provide a clear measure of variable importance like VIMP in the Random Survival Forests (RSF) model because it relies on regression coefficients to determine the impact of variables on risk without offering a unified assessment of their importance. While the RSF model directly shows the effects of each variable through VIMP values, CPH requires additional analyses to understand variable impacts, making importance assessment more complex. CPH results indicated that important variables for males include age, BMI, TC, TG, HDL-C, LDL-C, GGT, and ALT, while for females, they include age, TC, TG, HDL-C, LDL-C, and GGT. Despite showing acceptable accuracy (C-statistics = 0.719 for males and 0.787 for females), the model lacks the ability to evaluate non-linear relationships between variables and outcomes. | The (RSF) model is capable of capturing these non-linear relationships, allowing it to adapt better to the data. Partial plots show that variables such as TC, TG, HDL-C, and LDL-C have non-linear effects on survival probabilities, enhancing the model's accuracy in predicting outcomes. In contrast, the Cox Proportional Hazards Model (CPH) assumes a linear relationship between predictors and the hazard of the event. This assumption can lead to oversimplification, as it may fail to account for the complexities present in the data, thereby reducing the model's effectiveness in capturing the intricacies of variable interactions. | The study developed a gender-specific risk model for predicting dyslipidemia using the Random Survival Forest (RSF) method. The RSF model demonstrated better discriminative performance compared to the Cox Proportional-Hazards (CPH) model. This finding suggests that RSF can effectively identify high-risk individuals during routine health check-ups, offering significant potential for the prevention of dyslipidemia and related diseases. The results emphasize the importance of using robust predictive models in clinical settings to improve health outcomes. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **2.Individual mortality risk predictive system of patients with acute-on chronic liver failure based on a random survival forest model [2]** | The Random Survival Forest (RSF) model demonstrated higher accuracy in predicting mortality rates among patients with Acute Chronic Liver Failure (ACLF) compared to the Cox Proportional Hazards model. RSF achieved AUROC (Area Under the Receiver Operating Characteristic Curve) values of approximately 0.916, 0.916, and 0.905 at 3, 6, and 12 months, respectively. These values indicate the model's ability to distinguish between patients with a high probability of death and those with a low probability, where an optimal AUROC value is close to 1. In contrast, the Cox model recorded AUROC values of 0.872, 0.866, and 0.848 for the same time periods, indicating a less effective performance in prediction. | The RSF model demonstrated better performance in terms of error rates compared to Cox. The RSF recorded Brier Score values of 0.119, 0.119, and 0.128 at 3, 6, and 12 months, respectively, indicating higher accuracy. In contrast, Cox had values of 0.138, 0.146, and 0.156, which indicate a higher error rate. The accuracy of RSF is attributed to its flexibility in handling complex data, improving risk predictions | In the study of acute liver failure, the RSF model demonstrated superiority in handling complex interactions among explanatory variables. While the Cox model requires fixed assumptions about the effects of variables, which may negatively impact its accuracy in analyzing the data, RSF effectively accounted for interactions such as hepatic encephalopathy, age, and sodium levels, leading to better estimations of risks associated with these complex variables. In contrast, the Cox model was less capable of analyzing these interactions, resulting in poorer performance in predicting mortality rates. This flexibility in RSF contributed to improved risk prediction accuracy in patients with acute liver failure. | The RSF model demonstrated superiority in handling complex interactions among explanatory variables. While the Cox model requires fixed assumptions about the effects of variables, which may negatively impact its accuracy in analyzing the data, RSF effectively accounted for interactions such as hepatic encephalopathy, age, and sodium levels, leading to better estimations of risks associated with these complex variables. In contrast, the Cox model was less capable of analyzing these interactions, resulting in poorer performance in predicting mortality rates. This flexibility in RSF contributed to improved risk prediction accuracy in patients with acute liver failure. | --- | In the analysis of acute liver failure, the Cox Proportional Hazards model utilized several important variables, such as age, serum sodium levels, the International Normalized Ratio (INR), hepatic encephalopathy (HE), and acute kidney injury (AKI). However, the Cox model relies on fixed assumptions, making it less capable of detecting nonlinear relationships between these variables and their impact on mortality risk. On the other hand, the Random Survival Forest (RSF) model offers a significant advantage through its ability to identify nonlinear interactions among variables, allowing for more accurate risk estimates. In addition to the variables used in the Cox model, RSF can also analyze red cell distribution width (RDW) and factors associated with complications, such as pulmonary infections and gastrointestinal bleeding. This expanded analysis of variables aids in better understanding the impact of these factors on patient outcomes. Due to this flexibility in handling nonlinear interactions, the RSF model was able to identify the most influential variables on patient outcomes more effectively, facilitating a better clinical understanding of risk factors compared to the Cox model, which may overlook opportunities to uncover these complex relationships. | The study showed that the Random Survival Forest (RSF) model was more accurate in predicting mortality rates for acute-on-chronic liver failure (ACLF) patients compared to the Cox Proportional Hazards model. This advantage was due to RSF's ability to handle nonlinear interactions between variables, while the Cox model relied on fixed assumptions, which reduced its accuracy in cases involving complex interactions. **Other models** used in the study included MELD (Model for End-Stage Liver Disease), which was used to assess liver disease severity based on creatinine, bilirubin levels, and INR. The results showed that while MELD was useful for assessing disease severity, it was less accurate than RSF in predicting mortality risk. The ABIC (Age-Bilirubin-INR-Creatinine) score, which uses age, bilirubin, creatinine, and INR to estimate mortality risk, also provided good results but was still less accurate than RSF. Additionally, the iMELD (Integrated MELD) model, an improved version of MELD that incorporates age and sodium levels, showed better accuracy than standard MELD but did not outperform RSF in predicting mortality. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **3.Development of an Electronic Frailty Index for Predicting Mortality and Complications Analysis in Pulmonary Hypertension Using Random Survival Forest Model [3]** | Performance was measured using cross-validation (5-fold cross-validation). The results showed that the RSF model was accurate in approximately 92% of positive cases, achieving an accuracy of 0.9263 compared to an accuracy of 0.8382 for the Cox model. In predicting cardiovascular complications, the RSF model had an accuracy of 92% versus 84% for Cox, while it also achieved an accuracy of 92% in predicting kidney complications compared to 83% for Cox, and an accuracy of 92% in predicting diabetes complications compared to 84% for Cox. These results reflect the superiority of the RSF model in providing more accurate predictions and reducing errors when predicting complications. | ---------- **Relationship of the Electronic Frailty Index (eFI) to the Models**: The Electronic Frailty Index (eFI) is an effective tool that enhances the performance of the Random Survival Forest (RSF) model in predicting the risks of mortality and complications among patients with pulmonary hypertension (PHTN). The results showed that the eFI value significantly differed between patients who died and those who did not, with a median eFI of 9.0 (interquartile range: 8.0–10.0) for patients who died, compared to 8.0 (interquartile range: 6.0–9.0) for those who survived, with a P-value < 0.0001. Moreover, the eFI demonstrated a strong correlation with mortality risk, with a hazard ratio of 1.25 (95% confidence interval: 1.22–1.29, P < 0.0001). These findings indicate that a higher eFI value is associated with increased mortality risk. By utilizing the eFI, healthcare providers can more accurately identify patients at greater risk for complications, which contributes to improved healthcare strategies and the tailoring of appropriate treatments based on assessed risks. | The RSF model demonstrated a better ability to handle variables compared to the Cox model. RSF was capable of capturing non-linear relationships and interactions between factors such as age at diagnosis and readmission intervals. While the Cox model relies on simple linear assumptions, RSF provided deeper insights into how these variables affect clinical outcomes, thereby enhancing the accuracy of predictions. | 1. **Recall:** In the study, the Recall for the RSF model was 0.9058, while the Cox model achieved 0.8992. This indicates that RSF was more effective in identifying positive cases, reducing the risk of missing patients at high risk of complications. 2. **Area Under Curve (AUC):** The Area Under the Curve (AUC) for the RSF model was 0.9478, while the AUC for the Cox model was 0.9051. The higher AUC of the RSF model signifies its excellent ability to distinguish between positive and negative cases, enhancing effective clinical decision-making. **3. C-index:** The C-index for the RSF model was 0.9361, compared to 0.9240 for the Cox model. This indicates that RSF has a better ability to rank patients based on their risks, helping to identify those most likely to experience complication | The RSF model utilized techniques such as importance measurement and minimal depth, allowing it to identify the most influential variables. The results showed that age at diagnosis, readmission intervals, and length of hospital stay were among the most important variables in predicting clinical outcomes. These variables play a critical role in predicting mortality and complications, helping doctors make informed decisions to improve patient care. | the Random Survival Forest (RSF) model demonstrated efficiency in handling non-linear relationships between variables. Unlike the Cox model, which relies on simple linear assumptions, RSF was able to identify complex interactions among factors such as age at diagnosis, readmission intervals, and length of hospital stay. This deep understanding of non-linear relationships contributed to enhancing the accuracy of clinical predictions. | The study focused on using the Random Survival Forest (RSF) model to predict the risks of patients with pulmonary hypertension (PHTN), demonstrating its superiority over the Cox model in several aspects. RSF achieved higher accuracy, as it was more effective in identifying positive cases and played a significant role in improving predictions of mortality and complications. The model also showed effective capability in handling non-linear relationships, which contributed to enhancing clinical outcomes. The Electronic Frailty Index (eFI), used in conjunction with the RSF model, helped improve prediction accuracy, as its elevated values were associated with an increased risk of mortality. These results highlight the effectiveness of RSF and its role in enhancing patient management and delivering better healthcare. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **4.Optimal Tuning of Random Survival Forest Hyperparameter with an Application to Liver Disease [4]** | In the study, the performance of the Cox Proportional Hazards (Cox-PH) model was evaluated against other parametric models using the Akaike Information Criterion (AIC). The results indicate that the Cox-PH model achieved the lowest AIC value, recording 185.7233, making it the most efficient compared to the other models. In contrast, the Exponential model had an AIC value of 427.7262, the Weibull model 417.2082, and the Lognormal model 415.6323. These values demonstrate that Cox-PH outperforms the other models, establishing it as the best choice based on the AIC criterion, which reflects the model's efficiency in prediction. | In the study, the predictive accuracy of several models was evaluated using Brier Score (IBS), where lower values indicate higher accuracy. The Tuning Random Survival Forest (TRSF) model demonstrated the best performance, achieving IBS = 0.044, indicating it was the most accurate in predicting survival time. It was followed by the Random Survival Forest (RSF), which recorded IBS = 0.045, and the Cox Proportional Hazards (Cox-PH) model, which scored IBS = 0.046, along with the Forest model, which also achieved IBS = 0.046. Finally, the Kaplan-Meier model had the lowest accuracy among all the models, with IBS = 0.059. These results highlight that the TRSF model was the most accurate in predicting survival time compared to the other models. | The Cox Proportional Hazards (Cox-PH) model demonstrated a limited capacity to analyze variables, as it identified only three key covariates that affect survival, while assuming that their effects remain constant over time. In contrast, Tuning Random Survival Forest (TRSF) was able to handle 16 covariates, including numerical, binary, and categorical variables, allowing it to identify both positive and negative effects of factors such as bilirubin levels and age. This difference reflects TRSF's ability to analyze complex relationships between covariates and survival time more flexibly, making it the better choice for clinical applications. | The (TRSF) model demonstrated high accuracy in predicting survival time, while the (Cox-PH) model also showed good performance. Through the analysis of Hazard Ratios (HRs), researchers were able to understand the impact of variables such as albumin levels and age on the risk of death, enhancing the predictive accuracy. | The study utilized Variable Importance (VIMP) within the Tuning Random Survival Forest (TRSF) model to determine the impact of 16 variables on the survival of patients with liver disease. The analysis showed that bilirubin levels and age had positive effects on survival, while low albumin levels and the presence of ascites had negative effects. VIMP aided in understanding the relationships between the variables and survival time, enabling physicians to make informed treatment decisions and enhancing the development of predictive models to improve treatment strategies. | The study mentions that Cox-PH relies on certain assumptions that may limit its effectiveness in capturing complex relationships between variables. In contrast, TRSF has the ability to accommodate non-linear interactions. From this, we conclude that the flexibility of TRSF allows it to improve its predictive accuracy by identifying the intricate effects of variables on survival outcomes, making it more suitable in survival analysis contexts where relationships are non-linear. | The study compares the Cox-PH model and the TRSF model. The results show that TRSF excels at handling complex and non-linear relationships, while Cox-PH is limited in its effectiveness due to its reliance on fixed assumptions. Additionally, the use of VIMP helped identify significant variables, indicating TRSF's ability to analyze a larger set of covariates. It is concluded that these comparisons enhance the understanding of the effectiveness of different models in survival analysis. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **5.Prediction of prognosis and survival of patients with gastric cancer by a weighted improved random forest model: an application of machine learning in medicine [5]**<br><br>**Note: "This paper did not include a comparison with the Cox model."** | In the study, three models were compared in terms of accuracy in predicting the survival of gastric cancer patients, and the TLWRF model clearly outperformed the others. The traditional Random Forest (RF) model achieved an accuracy of 85.12% in predicting patient outcomes, while the OOBWRF model showed a slight improvement with an accuracy of 85.55%. The TLWRF model surpassed both, achieving the highest accuracy at 85.91%. Additionally, in terms of the AUC, which reflects the model's ability to distinguish between surviving and non-surviving patients, TLWRF scored 80.88% compared to 70.29% for the RF model. Overall, TLWRF had an average AUC of 78.94%, compared to 76.61% for the RF, demonstrating that TLWRF was the most accurate and effective model for predictions. | This study did not mention or use Brier Scores in the comparisons between the models. | The TLWRF model demonstrated its superiority in handling variables by accurately identifying important variables, as it was able to analyze a larger set of variables and select those most impactful on the survival of gastric cancer patients, such as age, gender, and disease stage, allowing it to provide more precise predictions. In contrast, the other two models (RF and OOBWRF) did not exhibit the same level of capability in identifying important variables with the same accuracy, as they relied on fixed weights that were not optimized based on the impact of the variables, reducing their precision in handling the data. Additionally, TLWRF utilized a weighting method that reflects the influence of each variable on prediction outcomes, which contributed to improved performance, while the other models did not adopt this strategy, resulting in less accurate prediction outcomes. | ---------- | In the study, the importance of variables (VIMP) was assessed as a key tool for determining the impact of each variable on predicting the survival of gastric cancer patients. VIMP measurement techniques were utilized in three models: Random Forest (RF), Out-of-Bag Random Forest (OOBWRF), and Tree-Level Weighted Random Forest (TLWRF). Analyses showed that the TLWRF model excelled in identifying high-impact variables, such as age, gender, and disease stage, which significantly affected patient survival. By effectively leveraging VIMP, TLWRF achieved an accuracy of 85.91%, outperforming the other models that were unable to utilize this information as effectively. Furthermore, the VIMP-based analyses provided valuable clinical benefits, allowing physicians to understand the key factors influencing patient outcomes, thereby aiding them in making more informed treatment decisions. | The study did not address non-linear relationships specifically, nor did it clarify how the Cox model handles these relationships within the context of the study. | In the study, three models of random forests were compared: the traditional Random Forest (RF), Out-of-Bag Random Forest (OOBWRF), and Tree-Level Weighted Random Forest (TLWRF). The results showed that the TLWRF model was the most accurate, outperforming the other two models in predicting the survival of gastric cancer patients. Additionally, TLWRF demonstrated a better ability to identify important variables through VIMP analysis.<br><br>However, the study was not solely focused on comparing the Cox model with the random forest models; it also included other models, such as logistic regression, which were not discussed in detail. This provides a broader context for understanding the relative performance of these models in analyzing patient survival data. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **6.Use of Survival-SVM combined with Random-Survival-Forest to predict the survival of nasopharyngeal carcinoma patients [6]** | The models in the study were evaluated using the C-index to determine their accuracy in predicting patient survival. The Cox Regression model achieved a C-index of 0.740 for the training set and 0.721 for the test set, indicating moderate accuracy. In contrast, Survival-SVM achieved the highest accuracy among the three models, with a C-index of 0.785, making it the most efficient. Random Survival Forest (RSF) ranked second with a C-index of 0.729, performing better than Cox Regression but less than Survival-SVM. | n the study, the Brier Score was used to evaluate the accuracy and stability of the Random Survival Forest (RSF) model. Although specific numerical values for the Brier Score were not provided, the study referred to OOB Brier and CRPS curves to illustrate the model's performance. The results indicated that the OOB error rate stabilized when the number of trees reached 60, suggesting that the model was stable and had good performance | The results showed that RSF was effective in handling complex and missing data, meaning it was not significantly affected by the issue of missing values, thus making it more reliable in analyzing patient survival. In contrast, Cox Regression was more affected by missing data, which could negatively impact prediction accuracy. These findings suggest that RSF has a notable advantage in environments characterized by limited or complex data, making it a better choice for analyzing patient survival in cases of incomplete data. | The stability of the RSF model was assessed through OOB error rates, which stabilized at 60 trees, indicating reliability in its predictions. In comparison, RSF provides greater flexibility and better performance in cases of high-dimensional data or nonlinear relationships, whereas Cox relies on assumptions that may negatively impact its accuracy. Therefore, RSF outperforms Cox in terms of stability. | In the study, the researchers utilized the Random Survival Forest (RSF) model to analyze data from nasopharyngeal carcinoma patients. The VIMP analysis results revealed that M stage*** was one of the most significant factors affecting patient survival. The RSF model allowed for an accurate assessment of variable importance, showing that patients classified as M0 had better survival expectations compared to those classified as M1. This indicates that early detection of cancer and appropriate therapeutic interventions can significantly impact patient outcomes, highlighting the importance of using RSF in survival analysis and identifying critical factors in prognosis. | It was found that Cox Regression was less accurate in cases where the proportional hazards assumption was not met, leading to inaccurate predictions of patient survival. In contrast, alternative models like Random Survival Forest (RSF) and Survival-SVM demonstrated a higher ability to handle data with nonlinear relationships, as they did not require linear assumptions, allowing them to capture complex patterns more effectively. The study concluded that Cox Regression may be limited in its analysis when the relationships between variables and survival time are nonlinear. On the other hand, RSF and Survival-SVM exhibited greater flexibility and adaptability to complexity, making them more accurate in predicting survival outcomes in cases involving factors such as tumor grade, metastasis to distant organs, and treatment factors. These variables may show nonlinear effects on survival time, enhancing the effectiveness of RSF and Survival-SVM in handling them. | The results showed that SVM was the best in terms of accuracy, especially in handling small sample sizes and nonlinear relationships, making it the ideal choice in complex scenarios. On the other hand, RSF outperformed Cox in its flexibility to analyze clinical and social factors, as it effectively handled complex and missing data. Although Cox was the fastest in execution, it was less accurate in analyzing all influential variables.<br><br>Ultimately, both RSF and SVM demonstrate better performance in certain contexts, with SVM excelling in accuracy, while RSF has greater adaptability to data complexities |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **7.Predicting Time to Diabetes Diagnosis Using Random Survival Forests [7]** | The Random Survival Forest (RSF) model achieved a C-Index of 0.84, reflecting high accuracy in predicting the time to diabetes diagnosis. The C-Index for the Cox model was not directly calculated in this study for comparison. | In the study, the Brier Score was mentioned generally as a tool for evaluating prediction accuracy, but specific values for the Brier Score were not provided. It was used as a metric to measure the performance of the models in predicting time-to-event outcomes, such as diabetes diagnosis | The RSF model can handle a large number of interrelated variables, such as A1c, FBS, and LDL, allowing for a comprehensive assessment of their impact on diabetes risk. In the study, 14 biomarkers and health-related variables were utilized for data analysis.<br>In contrast, the Cox model is more effective with simpler relationships and may struggle to manage multiple variables, which could negatively impact prediction accuracy.<br>Therefore, RSF is considered more efficient in analyzing medical data that involves complex interrelated factors. | Iterative Imputer was used to handle missing data, improving the quality of the data utilized in the models. The connection here is clear: the better the data quality, the higher the accuracy of models like RSF and Cox, which reduces error rates in predictions. Therefore, effective handling of missing data is a crucial component in enhancing model performance and minimizing prediction-related errors.<br>Additionally, the RSF model allows for precise estimation of the time until diabetes diagnosis for each patient, with predictions ranging from 1 to 15 years after the initial biometric measurement. This capability aids physicians in providing accurate timelines regarding patients' risk of developing diabetes. In contrast, the Cox model does not offer the same level of precise temporal estimation, as it relies more on linear assumptions and lacks flexibility in predicting event timing. Therefore, RSF is considered more effective in providing valuable information about the temporal risks associated with diabetes diagnosis | The (RSF) model outperforms the Cox model in determining the importance of the variables used to predict the time to diabetes diagnosis. The study revealed that A1c was the most important variable with a value of 0.12, followed by FBS with an importance of 0.072, and then Total Cholesterol with an importance of 0.0030. These results indicate RSF's ability to effectively identify the most impactful variables on prediction accuracy. Thus, these figures reinforce the effectiveness of RSF as an advanced analytical tool compared to Cox. | The RSF model can handle non-linear relationships between variables, while the Cox model assumes a linear relationship, making RSF more effective in predicting diabetes from complex data.<br>Additionally, RSF demonstrates a high capacity to manage right censoring, allowing it to analyze cases where the event (diabetes diagnosis) is not observed during the study period. In contrast, the Cox model relies on specific assumptions about data distribution, making it less flexible in these situations. | The study indicates that the (RSF) model demonstrates good performance in predicting the time to diabetes diagnosis, achieving a C-Index of 0.84, which reflects high accuracy in prediction. The RSF is considered more flexible in handling complex data, as it can effectively analyze multiple variables and non-linear relationships.<br><br>Additionally, the RSF allows for the estimation of the expected time to diabetes diagnosis, aiding physicians in providing accurate information to patients about their risks. The research shows that the RSF has significant potential to enhance healthcare by supporting clinical decision-making and personalizing prevention plans for patients. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **8.Explainability of random survival forests in predicting conversion risk from mild cognitive impairment to Alzheimer's disease [8]** | The RSF model achieved an accuracy of 89% in predicting the risk of conversion from mild cognitive impairment (MCI) to Alzheimer's disease (AD) using the c-index, while the Cox model achieved a lower accuracy of 81.8%. This indicates that RSF was more effective in predicting patient conversion compared to the Cox model, enhancing RSF's ability to distinguish patients who may experience this transition. | Regarding the Brier Score, both the RSF model and the Cox model achieved a value of 0.09, indicating good accuracy in predictions related to survival and disease progression. Neither model exceeded the critical threshold of 0.25 over the 48-month period, demonstrating acceptable performance from both models in providing accurate predictions about the development of Alzheimer's disease. | The RSF) model effectively handled variables even in cases of high correlation due to its tree-based design. In the study, a set of important variables was evaluated, such as FDG, which measures metabolic activity in the brain, ABETA42, which is associated with plaque formation, and HCI, which indicates levels of metabolic activity. The model utilized the SHAP method to provide accurate explanations of how each variable influenced the final outcomes. This ability to manage correlated variables enhances the reliability of predictions regarding the progression of Alzheimer's disease. | the Cox model demonstrated lower absolute error rates compared to the RSF model, indicating greater accuracy in its predictions. The evaluation also showed that the RSF model remained within the 95% confidence interval of the Kaplan-Meier curve, reflecting its capability to estimate survival rates effectively. Importantly, neither model exceeded the critical Brier score threshold (0.25) at any time point over the 48-month period, suggesting that both models performed well within acceptable limits. Overall, the Cox model showed higher precision in error rates, while both models maintained satisfactory performance in survival analysis. | The importance of variables was assessed using three different methods: RSF feature importance, permutation importance, and SHAP importance. The results showed that the three most important variables were FDG, ABETA42, and HCI, which maintained the same ranking across all methods used. The correlation between the three assessments of variable importance exceeded 90%, indicating the stability of the important variable rankings across different models | The study noted that the RSF model handled non-linear relationships between variables well, reflecting its ability to model complex interactions without requiring strict assumptions about the form of those relationships. Due to its non-parametric nature, the model can flexibly analyze complex data. Additionally, the study utilized the SHAP method to interpret predictions, revealing that the impact of non-linear relationships was limited in the context of predicting the risk of disease progression. There was no comparison made regarding the performance of the Cox model in this regard. | The study utilized the SHAP (Shapley Additive explanations) method to explain the results of the RSF model, providing a better understanding of the impact of variables on predictions. SHAP relies on game theory to offer accurate individual explanations for each patient, aiding clinicians in identifying the factors influencing the risk of patients progressing from mild cognitive impairment (MCI) to Alzheimer's disease (AD). Overall, the results confirm that the RSF model is an effective and reliable tool for estimating the risk of MCI patients transitioning to Alzheimer's disease, achieving high accuracy and providing precise explanations for the factors influencing this transition. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **9. A Random Forest Model for Post-Treatment Survival Prediction in Patients with Non-Squamous Cell Carcinoma of the Head and Neck [9]** | The random forest model consistently outperformed the traditional TNM-based Cox regression model in terms of AUC (Area Under the Curve) values. | The Brier score was not explicitly mentioned, but the model's performance was validated using AUC, calibration, and decision curves, which indicated good calibration and a high degree of accuracy in survival predictions【10:4†source】 | Error rates were indirectly assessed via the performance metrics such as the AUC, where higher values signify lower classification errors in the random forest model compared to the traditional TNM-based models【10:4†source】. | The random forest model, an ensemble learning algorithm, can capture non-linear relationships between variables, unlike traditional linear regression models like the TNM-based Cox regression. This flexibility allows the random forest to better account for the complex interactions between variables in survival prediction | The random forest model handled a broader set of variables including tumor size, involved node levels, pathology type, and age. Importantly, it did not rely on TNM classification stages, allowing it to capture more nuanced and detailed relationships between clinical features | The most critical variables in the random forest model were determined by Gini coefficients, which highlighted:<br>- Tumor pathology (type and grade)<br>- Levels of involved cervical nodes<br>- Tumor size and patient age<br>- These factors contributed the most to survival prediction accuracy. | The random forest model demonstrated superior accuracy, flexibility, and the ability to handle complex variable interactions compared to the TNM-based Cox regression. It captured non-linear relationships and identified the most important variables for prediction, such as pathology, node levels, and tumor size, which traditional models struggled with |

* The Brier Score is a crucial metric for evaluating the accuracy of predictive models in clinical settings.

** Non-linear relationships refer to the complex interactions between predictor variables and survival outcomes that are not adequately captured by linear models.

*** M stage in cancer classification refers to the extent of tumor spread to other parts of the body and is part of the TNM (Tumor, Node, Metastasis) system used to assess cancer stages. M0 indicates no distant metastasis, reflecting better survival chances for patients, while M1 indicates the presence of distant metastasis, increasing the severity of the condition and reducing survival prospects.

Table 3.4.2: Comparison between survival models from literature review

## 3.5 Conclusion

Studies have demonstrated the superiority of the Random Survival Forest (RSF) model over traditional models, such as Cox Proportional Hazards (CPH), in predicting time-to-event risks and assessing patient outcomes. RSF excels in handling non-linear relationships and complex interactions between variables, in addition to its flexibility in analyzing missing and high-dimensional data. It also provides tools like VIMP to identify the most influential variables, enhancing prediction accuracy and improving clinical decision-making. Moreover, the model has proven its efficiency through various performance metrics, including AUC, C-Index, and Brier Score, making it a robust tool for analyzing complex clinical data and offering precise insights that aid in improving healthcare outcomes and tailoring treatment plans.

# CHAPTER 4

# SUMMARY & FUTURE WORK

## 4.1 Summary of Previous Chapters

This research investigates the application and significance of Random Survival Forests (RSF) in survival analysis, building upon insights and comparisons with traditional models like the Cox Proportional Hazards (CPH) model, as detailed in the previous chapters.

- **Chapter 1** introduced survival analysis, emphasizing its role in predicting time-to-event outcomes in various fields. It highlighted the limitations of traditional models, such as the CPH model, particularly in handling non-linear relationships and complex interactions between variables. The flexibility and robustness of RSF were identified as key strengths, showcasing its ability to manage high-dimensional datasets and censored data effectively.
- **Chapter 2** provided a literature review, discussing significant studies that demonstrated RSF's advantages. These include its capability to reduce competition among variables, evaluate variable importance (VIMP), and manage non-linear interactions. For example, RSF has been successfully applied in predicting type 2 diabetes, liver diseases, and survival outcomes in nasopharyngeal carcinoma and Alzheimer's disease. The review highlighted RSF's higher predictive accuracy and robustness compared to the CPH model, especially in datasets with non-linear relationships.

  Challenges such as computational demands and interpretability were also discussed, emphasizing the role of tools like SHAP in improving RSF's explain ability. Furthermore, recent studies have shown how RSF can integrate diverse clinical and biomarker data to provide personalized predictions, reinforcing its utility in diverse clinical applications. Finally, the chapter underscored the potential of combining RSF with advanced techniques like deep learning to enhance its performance further.
- **Chapter 3** focused on a methodological comparison between RSF and CPH. It highlighted RSF's independence from hypothesis testing, its reliance on ensemble learning techniques (e.g., bootstrap sampling and decision tree aggregation), and its ability to improve predictive accuracy while reducing variance. The chapter also discussed how RSF calculates evaluation metrics like Brier scores and the C-index, which consistently

outperform the metrics of the CPH model. Additionally, a tabular analysis summarized findings from various studies, reinforcing RSF's advantages in handling complex datasets, reducing prediction error, and managing censoring.

### 4.2 Future Work

Due to time constraints, this project primarily focused on illustrating and reviewing RSF using selected studies without extensive application on real-world datasets or exploring diverse populations. Additionally, the work did not delve deeply into the impact of non-linear relationships between variables, which is a critical advantage of RSF.

To build upon the findings of this research, future work should address the following:

1) Testing the Model on Real Medical Data: Applying RSF to real-world datasets across various domains to evaluate its generalizability and reliability in clinical applications.
2) Exploring Non-Linear Effects: Conducting in-depth studies on the impact of non-linear interactions and variable importance to improve the model's interpretability and insights.
3) Integrating Explainable AI Tools: Utilizing tools like SHAP**** to make RSF outputs more interpretable for stakeholders, particularly in medical contexts.

**** Integrating explainable AI tools such as SHAP helps make the results of models like RSF more understandable and transparent. Due to the complexity of RSF and its reliance on ensemble learning, it can be challenging to interpret how each variable influences the outcomes. Tools like SHAP provide clear insights into the importance and impact of individual variables on predictions, enhancing trust in the model and supporting decision-making. This is particularly valuable in medical fields, where users require precise justifications for predictions to improve clinical decisions and facilitate the effective adoption of the model.

### 4.3 Final Remarks

This chapter consolidates the key findings from the research, offering a clear comparison of RSF's strengths and limitations relative to the CPH model. By addressing the outlined future work, RSF can further solidify its position as a transformative tool in survival analysis, contributing to more accurate predictions and better decision-making across diverse fields.

## References

**[1]** Dauda K. A. (2022). Optimal Tuning of Random Survival Forest Hyperparameter with an Application to Liver Disease. *The Malaysian journal of medical sciences : MJMS*, *29*(6), 67–76. https://doi.org/10.21315/mjms2022.29.6.7

**[2]** Saha, P., Marouf, Y., Pozzebon, H., Guergachi, A., Keshavjee, K., Noaeen, M., & Shakeri, Z. (2024). Predicting time to diabetes diagnosis using random survival forests. *MedRxiv*, 2024-02.

**[3]** Xiao, Z., Song, Q., Wei, Y., Fu, Y., & Huang, D. (2023). Use of Survival-SVM combined with Random-Survival-Forest to predict the survival of nasopharyngeal carcinoma patients.

**[4]** Sarica, A., Aracri, F., Bianco, M. G., Arcuri, F., Quattrone, A., Quattrone, A., & Alzheimer's Disease Neuroimaging Initiative. (2023). Explainability of random survival forests in predicting conversion risk from mild cognitive impairment to Alzheimer's disease. *Brain Informatics*, *10*(1), 31.

**[5]** Zhang, X., Liu, G., & Peng, X. (2023). A Random Forest Model for Post-Treatment Survival Prediction in Patients with Non-Squamous Cell Carcinoma of the Head and Neck. *Journal of Clinical Medicine*, *12*(15), 5015.

**[6]** Xu, C., Wang, J., Zheng, T., Cao, Y., & Ye, F. (2021). Prediction of prognosis and survival of patients with gastric cancer by a weighted improved random forest model: an application of machine learning in medicine. *Archives of Medical Science: AMS*, *18*(5), 1208.

**[7]** Zhang, X., Tang, F., Ji, J., Han, W., & Lu, P. (2019). Risk prediction of dyslipidemia for Chinese Han adults using random Forest survival model. *Clinical epidemiology*, 1047-1055.

**[8]** Zhang, Z. Q., He, G., Luo, Z. W., Cheng, C. C., Wang, P., Li, J., ... & Ye, Y. N. (2021). Individual mortality risk predictive system of patients with acute-on-chronic liver failure based on a random survival forest model. *Chinese Medical Journal*, *134*(14), 1701-1708.

**[9]** Zhou, J., Chou, O. H. I., Wong, K. H. G., Lee, S., Leung, K. S. K., Liu, T., ... & Zhang, Q. (2022). Development of an electronic frailty index for predicting mortality and complications analysis in pulmonary hypertension using random survival forest model. *Frontiers in Cardiovascular Medicine*, *9*, 735906.

**[10]** Clinical Biostats. (2023, June 21). *Cox Proportional Hazards Model Tutorial - Clinical Biostats*. https://www.clinicalbiostats.com/tutorials/cox_proportional_hazards/

**[11]** Abd ElHafeez, S., D'Arrigo, G., Leonardis, D., Fusaro, M., Tripepi, G., & Roumeliotis, S. (2021). Methods to Analyze Time-to-Event Data: The Cox Regression Analysis. *Oxidative medicine and cellular longevity*, *2021*, 1302811. https://doi.org/10.1155/2021/1302811

**[12]** Kleinbaum, D. G., & Klein, M. (2006b). *Survival analysis: A Self-Learning Text*. Springer.

**[13]** t-Test, Chi-Square, ANOVA, Regression, Correlation... (n.d.). Datatab.net. https://datatab.net/tutorial/cox-regression

**[14]** Rezaei, M., Tapak, L., Alimohammadian, M., Sadjadi, A., & Yaseri, M. (2020). Review of Random Survival Forest method. *Journal of Biostatistics and Epidemiology*. https://doi.org/10.18502/jbe.v6i1.4760

**[15]** Weathers, B. (2017). Comparison of Survival Curves Between Cox Proportional Hazards, Random Forests, and Conditional Inference Forests in Survival Analysis.

**[16]** FasterCapital. (2024, June 19). التنبؤ بمعدل الخطر مع المتغيرات المشتركة :COX نموذج المخاطر النسبية. *FasterCapital*. https://fastercapital.com/arabpreneur/التنبؤ-بمعدل-الخطر---COX-نموذج-المخاطر-النسبية-مع-المتغيرات-المشتركة.html