University of Cape Town

STA5058W

Master of Science in Biostatistics

Statistical model selection techniques for the Cox proportional hazards model: a comparative study

---

Author:
Jolando Njati

Supervisor:
Freedom Gumedze

A dissertation submitted to the Faculty of Science of the University of Cape Town in partial fulfilment of the requirements for the degree of Master of Science in the Department of Statistical Sciences.

November 2021

Abstract

The advancement in data acquiring technology continues to see survival data sets with many covariates. This has posed a new challenge for researchers in identifying important covariates for inference and prediction for a time-to-event response variable. In this dissertation, common Cox proportional hazards model selection techniques and a random survival forest technique were compared using five performance criteria measures. These performance measures were concordance index, integrated area under the curve, and R^2. To carry out this exercise, a multi-centre clinical trial data set was used. A simulation study was also implemented for this comparison. To develop a Cox proportional model, a training dataset of 75% of the observations was used and the model selection techniques were implemented to select covariates. Full Cox PH models containing all covariates were also incorporated for analysis for both the clinical trial data set and simulations. The clinical trial data set showed that the full model and forward selection technique performed better with the performance metrics employed, though they do not reduce the complexity of the model as much as the Lasso technique does. The simulation studies also showed that the full model performed better than the other techniques, with the Lasso technique over-penalising the model from the simulation with the smaller data set and many covariates. AIC and BIC were less effective in computation than the rest of the variable selection techniques, but effectively reduced model complexity than their counterparts for the simulations. The integrated area under the curve was the performance metric of choice for choosing the final model for analysis on the real data set. This performance metric gave more efficient outcomes unlike the other metrics on all selection techniques. This dissertation hence showed that variable selection techniques differ according to the study design of the research as well as the performance measure used. Hence, to have a good model, it is important to not use a model selection technique in isolation. There is therefore need for further research and publish techniques that work generally well for different study designs to make the process shorter for most researchers.

Keywords: survival analysis; simulation; Cox proportional hazard model selection; integrated area under the curve

Acknowledgments

Firstly, I would like to thank the Lord God Almighty for His grace, abundant supply and guidance and giving me a worthy support system. I would also like to thank Reverend Doctor Joel Roosevelt Gondwe for his mentorship, and for encouraging me to pursue a Masters degree. My special thanks also goes to my parents for loving me and for being patient with me throughout my University of Cape Town journey.

To Associate Professor Freedom Gumedze, thank you so much for providing me with the IMPI data set I am forever grateful for the valuable criticisms, suggestions and guidance that helped improve the material of this dissertation. My gratitude also goes to Professor Francesca Little for her support and letting me use the department's laptop from Honours to Masters. Allan Clark, thank you so much for always checking up on the progress of my studies. My special gratitude goes to Stellies and Osborne houses, thank you so much for being awesome and supporting me greatly through my journey. To the lecturers, examiners, classmates, church family and friends, too numerous to include here, thank you for playing a part in my success story.

This work was partially funded by a Statistical Sciences Departmental scholarship from the University of Cape Town.

Contents

List of Figures

List of Tables

Acronyms

AFT: Accelerate failure time.

AIC: Akaike Information Criteria.

aLASSO: Adaptive Least absolute shrinkage & selection operator.

BIC: Bayesian Information Criteria.

C-index: Concordance index.

CDA: Coordinate decent algorithm.

CHF: Cumulative hazard function.

CI: Confidence interval.

CV: Cross validation.

DAC: divide-and-conquer.

FNSS: Forward nested subset selection.

GLM: Generalised linear models.

HIV: Human immunodeficiency virus.

iAUC: Integrated area under the curve.

IMPI: Investigation of the management of pericarditis in Africa.

KM: Kaplan-Meier.

LASSO: Least absolute shrinkage & selection operator.

LRT: Likelihood ratio test.

MAR: Missing at random.

MIP: Mycobacterium Indicus Pranii.

Mw Mycobacterium Indicus Pranii.

NYHA: New York heart association.

OLS: Ordinary least squares.

OOB: Out-of-bag.

PH: Proportional hazards.

RF: Random forest.

ROC: Receiver operator curve.

RSF: Random survival forest.

SIS: Sure independence screening.

STMC: Stepwise tuning in maximum concordance index.

TB: Tuberculous.

VIF: variance inflation factor.

VIMP: variable importance.

Chapter 1

Introduction

Model selection is an integral part of a statistical model-building process. It is one of the model-fitting steps required to ensure valid statistical inferences. The advancement in technology has seen improvements in data collection, which has in turn led to the presence of many covariates also known as prognostic variables or risk factors which may influence a response of scientific interest. For ease of interpretation of the models, a parsimonious model with few but important covariates that fit the observed data well is desirable. Researchers usually develop prediction models for the response variables to be used on unseen data. Thus, it is imperative to utilise techniques that minimise the error in precision of estimation and prediction of new responses [68].

The Cox proportional hazards model (Cox PH) [15] is usually adopted to assess the influence of covariates on a time to event outcome in survival analysis [32; 83]. Censoring makes the nature of data in survival analysis differ from that of standard regression analysis, because the actual failure times are not known for censored observations. Hence, calculating the least-squares estimator is challenging. However, extensions of the least-squares principle for censored data exist [9; 43; 54; 53], but their development is limited to the academic sphere as the algorithms are said to be computationally intensive with no guarantee of a consistent solution [39]. Therefore, model selection techniques used in standard regression have been extended to Cox PH models.

Techniques to be considered in this dissertation are: stepwise regression (forward selection, backward elimination), best subset selection, Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), shrinkage methods (Lasso and Ridge regression), and random survival forests. These techniques are common procedures that are adopted in most survival data analyses. These techniques will be implemen-

ted in statistical software R [72] and applied on a real data set from a multicentre clinical trial, and a simulated data set. To identify the "best" technique for selecting a final model in this dissertation, comparative performance methods will be employed. One thing observed from literature is that there is no agreement as to which comparison method is better, hence three different methods will be used. The background and description of the real data set is given in Chapter 2 of this dissertation.

1.1 Motivation

Modern data sets tend to have many covariates with many possible interactions, which may affect the performance of a model. There is, therefore, a need to find informative covariates that affect survival time given an event of interest, leading to a parsimonious model that can be used for prediction. The selection of models is of great importance in finding the number of covariates that are informative for a model to perform relatively better in estimation of effects than when non-informative covariates were included. It helps to have a model that is simple to implement practically and importantly, easy to interpret clinically.

There is a wide array of Cox PH model selection techniques that have been developed. Most of these techniques are extensions of linear regression model selection techniques. This dissertation aims at bringing together some of the common Cox PH model selection techniques and comparing them using a clinical trial data set and a simulated data set. Understanding this may lead to providing niches that will encourage creative development in finding techniques that work generally well for different scenarios and which will help future researchers in the field.

1.2 Aims and objectives

The main aim of this dissertation is to investigate standard and modern selection methods for the Cox PH model that will result in a model that is an appropriate representation of the data. Thus, finding a technique that "correctly" identifies important risk factors for a model is imperative for estimation and/or prediction. These techniques will be compared for a time to event health outcome through a real data set and carefully constructed simulations.

The primary objective is to compare different model selection techniques in the survival analyses of the multicentre clinical trial data set. The focus will be on time to composite endpoint of death, cardiac tamponade requiring pericardial drainage or constrictive pericarditis. A composite endpoint is an outcome in which competing

events are combined and the event of interest is observed when either one of the combined events occurs first. The survival analyses will entail the investigation of the effect of baseline covariates on the following events of interest:

- time to a composite event
- time to death
- time to constrictive pericarditis

A secondary objective of this dissertation will be to compare the model selection techniques in a simulation study with three varying censoring levels.

1.3 Structure of the dissertation

The outline of this dissertation is as follows: Chapter 2 gives the background and descriptive statistics to the multicentre clinical trial data set. Chapter 3 reviews some literature that is available in the subject area of survival analysis and Cox PH. Chapter 4 discusses the theory of model selection techniques and some comparative methods used for survival data. This will be followed by a summary of model selection estimation results of the Cox PH models from a multicentre clinical trial data set in Chapter 5. It also illustrates plots for checking overall model fit. This chapter will also present comparative study results on all the model selection techniques being investigated for the clinical trial data set and simulated data. Chapter 6 follows with the discussion of the results. The conclusion and some areas for future research are presented in Chapter 6.

Chapter 2

Background to IMPI trial data set

In this chapter, the background and summary statistics for the IMPI clinical trial data set used in this dissertation are given. This data set will be revisited in Chapter 5.

2.1 IMPI clinical trial data set

The IMPI clinical trial was a double-blind 2-by-2 factorial design randomised trial of 1400 adult patients in 19 centres across eight sub-Saharan countries with definite and probable Tuberculous (TB) pericarditis. The number of patients per centre ranged from 1 to 350 patients, aged 18 and older. Carried out between January 2009 and February 2014, the trial followed the cohort of patients randomly assigned to either 1) prednisolone or placebo (120 mg/day to 5 mg/day) over 6 weeks and 2) Mycobacterium indicus pranii (Mw) or placebo, administered in five injections over the course of 3 months. The trial hence had four arms: Mw^+P^+, Mw^-P^-, P^-Mw^+ and P^+Mw^-, where - means placebo and + means active treatment. For example, Mw^-P^- means the arm had placebo for both Mw and prednisolone treatments. In this study, the focus will be on comparing all patients P^+ vs all patients P^- because the initial IMPI study flagged Mw as redundant.

The trial used central concealed randomisation to group the patients into experimental and control groups. The experimental groups received the treatments and the control groups received the placebo. Of the 1400 patients, two-thirds - 939 to be precise - had associated human immunodeficiency virus (HIV) infection [51]. The data set had 44 covariates, all of which can be deemed to be of clinical importance and were included for analysis [61].

The IMPI trial had the objective of determining the effects of ancillary treatment

on major effectiveness and safety outcomes measured by the effect on opportunistic infections [51]. The primary outcome analyses between patients who received either treatments showed no significant differences in the combined outcome of death from all causes, cardiac tamponade, or constriction. [51]. A secondary objective of the same data set was to measure the safety of immunomodulatory treatment and it was concluded that the use of prednisolone therapy was associated with an increased risk of cancer, mostly related to HIV among the HIV-infected patients. However, the therapy was associated with a significant reduction in pericardial constriction [51].

Description of baseline characteristics in the IMPI clinical trial data set

Of the 1400 patients enrolled for the study, 939 were HIV positive and 431 were HIV negative. Of these, 706 were assigned prednisolone and 694 the placebo. Out of the 1400 patients, 336 patients experienced the composite event, (i.e. death, cardiac tamponade or constriction), 246 patients died and 85 patients experienced constriction. Patients lost to follow-up or whose event was observed beyond the study period were right-censored.

As with most data sets, the detection of missing data is imperative. The assumptions that data is missing at random (MAR) and that censoring is non-informative are used. A preliminary univariate analysis was done for each of the three response variables on the 44 clinically appropriate covariates. Any covariates that had more than 40% missing values for the IMPI data set were dropped because imputing on these covariates would introduce bias in the results due to larger error terms generated. Patients (row observations) with missing values (about 9%) for the remaining covariates were set to be deleted. This was because the selection techniques currently employed only work with complete data. Therefore, only 24 covariates were included for further analysis. The covariates that had missing data are given in the Appendix A. A description of some of the candidate covariates in the data set appear below in Table 2.1.

Baseline characteristic distributions for the treatment group

To compare the distributions of the baseline characteristics for the prednisolone treatment group, the Mann-Whitney test was used for the continuous variables and the Chi-Square test or Fisher's exact test for the categorical variables. The characteristics for the baseline categorical covariates are given in Table 2.1 and the continuous covariates in Table 2.2 below. The p-values of the covariates (p-vaue ≥ 0.05) suggest that at baseline, all covariates showed no significant differences in the distribution of the baseline characteristics of the participants by treatment

group.

Table 2.1: Baseline characteristics for categorical covariates of patients randomised either to the prednisolone or placebo group

Covariate	**Placebo (N = 694)** N (%)	**Prednisolone (N = 706)** N (%)	**Overall (N = 1400)** N (%)	**p-value**
New York Heart Association Class (NYHA) at study entry				0.5663
I	119 (8.52)	137 (9.81)	256 (18.32)	
II	352 (25.20)	342 (24.48)	694 (49.68)	
III	167 (11.95)	163 (11.67)	330 (23.62)	
IV	54 (3.87)	63 (4.51)	117 (8.38)	
Pericardiocentesis at randomisation				1.0000
No	275 (19.64)	279 (19.93)	554 (39.57)	
Yes	419 (29.93)	427 (30.50)	846 (60.43)	
Sex				0.5280
Female	299 (21.36)	317 (22.64)	616 (44.00)	
Male	395 (28.21)	389 (27.79)	784 (56.00)	
Country				0.8253
Other	189 (13.50)	197 (14.07)	386 (27.57)	
South Africa	505 (36.07)	509 (36.36)	1014 (72.43)	
ARV status at study entry				0.8286
Unknown	236 (16.86)	237 (16.93)	473 (33.79)	
No	354 (25.29)	370 (26.43)	724 (51.71)	
Yes	104 (7.43)	99 (7.07)	203 (14.50)	
Palpable pulsus paradoxus				0.5563
No	565 (40.36)	565 (40.36)	1130 (80.71)	
Yes	129 (9.21)	141 (10.07)	270 (19.29)	
Peripheral oedema				0.9598
No	408 (29.14)	417 (29.79)	825 (58.93)	
Yes	286 (20.43)	289 (20.64)	575 (41.07)	
Systolic blood pressure				0.3043
$\leq$ 90 mm Hg	47 (3.36)	59 (4.22)	106 (7.58)	
$>$ 90 mm Hg	647 (46.25)	646 (46.18)	1293 (92.42)	
Heart Rate				0.4293
$\leq$ 100	294 (21.03)	314 (22.46)	608 (43.49)	
$>$ 100	400 (28.61)	390 (27.90)	790 (56.51)	
Haemoglobin				0.4867
$\leq$ 10	363 (26.06)	385 (27.64)	748 (53.70)	
$>$ 10	326 (23.40)	319 (22.90)	645 (46.30)	
White Blood Count				0.1215
$\leq$ 10/mm3	635 (45.49)	662 (47.42)	1297 (92.91)	
$>$ 10/mm3	57 (4.08)	42 (3.01)	99 (7.09)	
Creatinine				0.9318
$\leq$ 105 umol/l	548 (43.08)	564 (44.34)	1112 (87.42)	
$>$ 105 umol/l	80 (6.29)	80 (6.29)	160 (12.58)	

Table 2.1 Categorical baseline characteristics continued

Covariate	**Placebo (N = 694)** N (%)	**Prednisolone (N = 706)** N (%)	**Overall (N = 1400)** N (%)	**p-value**
Left ventricular ejection fraction				0.3742
$\leq 55\%$	1 (0.07)	4 (0.29)	5 (0.36)	
$> 55\%$	693 (49.50)	702 (50.14)	1395 (99.64)	
Effusion size				0.6720
Large (≥ 2 cm)	460 (33.85)	462 (34.00)	922 (67.84)	
Medium ($1 - 2$ cm)	159 (11.70)	172 (12.66)	331 (24.36)	
Small (< 1 cm)	56 (4.12)	50 (3.68)	106 (7.80)	
Tamponade at study entry				0.9638
No	210 (21.45)	213 (21.76)	423 (43.21)	
Yes	278 (28.40)	278 (28.40)	556 (56.79)	
Constriction at study entry				0.5216
No	280 (28.00)	279 (27.90)	559 (55.90)	
Yes	211 (21.10)	230 (23.00)	441 (44.10)	
HIV Status				1.0000
HIV-	213 (15.55)	218 (15.91)	431 (31.46)	
HIV+	465 (33.94)	474 (34.60)	939 (68.54)	
Chest Xray pulmonary infiltrate				0.4622
No	434 (34.01)	422 (33.07)	856 (67.08)	
Yes	203 (15.91)	217 (17.01)	420 (32.92)	
Atrial fibrillation on ECG				0.3096
No	500 (47.26)	501 (47.35)	1001 (94.61)	
Yes	24 (2.27)	33 (3.12)	57 (5.39)	
Definite TB pericarditis status				0.6164
No	572 (40.86)	590 (42.14)	1162 (83)	
Yes	122 (8.71)	116 (8.29)	238 (17.00)	
Pericardial thickness				–
Normal (≤ 2 mm)	693	702	1395	

Table 2.2: Baseline characteristics for continuous covariates of patients randomised either to the prednisolone or placebo group

Covariate	**Placebo (N = 694)** **Median (IQR)**	**Prednisolone (N = 706)** **Median (IQR)**	**Overall (N = 1400)** **Median (IQR)**	**p-value**
Age (years)	35.39 (17.68)	35.90 (17.63)	35.56 (17.65)	0.7625
Weight (kg)	58.00 (15.65)	57.35 (16.00)	58.00 (16.00)	0.7047
Duration of symptoms (days)	30 (21.75)	30 (46.00)	30.00 (28.00)	0.3961

Chapter 3

Literature review

In this chapter, we cover the literature on the basic concepts of survival analysis, the Cox PH model and some of its extensions, and some previous model selection comparative studies for the Cox PH model.

3.1 Basic concepts of survival analysis

Survival analysis deals with the analysis of time to events of interest. The time until an occurrence of one or more events is the outcome variable of interest. Survival analysis can be applied to many problems in different fields. For example, an event can be death, hospitalisation, graduation, retirement, expiry of a product or divorce. In this dissertation, the focus will be on the application of survival methods when analysing biomedical data. The main objectives of survival analysis are descriptive and modelling.

Some of the important functions in survival analysis are the survival function and the hazard function. The theoretical description of these functions (technical or otherwise) has been discussed by several authors [15; 17; 32; 55].

Let T be a non-negative valued random variable associated with survival time; of which t is the actual survival time for an individual. *F(t)* is the distribution function which describes the probability that the survival time is less than some value, t, given by:

$$\begin{aligned} F(t) &= P(T < t) \\ &= \int_0^t f(v)\, dv, \end{aligned} \tag{3.1}$$

where *f(v)* is the probability distribution function of T for some variable v.

The probability that an individual will survive from the time of origin to a time point beyond t gives the survival function *S(t)*:

$$\begin{aligned} S(t) &= P(T \geq t) \\ &= 1 - F(t). \end{aligned} \tag{3.2}$$

The survival function is usually defined in terms of the hazard function also known as the instantaneous failure-rate or force of mortality in time t, $h(t)$, which can also be written as:

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \ < \ T \ < \ t + \Delta t \mid T \ > \ t)}{\Delta t}, \tag{3.3}$$

where Δt is the instantaneous change in time, t,

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \left\{ \frac{1}{\Delta t} \frac{\Pr(t \leq T < t + \Delta t)}{\Pr(T \geq t)} \right\} \\ &= \lim_{\Delta t \to 0} \left\{ \frac{1}{\Delta t} \frac{F(t + \Delta t) - F(t)}{S(t)} \right\} \\ &= \lim_{\Delta t \to 0} \left\{ \frac{F(t + \Delta t) - F(t)}{\Delta t} \right\} \frac{1}{S(t)} \\ &= \frac{f(t)}{S(t)}. \end{aligned} \tag{3.4}$$

The hazard function represents the risk or hazard of observing an event of interest at some time t, conditional on the individual having survived to that time. Another relationship between $h(t)$ and $S(t)$ is derived by considering the logged relationship of $S(t) = 1 - F(t)$ and then using the chain rule of differentiating composite functions:

$$\begin{aligned} \frac{d}{dt} \log(1 - F(t)) &= \frac{\frac{d}{dt}(1 - F(t))}{1 - F(t)} \\ &= -\frac{f(t)}{S(t)}, \end{aligned}$$

to get

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= -\frac{d}{dt}\left\{\log S(t)\right\}. \end{aligned} \tag{3.5}$$

It is an intermediary measure for estimating $h(t)$ and is a diagnostic tool in assessing model validity. It can be thought of as the number of events that would be expected for each individual by time t if the event were a repeatable process [12].

$$\begin{aligned} H(t) &= \int_0^t h(u)\,du \\ &= -\log S(t). \end{aligned} \tag{3.6}$$

3.1.1 Censoring in survival data

It is common to not observe precisely the starting and ending points of events in survival data, that is, the event of interest is not observed. This leads to censored data, meaning that the data has incomplete observation of event times where survival times will be unknown for a subset of the study group [32]. In survival analysis, there are three common forms in which data is censored; right censoring, left censoring and interval censoring. In this dissertation, the emphasis is that the data is right censored for both the clinical trial and simulated data sets.

Right censored data implies that a patient leaves the study before an event occurs (such as lost to follow-up, withdrawal/dropout, a competing event experienced making it impossible to follow up), or the study ends before the event of interest has occurred. *Left censoring* is when the true survival time has already occurred prior to observing the event of interest i.e. the time an event actually occurred cannot be observed. *Interval censoring* describes the situation where an event of interest did not occur at some time point t_1 but at a further time point t_2 the event is observed to have occurred. All that is known is that the event occurred between t_1 and t_2, but with no knowledge of the exact timing of the event. Left censoring and interval censoring are described in more detail elsewhere [12; 17; 32; 46].

Censoring needs to be taken into account when analysing events of interest to avoid creating bias in the model. There are different censoring mechanisms, of which most assume *non-informative censoring* [46], where the censoring mechanism is independent of the event process. However, the details of these mechanisms will not be discussed in this dissertation. This dissertation will assume that the censoring is

non-informative. Additionally, the assumption that censoring is generated truly at random is used, since the case study deals with patient data and the start time of the disease cannot be prespecified.

If T^* is a random variable representing time to an event and U is a random variable representing the censoring time [55], $T = \min(T^*,\ U)$ is what is observed, with a censoring indicator;

$$\delta = I[T^*, U] = \begin{cases} 1 & \text{censored time} \\ 0 & \text{failure time} \end{cases}$$

Censoring underestimates the true (but unknown) failure time of an event. Therefore, standard methods of data exploration and analysis lead to misleading results, hence special methods must be employed [12].

3.1.2 Analysis of survival data

The analysis of time-to-event data can be characterised into three main categories; parametric, semi-parametric and non-parametric methods [17; 55]. The objectives of a survival analysis are describing the survival experience and modelling the survival time. Kaplan and Meier [41] introduced a descriptive non-parametric estimator for survival probabilities which produce survival curves. The Kaplan-Meier curve considers the survival time of subjects whose event of interest is yet to be observed e.g. death, while accounting for some forms of censored data. This enables one to estimate the survival rate at different survival times before the event of interest. However, it cannot estimate the survival function adjusted for covariates [41; 64]. The Kaplan-Meier estimate for survival function is given as;

$$S(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right), \tag{3.7}$$

where d_j is the number of failures at a time t_j and n_j is the number of subjects alive just before t_j.

Mantel [50] introduced the log rank test statistic that quantifies and tests estimates of hazard functions by comparing two groups at each observed time, and which can be extended to a comparison of more than two groups [17]. In broad terms, this test statistic compares survival distributions and is like a modified chi-squared test [44],

stated simply as;

$$\chi^2_{LR} = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B} \sim \chi^2_{(1)}, \tag{3.8}$$

where χ^2_{LR} is the test statistic for the log-rank test. O and E are the sums of observed and expected events in a particular group, respectively. For k number of groups, the statistic can easily be generalised if the summation is extended to cover more than two groups with $k-1$ degrees of freedom.

However, with the advancement in technology, time-to-event data tends to be large in volume, hence it is vital to assess the importance of certain explanatory factors also known as covariates such as sex, race and age in predicting the survival outcome when analysing models [60]. Cox [15] then introduced the Cox PH model, making significant contributions to the development of the field [23].

Additionally, the set of most influential variables on failure time will need to be identified, with the objective of selecting a small set of covariates that best predict the outcome of interest. The problem now becomes that of choosing one model from among a set of candidate models. To achieve this objective, variable selection techniques are employed with the aim of controlling the removal of redundant or non-informative variables from the model which describes the data, thus facilitating practical interpretation, which can enable sound decision-making. This dissertation will focus on variable selection in Cox PH models.

3.2 The Cox proportional hazard model

The Cox PH model allows for the inclusion of covariates for the subjects into a model [15]. These models are the most popular form of semi-parametric models, which estimate relative risk, with no knowledge assumption of absolute risk [31]. They are used to assess the effects of several risk factors/covariates on survival and the technique resembles that of regression analysis, with a vital difference in that the outcome variable of interest (*time*) is always non-negative and most often censored [55]. Hosmer et al [32] ascertain that the response variable *time* is from an unambiguously defined origin (that is, when the "clock starts") to the occurrence of a well-defined event (that is, when the "clock stops").

The Cox PH model is the standard method for analysing the relationship of a time-to-event variable to one or more covariates [55]. It adjusts for the impact of the

other covariates in a regression. Cox [15] rendered the model as:

$$h_i(t) = h_0(t)\, e^{\boldsymbol{\beta}^T \boldsymbol{X}}, \tag{3.9}$$

where $h_i(t)$ is the hazard function for an individual i at risk at time t, with a vector of covariates $\boldsymbol{X} = (x_{i1},\ \ldots,\ x_{ip})$ assumed to be fixed over time with an associated vector of unknown parameters $\boldsymbol{\beta}^T = (\beta_1,\ \ldots,\ \beta_p)$ which measure the impact/effect size of the covariates. The assumption that the effects of the different covariates on survival are constant over time renders the ratio of the hazards to be proportional.

The baseline hazard denoted by $h_0(t)$ is an arbitrary non-negative function of time that describes the risk for individuals when $x = 0$. The shape of the baseline hazard is not specified, therefore no assumption of the distributional form of $h_0(t)$ is made, making the model semi-parametric. The estimate for the intercept is absorbed in $h_0(t)$, thus it is usually not estimated for inferential purposes of the effects of the explanatory variables on the relative hazard. Therefore, the survival function cannot easily be estimated but the hazard ratios can be estimated [17].

The quantity $e^{\boldsymbol{\beta}^T \boldsymbol{X}}$ is the relative risk or hazard ratio, and it captures the proportionate decrease or increase in risk associated with a given set of covariates. For instance, a hazard ratio above 1 signifies a covariate that is positively associated with the event probability, and thus having an inverse relationship with the length of survival, i.e. increased risk of the event of interest. A ratio less than 1 associates the covariate with improved survival, whilst one close to 1 is associated with no effect on survival. Loosely, the hazard function can be interpreted as the risk of observing an event (such as death) at time t.

The model can be re-expressed as a linear model for the logarithm of the hazard ratio, in the form:

$$log\left(\frac{h_i(t)}{h_0(t)}\right) = \beta_1 X_1 + \beta_2 X_2 + \ \ldots,\ + \beta_p X_p, \tag{3.10}$$

where β refers to the log of the hazard ratio associated with a 1 unit increase in the continuous X.

3.2.1 Partial likelihood function of the Cox PH model

The β-coefficients can be estimated using the method of maximum likelihood. Since the baseline hazard may not be specified when the main interest is the regression parameters [15], then only the likelihood for a description of part of the data is considered, hence using partial likelihood estimation.

However, there are ways in which the functional form of the baseline hazard can be specified fully in a parametric approach, but this requires strong assumptions about the form of the underlying survival distribution. Where the form of the baseline hazard can be assumed to be constant in each time interval a piece-wise exponential model in a semi-parametric setting is assumed. An in-depth discussion of these settings is beyond the scope of this dissertation but can be found in Cox [15; 16].

The partial likelihood depends only on the ranking of the actual death times and hence inference about the explanatory variables depends only on the rank order of the survival times [17]. Therefore, for ordered death times, the probability that the failure occurs to i at t_j given the risk set $R_{(t_j)}$ is of the form [15]:

$$L_j(\beta) = \frac{\exp\left(\beta' x_{(j)}\right)}{\sum_{I \in R_{\left(t_{(j)}\right)}} \exp\left(\beta' x_I\right)}. \tag{3.11}$$

For r distinct, unique death times with $x_{(j)}$ vector of explanatory variables for the individual that dies at time $t_{(j)}$, the partial likelihood function becomes:

$$L(\beta) = \prod_{j=1}^{n} \frac{\exp\left(\beta' x_{(j)}\right)}{\sum_{I \in R_{\left(t_j\right)}}\right) \exp\left(\beta' x_l\right)}. \tag{3.12}$$

Taking into consideration right-censoring of an individual i at time t_i with an event indicator of δ_i for n observed survival times, the partial likelihood function can then be expressed as:

$$\prod_{j=1}^{r} \left\{ \frac{\exp\left(\beta' x_{(j)}\right)}{\sum_{I \in R_{(t_i)}}\right) \exp\left(\beta' x_l\right)} \right\}^{\delta_i}. \tag{3.13}$$

It is computationally convenient to maximise the logarithmic of the likelihood function, which corresponds to:

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left\{ \boldsymbol{\beta}' \boldsymbol{x}_i - \log \sum_{l \in R(t_i)} \exp\left(\boldsymbol{\beta}' \boldsymbol{x}_l\right) \right\}. \tag{3.14}$$

Maximisation of the log-likelihood function is achieved using the Newton-Raphson procedure. This can be implemented in a computer software like R or STATA using the necessary package.

3.2.2 Model checking procedures for the Cox PH model

The appropriateness of a fitted Cox PH model needs to be checked to validate the statistical estimation and inference. The assumption of proportionality of any Cox PH model first must be tested before any inferences are made [15; 17; 31; 32]. In a standard Cox PH model the assumption is that all values of covariates are measured only at baseline (at entry) - no updating of values or repeated measurements over time. Then, the subjects are observed for the duration of the study to see if the response variable occurs or not [12].

The standard Cox PH model has no time-dependent covariates, thus it assumes that the hazard ratio (covariate effect) is constant over the entire study period. That is to say, the hazard of the event in any group is a constant multiple of the hazard in any other. This leads to the key assumption of a Cox model that the hazard curves for the patient groups should be proportional and hence cannot cross i.e. the hazards ratio for any two patients is constant over time. This is known as the constant-time assumption or the proportional hazards function assumption (PH assumption) [17]. This kind of model is known to be easy to interpret for the average user.

Furthermore, it is essential to examine the correctness of the functional form of continuous covariates being used and to check whether any important covariates have been omitted. One also needs to check if any observations are influential or outlying. Procedures used for checking the validity of a model are called diagnostic methods. Most Cox PH model diagnostic methods for assessing goodness-of-fit are based on residuals [17; 81]. In literature, various residuals and methods have been defined for Cox modelling diagnostics but the common ones are outlined below.

To assess the validity of the proportionality assumption, the Schoenfeld and score residuals based on the chi-squared goodness-of-fit test statistic is used. Other diagnostics for checking the proportionality assumption are the Cox-Snell residual and the martingale residuals. Graphical methods are also used but are known to work well when applied to categorical covariates with few levels [17; 81]. The plot of martingale residuals can be used for assessing the functional form of covariates, but works well when the covariate of interest is not correlated with the other covariates.

Plotting the martingale residuals against the risk score or linear prediction can help in identifying outlying observations. One can also use deviance residuals in identifying outliers. Lastly, score vector U can be used to assess influential observations [28]. However, care needs to be taken as to whether the influence is on regression coefficients or overall influence before applying any remedy on the influential observations [81]. This dissertation will focus on assessing overall model fit using Cox-Snell residuals.

3.2.3 Extensions of the standard Cox PH model

In some instances, the constant-time assumption does not hold, hence flexible models should be employed to avoid erroneous inference which lead to misleading conclusions. One may consider modelling with time-covariate interactions, or functions of time, which lead to a changing hazard ratio over time. Conversely, the stratified Cox PH model stratifies on the covariate that fails to meet the proportionality assumption while including covariates that satisfy the assumption.

Other models include the accelerated failure time (AFT) and the extended Cox model and other advanced models that deal with competing risks and multistate models, time-dependent models and frailty models [67]. However, as stated earlier, this dissertation will only consider the standard Cox PH model.

3.3 Previous model selection comparative studies

Where the PH assumption holds, the main concern lies in choosing covariates that will lead to a final best fit Cox PH model. Hence, model selection techniques are employed for this process. Model selection helps minimise the issue of data-dependent model building, such as overfitting the data or having biased estimates of parameters through model misspecification. The Cox PH model is not exempt from model misspecification because of its linear assumption. This linear assumption implies that each covariate makes a linear contribution to the model. It is therefore imperative that a useful model should correctly contain the important covariates that are relevant to the problem setting and give valid estimated effects.

There have been extensive studies in literature on model selection techniques for various statistical models [19; 82], most of which have been extended to survival analysis. The aim of these techniques is to propose combinations of covariates that might assist in predicting patient survival.

Peterson and Sehlstedt [60] found that the best subset performs better in selection than the all subset selection, but performs poorly computationally against the least absolute shrinkage and selection operator (Lasso). However, the results of the former are more systematic than the latter. Ekman [22] found that by simulating survival data, the Lasso performed equivalently or slightly better than the stepwise technique when the data consists of weak effects but not so much vice versa.

There have been developments in techniques that focus on optimising predictive performance such as the Stepwise Tuning in the Maximum Concordance Index (STMC) with Forward Nested Subset Selection (FNSS) in two stages [11]. In dealing with the issue of low accuracy and applicability which the Cox PH model brings when

working with high-dimensional covariates, Zhao et al [83] proposed the use of a technique based on both the least absolute shrinkage and selection operator (LASSO) and the Coordinate Decent Algorithm (CDA), termed the LASSO-CDA. Low predictive ability of a model acts in such a way that patients with similar measured covariates exhibit larger variability in their survival.

To conduct model selection in survival analysis data sets with many covariates, other authors have further proposed algorithm configuration approaches like Iterated-racing [45], Sure Independence Screening (SIS) based on marginal correlation ranking and Iterative ISIS where some predictors are marginally uncorrelated with the response [25]. The (I)SIS can be used for data sets which have more covariates than observations. It filters unimportant covariates through conditional marginal utility, then does selection via the penalised partial likelihood method which further filters out unimportant covariates [25].

Chapter 4

Model selection techniques for the Cox PH model

In this chapter, the theory of model selection techniques in Cox PH model employed in this dissertation is discussed. This is followed by a description of the model selection performance evaluation methods covered in this dissertation.

4.1 Model selection techniques

The focus in this dissertation will be on the widely used model selection techniques. The techniques in question are the classical techniques (Stepwise, forward, backward elimination and best subset selection methods), information criteria techniques based on likelihood functions (AIC and BIC), shrinkage techniques (Lasso and Ridge) [25] and the random forest technique.

Guided by George Box's aphorism that "all models are wrong but some are useful" it is important to note that models are not the reality of the phenomenon under study rather an approximation [7]. This is largely attributed to limitation in resources to capture all covariates driving the response variable as well as a limitation in tools to analyse complex models. This was captured concisely by Von Neumann [78] in his statement that "truth ... is much too complicated to allow anything but approximations". However, there must be a balance between simplicity and complexity to avoid the incidence that Paul Valery [76] pointed out that "What is simple is always wrong. What is not is unusable", hence the emphasis to have parsimonious models and being alert to what is importantly wrong [7].

In a quest to find a parsimonious model, this dissertation will focus on the application and performance of the specified variable techniques. The techniques will be used

on the IMPI clinical trial data set and a simulated data. Analysis will be carried out in the R programming language using RStudio [72].

4.1.1 Standard model selection techniques

Classical techniques are a set of search algorithms used to pick a "good" model for a response variable on a basis of some measured covariates [21; 40]. Commonly known as automated variable selection methods, they are useful when the number of covariates is large and fitting all possible models is infeasible. They consider all possible subsets from available covariates to include in a model that best fits the data according to some criteria, and this reduces the dimensionality of the model. These criteria assign scores to each model and chooses the one with the best score. James et al. [38] has an in-depth description of the algorithms as well as the criteria of choosing the "best" model after the model searching, such as Adjusted R^2, cross-validated prediction error, AIC or BIC.

A challenge with these techniques is that they are discrete processes, hence tend to be unstable, computationally intensive and said to be greedy in their variable selection approach [48]. Greedy means that the algorithm chooses the most optimal element at any given step with the assumption that this approach will eventually lead to an optimised model, which might not always work [3]. Additionally, since these techniques were developed for normal linear models; hence care must be taken on interpretation as they don't directly translate to survival analysis.

4.1.1.1 Best subset selection

The best subset selection is one of the most popular variable selection techniques [48]. It iterates over all possible variable combinations to select the best one. It picks the subset of size k that gives the smallest residual sum of squares, for every $k \in 1, 2, .., p$, where p is the total number of available variables. However, it has a shortfall in its use of the residuals as a criterion for determining k which decreases as the number of variables are included in the model, which might influence the final model chosen. As the number of covariates increases, so does the computational complexity [48] as there are 2^p possibilities of selecting the best model [38] hence makes it infeasible. The branch-and-bound techniques used to circumvent the complexity as covariates increase have only been researched to work for least squares linear regression [38] but still result in models with low predictive power. Additionally, the impact of the variables excluded from the model on the response variable remains a mystery.

4.1.1.2 Stepwise selection

On the other hand, stepwise selection differs from best subset in that it restricts the number of possible subsets enumerated, making it a computationally efficient alternative. Stepwise technique refers to backward-forward stepwise selection. It considers which covariates to add or subtract depending on the type of stepwise method being used until there is a "best" performing model that lowers prediction error [38]. With the aid of substantive knowledge, at a chosen alpha level α, manually identify the best candidate final regression model by dropping the covariates with p-value $> \alpha$ one at a time until all regression coefficients are significantly different from 0 [13].

The backward elimination starts with a full model, calculates the information criterion at each step and then compares the subsets. It iterates until the information criteria does not improve, see Algorithm 1 [22].

Algorithm 1: Backward stepwise selection

1. Let M_p denote the full model with all p covariates.
2. For $i = p, \ldots, 1$:
   (a) Fit all i models that contains all but one of the covariates in M_i
   (b) Denote the best among these models $M_{\text{i-1}}$.
3. Select the best model from among $M_0, \ldots, M_p$ using either CV, AIC or BIC

Conversely, forward selection starts with the intercept model and adds on variables iteratively while calculating the information criterion for each subset [22].

Algorithm 2: Forward stepwise selection

1. Let M_0 denote the base model with no covariates.
2. For $i = 0, \ldots, p-1$:
   (a) Fit all $p - i$ models that add one extra covariate to M_i
   (b) Denote the best among these models $M_{\text{i+1}}$.
3. Select the best model from among $M_0, \ldots, M_p$ using either CV, AIC or BIC

In step 2, the model with the highest likelihood is often defined as the best. The multiple testing problem is not of concern since in the stepwise variable selection technique there is conditioning on other covariates in the regression model at each step [13]. However, these techniques might overlook a great model because it does not exhaust all possible subsets. Additionally, stepwise is known to usually result output biased coefficients (inflated) and p-values (deflated) [75] but are easy to implement and understand.

4.1.2 Information criteria measures

AIC [2] and BIC [69] provide a way of scoring a model based on its log-likelihood and complexity. They attempt to quantify both the model performance on the dataset and the complexity of the model i.e. they judge the quality of a model [40]. They are best used for comparing non-nested models other than providing information on how well a model fit. To avoid overfitting which affects the ability of a model to generalise through predictions, these models introduce a penalty term in relation to the number of parameters in a model.

4.1.2.1 Akaike information criterion (AIC)

The AIC is derived from a frequentist framework and for model selection, the model giving the smallest AIC over all sets of models is considered:

$$AIC = -2pl\left(y\middle|\hat{\beta}\left(y\right)\right) + 2k, \tag{4.1}$$

where k is the number of degrees of freedom used or the number of parameters included in the model and $pl\left(.\right)$ is the partial log likelihood. A detailed description of the AIC can be found in Stoicha and Selen [1; 70]. The model with the lowest AIC is considered as the one approximately closer to the truth. Hence, a covariate is added to the model if it contributes enough information to make up for the increased penalty. It penalises complex models less i.e. tends to select complex models.

Hurvich and Tsai [34] showed that the AIC may be drastically biased for the linear model in small samples. However, the dataset to be considered in this dissertation has a sample size of 1400.

4.1.2.2 Bayesian information criterion (BIC)

The BIC differs from the AIC in the way the penalty for including more variables behaves. An extension of the BIC to the Cox PH model can be found in Volinsky and Raftery [77], where a modification of the penalty term was proposed so that it is defined in terms of the number of uncensored events instead of the number of observations n [47].

$$BIC = \ -2pl\left(y\middle|\hat{\beta}\left(y\right)\right) + In\left(d\right)k, \tag{4.2}$$

where d is the "effective" sample size i.e. the number of events for censored survival data. k is the number of parameters included in the model.

It penalises models with more parameters to a greater extent. Considering that $In\,(d) > 2$ for any $d > 7$ BIC is known to be harsher on adding more coefficients than AIC, thus tends to choose parsimonious models [55].

Bootstrap method

As proposed by Efron, to assess the variability of test statistics, the bootstrap is used [20; 75]. With studies showing the challenge of assessing the instability of automated model selection methods which identify noise variables as independent predictors, it is vital to include a method that repeatedly draws with replacement from the original dataset and assesses the strength of the identified independent variables [68]. Bootstrap also asists with model prediction which is an integral part of the model building process [63].

Therefore, bootstrap method can be employed to assist in assessing the distribution of an indicator variable denoting the inclusion of a specific predictor variable in a model to obtain a parsimonious model using stepwise methods [27; 68]. The expectation is that that noise variables will be identified as independent predictors in a minority of samples. One such bootstrap algorithm is as specified below [22]:

Algorithm 3: Model selection with bootstrap

Let $V_1, V_2, \ldots, V_n$ be vectors containing time-to-event, censoring indicator and the p covariates for the n observations.

1. For $b = 1, 2, \ldots, B$:
   (a) Randomly generate with replacement a bootstrap sample $V_1^*, V_2^*, \ldots, V_n^*$ from the original sample.
   (b) Perform variable selection on the bootstrap sample using a stepwise method.
   (c) Construct a vector, I_b, of length p where;

$$I_b = \begin{cases} 1 & \text{if covariate i in model} \\ 0 & otherwise \end{cases}$$

2. Construct a vector, I_{sum}, where element i is the sum of all i_{th} elements of the vectors I_b, $b = 1, 2, \ldots, B$.
3. If element i of vector I_{sum} is larger than the pre-specified inclusion frequency, include covariate i in the final model.

The stability of a model can also be assessed by evaluating accuracy of estimated coefficients in bootstrap methods unlike in traditional selection techniques where spurious covariates may be included resulting in inferior prediction in an unseen data set. These automated techniques usually result into associated p-values that

vary sample to sample, and are biased downwards by decreasing at each run, owing to the fact that the p-value is a binary decision (yes/no), so might be unreliable [68]. In a bootstrap, for instance, if an independent predictor variable had half of the estimated coefficients as negative and half as positive in the bootstrap samples, then further investigation on that variable is a necessity.

In bootstrap, the correlation structure is usually of main interest because the issue of multicollinearity can cause correlated variables to compete for inclusion in a model while excluding the other in half of the models. This results in neither variable being identified as independent predictors leading to less precise results . Hence, it is vital to examine any collinear variables in the samples to avoid their faulty exclusion. It is thus of importance to think carefully on what one is bootstrapping on, as it is easy to leave out important covariates if bootstrapping on covariates [68]. If one is conducting a survival analysis bootstrap, the challenge would arise on how censoring is going to be handled. In this dissertation, bootstrap is used for the estimation of regression parameters and standard errors [5]. We employ a bootstrap method that performs stepwise model selection by AIC using the *boot.stepAIC* function in R. The function deals with the multicollinearity problem by simplifying the model without impacting much on the performance, whilst the amount of information lost due to this simplification is quantified by AIC.

4.1.3 Penalised Cox PH Regression

These methods are based on the Cox PH model adding a penalty term in the Cox partial likelihood function to control for over-fitting [44]. They fit a model with all covariates included reducing estimation variability unlike classical methods [38]. They then shrink (also known as regularisation) some coefficients of the model towards zero. This leads to a reduction in parameters' variance hence a parsimonious model, less sensitive to extreme variance, that generalises better [38].

Additionally, in working with biological data from the same subject, some covariates may be highly correlated because their functions might interact in the body or may belong to the same biological pathway [44]. Hence, these penalised models are aimed at addressing the problem of multi-collinearity, which left unchecked results in model estimates with large variance.

4.1.3.1 Ridge regression

Ridge regression adds the sum of the squared magnitude of coefficient values (L2 penalty) as penalty term to the loss function, enforcing them to be small. It keeps all variables in the model as the complexity of the model decreases in a continuous way, hence can lead to cumbersome models that are a challenge to interpret [40].

It minimises the sum of squared residuals as well as penalise the size of parameter estimates.

$$\begin{aligned} L_{ridge}\left(\hat{\beta}\right) &= \sum_{i=1}^{n}\left(y_i - x_i\hat{\beta}\right)^2 + \lambda\sum_{j=1}^{m}\hat{\beta}_j{}^2 \\ &= \left|\left|y - X\hat{\beta}\right|\right|^2 + \lambda\left|\left|\hat{\beta}\right|\right|^2 \end{aligned} \tag{4.3}$$

Solving the above penalized likelihood estimator minimization problem gives the analytical formula for the β's as:

$$\hat{\beta}_{ridge} = \left(X^T X + \lambda I\right)^{-1} X^T y \tag{4.4}$$

Note that, I is an identity matrix, λ is a non-negative *tuning parameter* imposed on the parameters and needs to be determined separately [60], with $\lambda\left|\left|\hat{\beta}\right|\right|^2$ as the shrinking penalty, m is the number of covariates. The larger the tuning parameter λ, the greater the impact of the penalty, hence the beta coefficients approach zero, and the impact of that variable on the response variable becomes insignificant. On the other hand, if λ is small, the penalty vanishes [38] and the output resembles that of ordinary least squares (OLS). Therefore, the choice of λ is vital to avoiding underfitting. However, ridge regression does not remove some covariates altogether, hence not a technique of choice for variable selection in case a large number of covariates observed.

4.1.3.2 Least Absolute Shrinkage and Selection Operator (Lasso)

Lasso aims at performing estimation and selection at the same time [84]. It penalizes the absolute values of the coefficients (L1 penalty) which can force some of them to be exactly zero. In comparison to Ridge, it penalises non-zero coefficients to the loss function i.e. it penalises the sum of their absolute values, which results in zeroed coefficients for high values of λ. The Lasso minimisation problem for survival data can be solved with numerical algorithms:

$$L_{\text{lasso}}(\hat{\beta}) = \arg\min_{\beta}\left\{-\sum_{\{i=1}^{n}\delta_i\left[x_i^T\beta - \text{In}\left(\sum_{j=1}^{m}e^{x_i^T\beta}\right)\right] + \lambda||\beta||_1\right\} \tag{4.5}$$

The δ_i is the censoring indicator, which is 1 if an event occurs and 0 otherwise. Unlike Ridge regression which works well when most covariates truly impact the

response, Lasso is known to work best when few covariates influencing the response as it yields sparse models – models that do not include all covariates [38]. The choice between the two is usually based on out-of-sample prediction error.

The Lasso is known to have biased estimates, which one can counteract by using the adaptive Lasso, a secondary stage of estimation, which is known to lead to more consistent variable selection. Hence is known to possess the oracle property – performs as well as if the true underlying model was known beforehand [84]. The adaptive Lasso (aLASSO), not to be explored further in this dissertation, simply adds weights to the penalty term of the Lasso minimisation problem. A strategy that is used to fit aLASSO penalized Cox model performs well under extraordinarily large data sets is the novel sequence linearisation divide-and-conquer DAC_{lin} selection strategy proposed by Wang et al [80]. This strategy is known to reduce the computational burden compared to common penalised techniques. This strategy gets a K subsets of the whole sample for which an estimator is obtained from one subset to optimise the partial likelihood (PL) . Then, the PL updates all subsets via one-step approximations [80].

Choosing a tuning parameter
Clearly, selection of a tuning parameter is critical to the quality of coefficient estimates in shrinkage methods. Introducing the penalty shrinks variance, and this comes at a cost of introducing some bias in the model through the choice of the tuning parameter. However,if the penalty reduces variance more than it does the introduced bias, the overall accuracy of the model will be improved.

Just as with classical methods; AIC, BIC or cross validation can be used for selecting the value of λ, with CV being the most popular choice in practice. Cross validation is used to determine a "good" λ, where a set of different λ values are tried and the one which minimizes the cross-validated error is chosen [38].

4.1.4 Random survival forest

Random forest (RF) is an algorithmic model fitting approach in the context of classification of which model selection is a subset [66]. The method is described briefly with the assumption that the reader is familiar with the basic ideas and terminology of tree-based methods. Unlike automated model selection techniques, it is said to have "no strong assumptions" made about the functional form of the model or the distribution of residuals, hence non-parametric [66]. It is completely data driven hence free from model assumptions [18] and subverts the limitation of most univariate regression of overfitting, multicollinearity, unreliable estimation of regression coefficients, inflated standard errors or convergence problems [79]. Multicollinearity occurs where one can accurately predict one covariate with significance from other

covariates.

This method incorporates bootstrapping and is used as an alternative to Cox PH model and is known to perform better in prediction, but the two methods can possibly be used "jointly". For time to event data, it uses a collection of decision trees to rank covariates by their importance, VIMP, which assists in identifying covariates of the response with predictive ability [36]. Non-predictive covariates are identified if they have zero or negative values. It's ability in representing complicated non-linear relationships between inputs and outputs makes it a good candidate to offer 'better' survival predictions [35]. Instead of working with a single random tree, which might be unstable in their predictive function, it draws B bootstrap samples from the data. Then for each bootstrap, a tree is grown without pruning, then the predictions are averaged to obtain a final prediction. This process makes the algorithm an ensemble method with survival trees [8]. A general description of the algorithm is as follows:

Algorithm 4: Random survival forests [36]

1. Draw B bootstrap samples from the original data with a certain percentage excluded as out-of-bag (OOB) data.
2. For each bootstrap sample, grow a survival tree, randomly selecting p candidate variables at each node of the tree. The candidate variable that maximizes survival difference between daughter nodes is used to split the node. To split the node, a criterion involving survival time and censoring status information is used.
3. Grow the tree to full size under the constraint that a terminal node should have no fewer than $d_0 > 0$ unique deaths.
4. Calculate a cumulative hazard function (CHF) for each tree, which is the Nelson–Aalen estimator for each tree. Average to obtain the ensemble CHF.
5. Using OOB data, calculate prediction error for the ensemble CHF.

In survival analysis, the term Random Survival Forests (RSF) is adopted. In terms of variable selection, literature shows that there are three main methods proposed under RSF called variable importance measures. Firstly, variable hunting, which is a forward stepwise regularization based variable selection method which uses minimal depth of maximal sub-trees to calculate variable importance. This method identifies and ranks the importance of tree pathways for their association with a survival outcome. Minimal depth measures variable importance in terms of the depth a variable first splits within a tree relative to the root node. Where a smaller depth is desirable as this implies the variable is more predictive [36; 37]. Effective and efficient for high-dimensional data, this method has shown to also work well for cancer data in identifying key pathways with a relatively small sample size [10; 79]. Chen et al

[10] found that this techniques outperformed standard well-known procedures like random-set [57], Fisher's exact test, Gene Set Enrichment Analysis (GSEA) [71] and Pwayrfsurvival [59].

Secondly, the Iterative Feature Elimination (IFE) which can incorporate multivariate correlations and does not require the user to set a cut-off for p-values and is mainly used on micro-array data. IFE is known to have the advantage of being able to identify a small set of genes while preserving the predictive accuracy for survival [79]. Thirdly, topological index selects important variables in their improper Bagging Survival Tree (iBST) using a topological index instead of performance. It is based on building bagging trees and can select the explanatory variables even in the presence of a high number of noise variables, but is computationally intensive because it uses permutation [52; 79].

As earlier pointed out, the challenge is basing the RF explicitly on the Cox PH models, especially in the aspect of how trees deal with missing values and time-varying effects [6]. Extensions have been developed for competing risks and unorthodox censoring [79]. Random forests are used to reduce the number of covariates that explain the response variable to improve survival prediction performance. It does not use Cox PH explicitly, rather Cox PH is run after the dimension of covariates has been reduced. It is proposed that it is an exploratory technique that majorly deals with very large data sets and is an area of active research for the analysis of survival data. Regardless of it being non-parametric, care should be taken in its use to avoid blind application [79]. A few analysis will be run using this technique to see what sort of results will be derived. The R package *Ranger* will be used.

4.2 Performance evaluation of survival models

Most of the model selection techniques are not "direct measures" of predictive power and/or performance. Thus, a need to have methods that can assess the performance of the models selected. Like model selection techniques, literature has a plethora of measures used to assess the performance of models and the development of new measures is an area that is actively being researched [62]. According to Englebert et al. [Englebert et al.], classical loss functions like L2 are not the best methods to use in assessing the predictive ability of a survival model. This is because, only the time at which an event actually occurs is known and not the curve which represents the probability of occurrence of an event for the data. For predictive modelling, cross validation (CV) is used as an estimate of predictive power/error [82]. CV introduces the concept of variance-bias trade-off and aims at achieving a "good" estimate for the performance of a model [63].

Kadane and Lazar [40] argues that cross validation can be used as a model selection technique. Zhang et al [82] argues that CV can be used across many of the model selection techniques above as an internal step in producing final estimator or identifying best candidate model or model selection technique for the data at hand. In the interest of time, CV was not employed as a model selection technique in this dissertation. It was otherwise used across some of the selection techniques as a statistical metric. It was used to choose the optimal model with the lowest predictive error in the shrinkage regression techniques. All measures considered in this dissertation were implemented in R using built-in functionalities.

4.2.1 Discrimination measures

To assess how well a model distinguishes between low risk and high risk patients [62], discrimination measures were used. Concordance index (C-index) and Integrated AUC (iAUC) are the measures of performance considered based on discrimination [11; 62].

4.2.1.1 Concordance index

Concordance is the probability of agreement that the observation with the shorter survival time of two randomly selected observations has the worst predicted outcome or larger risk score [36; 73]. For the predicted survival times, the ordering is evaluated, by quantifying the rank correlation between the predicted risk and the observed survival times. The C-index usually takes values between 0 and 1. A value of 0 suggests poor discrimination, 0.5 is a result from a random prediction and 1 suggests perfect discrimination [62]. The C-index is defined as,

$$\text{C-index} = \frac{\sum_{ij\epsilon\Omega} I\{t_i > t_j\}}{|\Omega|},$$

where Ω is the set of all pairs of patients $\{i, j\}$ and $I\{\}$ is the indicator variable such that,

$$I\{\} = \begin{cases} 1 & \{t_i > t_j\} \\ 0 & \text{otherwise.} \end{cases}$$

The C-index is popular, easy to interpret and robust [11], hence it will be used as one of the primary criterion to compare model performance in this dissertation. It also takes censoring into account [36].

4.2.1.2 Integrated Area Under the Curve (iAUC)

ROC curves are used to assess the predictive accuracy of a survival model. For a risk score R denoted as cutoff values c of a continuous variable with event status $D(t)$, the ROC curve is used to assess sensitivity and 1-specificity, defined as,

$$Sensitivity(c,t) = Pr\{R > c|D(t) = 1\}$$
$$Specificity(c,t) = Pr\{R \leq c|D(t) = 0\}$$

For individual i at time t the event status is denoted by;

$$= \begin{cases} D(t) = 1 & \text{if } T_i \leq t \\ D(t) = 0 & \text{if } T_i > t \end{cases}$$

The time-dependent area under the ROC curve AUC(t) can be summarised by the integrated area under the curve (iAUC), given by the area under the ROC(t) over event time [11]. iAUC=1 is desirable as it suggests perfect prediction accuracy, whilst iAUC=0.5 suggests a random guess over time.

4.2.2 Overall performance measures

Measures based on explained variation, explained randomness and coefficient of determination were explored ('R^2-type' measure). The measure of explained randomness in proportional hazards models used was Kent and O'Quigley's R^2_{PM} [42]. For explained variation for the Cox PH model Royston measure was used and Nagelkerke measure as the measure of the coefficient of determination [56; 65]. These measures are based on quantifying the improvement in likelihood and prediction accuracy between the fitted covariate model and null model. They usually take values between 0 and 1 with higher values indicating better model fit [62].

4.2.2.1 Coefficient of determination measure

This is a general measure of the dependence strength of the outcome on the covariates in a regression model defined as

$$R^2 = 1 - exp\Big\{ -\frac{2}{n}\Big(l_{\hat{\beta}} - l_0\Big)\Big\}$$
$$= 1 - exp\Big(-\frac{X^2}{n}\Big),$$

where n is the number of observations in the dataset and l_0 denotes the log partial likelihood for the null model and $l_{\hat{\beta}}$ is the the log partial likelihood for the model with covariates. Additionally, $X^2 = 2(l_{\hat{\beta}} - l_0)$ is the likelihood ratio statistic for comparing the covariate model with the null model distributed as $\chi^2_{dim(\beta)}$.

4.2.2.2 Explained randomness measure

O'Quigley, Xu, and Stare [58] noted that R^2 is negatively correlated with the proportion of censored observations and tends to 0 as that proportion tends to 1, for any given model and dataset. Therefore, they proposed using number of uncensored observations, e, for the denominator in R^2 instead of number of observations in a data set, to give

$$\begin{aligned} R^2_{mer} &= 1 - exp\Big\{ -\frac{2}{e}\Big(l_{\hat{\beta}} - l_0\Big)\Big\} \\ &= 1 - exp\Big(-\frac{X^2}{n}\Big). \end{aligned}$$

4.2.2.3 Explained variation measure

Explained variation and explained randomness are the same for linear models. However, for PH models the assumption of normal-errors used in linear models makes the concepts different and that of explained variation tricky, because explained randomness always exceeds explained variation for PH models [65]. Hence, Royston proposed a measure of explained variation that takes into account censoring and based on the measure of explained randomness to give

$$R^2_{mev} = \frac{R^2_{mer}}{R^2_{mer} + (\pi/6)(1 - R^2_{mer})}.$$

This measure adjusts the measure of explained randomness for the mild upward bias which can be caused when large censored events are observed [65].

Chapter 5

Results

In this chapter, the estimation results from the model selection and performance evaluation methods for the IMPI trial data set are outlined. Additionally, each model is checked for overall fit. Furthermore, the performance evaluation results for the simulation studies are given.

5.1 Cox modelling of IMPI clinical data set

In Chapter 2 we described the multicentre clinical trial data set. The model selection techniques under study in this dissertation are applied to the IMPI trial data set. Cox PH regression with no assumption of interaction or non-linearity was used to fit the models in conjunction with model selection techniques. The techniques used were classical regression (stepwise, forward, and backward selection), AIC with bootstrap, BIC, penalised Ridge and Lasso regressions, and random survival forest.

Of the covariates specified in Chapter 2, a univariate Cox PH analysis selected 18, 15 and 19 candidate covariates for time to composite, time to death and time to constriction events, respectively. For each model selection technique, models using the candidate covariates were fitted for each event of interest.

This section first covers the interpretation of some estimation results for the covariates under all selection techniques. As a summary, for illustration, only the estimation results for the time to composite event are specified in Tables 5.1 to 5.5. The comparative results for all response variables are also given in this section.

The selection techniques used complete-case data which was sampled randomly into a 75% training set and 25% testing set for performance evaluation. Regression models were run on the training set for each of the three time-to-event variables

using their respective covariates. The aim was to use the testing set for evaluating the iAUC performance measure.

5.1.1 Cox modelling of time to composite event

In this section we will look at the measure of effect, the hazard ratio, of some of the selected covariates onto the composite event. These are the regression coefficients which quantify the association between each of the covariates and the event of interest. They represent the change in the expected log of the hazard to a unit change in a continuous covariate, all other covariates being constant. Measuring effect sizes helps us know how much actual change a covariate makes on a response variable, hence it is vital for inferring the clinical importance of a covariate [30].

The confidence interval of the hypothesis tests will be used to assess statistical significance of the effect of covariates on time to composite. In the interest of time, only significant covariates will be interpreted. The confidence interval is interpreted as 95% of intervals which would have the true estimate if a study were repeated many times, and on each study a 95% confidence interval was calculated [30]. If an interval includes 1, then it suggests that it is not significant, if it does not then the covariate is statistically significant. Wide intervals suggest that the effect size might not be known precisely, as the margin of errors is large. Therefore, one may consider getting further information to establish whether a small sample size or variability in the data might be the determining factor for the wide intervals [30].

Naive Cox PH model estimation results

We fitted a model which did not employ any selection technique and included all covariates, the full model. Estimation results are given in Table 5.1 with their corresponding confidence intervals.

The confidence intervals of Table 5.1 show that not all covariates make a significant contribution to the hazard of the event of interest except *creatanine* (95% CI: 1.149 - 2.289), *weight* (95% CI: 0.789 - 0.981), *definite TB pericarditis status* (95% CI: 1.109 - 2.092) and *peripheral oedema* (95% CI: 1.091 - 1.840). You will notice that the NYHA Class at study entry covariate is statistically significant for 2 categories but one. The whole covariate will be included as one cannot just exclude only the categories having insignificant difference. Excluding a category would combine the insignificant level with the reference level, hence changing the interpretation or would not make sense depending on the nature of the covariate.

Table 5.1: Naive Cox PH model estimation results

Variable	HR (95% CI)	p-value
Prednisolone	0.943 (0.735 - 1.211)	0.647
Pericardiocentesis at randomisation (Yes)	0.756 (0.549 - 1.040)	0.086
Duration of symptoms (days)	1.003 (0.999 - 1.006)	0.169
Age (years)	1.006 (0.995 - 1.016)	0.278
NYHA Class at study entry:		
II	1.429 (0.919 - 2.221)	0.113
III	2.416 (1.504 - 3.881)	0.000
IV	2.829 (1.651 - 4.849)	0.000
Creatanine (≥105)	1.622 (1.149 - 2.289)	0.006
HIV status (HIV+)	0.894 (0.656 - 1.219)	0.479
Weight (kg)	0.879 (0.789 - 0.981)	0.021
Palpable pulsus paradoxus (Yes)	0.866 (0.626 - 1.198)	0.386
Systolic blood pressure (≥90)	0.681 (0.459 - 1.009)	0.056
Heart rate (≥ 100)	1.141 (0.876 - 1.487)	0.328
Effusion size:		
Medium (1-2cm)	0.998 (0.713 - 1.397)	0.991
Small (<1cm)	0.623 (0.335 - 1.161)	0.137
Chest Xray pulmonary infiltrate (Yes)	1.187 (0.915 - 1.539)	0.196
Definite TB pericarditis status (Yes)	1.523 (1.109 - 2.092)	0.009
Haemoglobin (≥ 10)	1.299 (0.988 - 1.708)	0.061
Peripheral oedema (Yes)	1.417 (1.091 - 1.840)	0.009

Holding the other covariates constant, for a patient with creatanine $\geq$ 105, the expected risk of either death or constriction or tamponade is increased by a factor of 1.622 compared to a patient with creatanine $<$ 105. There is an increased risk of the composite event of 52.30% (HR = 1.523) in patients with definite TB pericarditis compared to those without it. Similarly, for patients with Peripheral oedema the expected risk of the composite event is increased by a factor of 1.417 compared to those without it.

For patients in category II of NYHA class at study entry - though not making a significant contribution to hazard - there is an increased risk of the composite event of 1.429-fold compared to class I patients. The same is true for patients in class III and class IV with a significant 2.416-fold and a 2.829-fold increased expected risk, respectively, compared to class I patients. One may consider carrying out tests like the likelihood ratio test to check whether NYHA class should be included in the model or not by building two models, one with and one without it. Conversely, for a unit increase in weight the risk of the composite event is reduced by 12.3%. This suggests that having more weight is associated with good prognosis.

Classical techniques

Table 5.2 contains the estimation covariate results for the stepwise, forward selection and the backward selection techniques. In the stepwise selection model, all covariates are significant but *pericardiocentesis* and NYHA class II. The interpretation of the

categorical covariates is as for that in the naive model with all covariates, except for systolic blood pressure (BP). Note that normal systolic BP reading ranges between 90 and less than 120. The datafile has systolic BP with mean 113.3 and a standard deviation of 17. Patients with systolic blood pressure ≥ 90 are suggested to have a small hazard of 0.678 (32.2% decreased risk) of the composite event compared to those with systolic blood pressure < 90. We also see that there is a 12.1% decrease in expected risk of the composite event for a unit increase in weight for patients. These results suggest that higher weight and systolic blood pressure are good prognostics.

Forward selection technique selected similar covariates as the stepwise selection technique in the same table (Table 5.2). The HR results can be interpreted in the same manner. The results suggest that the risk of the composite event is decreased by 12.1% (95% CI: 0.789 - 0.981) per unit increase in weight for patients. The risk of the composite event for patients with greater than 90 systolic blood pressure decreases by 31.9% compared with patients with lower than 90 blood pressure. All the other covariates increase the risk of the event compared with their respective reference levels.

Table 5.2: Estimation results for classical Cox PH model selection techniques

	Stepwise model selection		**Forward model selection**		**Backward model selection**	
Variable	**HR (95% CI)**	**p-value**	**HR (95% CI)**	**p-value**	**HR (95% CI)**	**p-value**
Pericardiocentesis at randomisation (Yes)	0.766 (0.571 - 1.028)	0.076	0.756 (0.549 - 1.040)	0.086	-	-
NYHA Class at study entry						
II	1.476 (0.954 - 2.285)	0.080	1.429 (0.919 - 2.221)	0.113	1.649 (0.988 - 2.762)	0.056
III	2.356 (1.489 - 3.726)	0.000	2.416 (1.504 - 3.881)	0.000	2.177 (1.264 - 3.759)	0.005
IV	2.716 (1.624 - 4.543)	0.000	2.829 (1.651 - 4.849)	0.000	2.662 (1.457 - 4.857)	0.001
Creatanine (≥ 105)	1.616 (1.153 - 2.266)	0.005	1.622 (1.149 - 2.289)	0.006	1.668 (1.139 - 2.440)	0.009
Weight (kg)	0.879 (0.791 - 0.977)	0.017	0.879 (0.789 - 0.981)	0.021	-	-
Systolic blood pressure (≥ 90)	0.678 (0.463 - 0.995)	0.047	0.681 (0.459 - 1.009)	0.056	0.513 (0.346 - 0.768)	0.001
Definite TB pericarditis status (Yes)	1.579 (1.154 - 2.159)	0.004	1.523 (1.109 - 2.092)	0.009	-	-
Haemoglobin status (≥ 10)	1.377 (1.069 - 1.775)	0.013	1.299 (0.988 - 1.708)	0.061	-	-
Peripheral oedema (Yes)	1.457 (1.130 - 1.878)	0.004	1.417 (1.091 - 1.840)	0.009	1.474 (1.104 1.971)	0.001

The results for the backward selection technique contained in Table 5.2 suggest that all covariates are statistically significant except for the NYHA class II level, as was observed for the estimation results of the other classical selection techniques. The estimation results suggest that patients with systolic blood pressure ≥ 90 have a hazard reduction of 48.7% - decreased expected risk of the composite event - compared to patients with a lower blood pressure.

Information criteria selection

Table 5.3 gives the estimation results for AIC and BIC. For the AIC technique, pericardiocentesis, systolic blood pressure and NYHA class II are not statistically significant, and only NYHA class II is not significant for the BIC technique.

Table 5.3: Estimation results for information criteria Cox PH model selection techniques

	AIC model selection		**BIC model selection**	
Variable	**HR (95% CI)**	**p-value**	**HR (95% CI)**	**p-value**
Pericardiocentesis at randomisation (Yes)	0.717 (0.512 - 1.005)	0.053	-	-
NYHA Class at study entry				
II	1.326 (0.824 - 2.134)	0.245	1.392 (0.866 - 2.237)	0.172
III	2.158 (1.308 - 3.559)	0.003	2.138 (1.299 - 3.518)	0.003
IV	2.535 (1.433 - 4.485)	0.001	2.454 (1.392 - 4.328)	0.002
Creatanine ($\geq$ 105)	1.599 (1.086 - 2.354)	0.017	1.635 (1.111 - 2.407)	0.013
Weight (kg)	0.824 (0.727 - 0.935)	0.003	0.817 (0.719 - 0.928)	0.001
Systolic blood pressure ($\geq$ 90)	0.727 (0.472 - 1.120)	0.148	-	-
Pulmonary infiltrate (Yes)	-	-	1.499 (0.667 - 2.001)	0.006
Definite TB pericarditis status (Yes)	1.584 (1.107 - 2.267)	0.012	1.418 (1.020 - 1.972)	0.038
Haemoglobin status ($\geq$ 10)	1.368 (1.019 - 1.836)	0.037	1.391 (1.035 - 1.871)	0.029
Peripheral oedema (Yes)	1.446 (1.077 - 1.940)	0.014	1.406 (1.049 - 1.886)	0.023

The results suggest that under the AIC technique, patients with creatanine $\geq$ 105 have a 1.599-fold increased risk of the composite event compared to those with creatanine $<$ 105. For patients with haemoglobin $\geq$ 10, the risk increases by a factor of 1.368. The risk also increases by a factor of 1.584 for patients with definite TB and by 1.446 for patients with peripheral oedema compared to those without these covariates. However, we observe that the risk of the composite event decreases by 27.3% (hazard factor of 0.727) for patients with systolic blood pressure $\geq$ 90 compared to those with $<$ 90. We also observe that the risk of the event is decreased by 17.6% (hazard factor of 0.824) for patients adding more weight. The effect sizes (HRs) of the BIC technique can be interpreted in the same manner.

Shrinkage methods

Table 5.4 gives estimation results for ridge and lasso regression models. For ridge, despite weight being non-significant, the risk of the composite event is decreased by 11.1% for a unit increase in weight among patients. Patients with systolic blood pressure $\geq$ 90 have a decreased risk of the composite event of 42.1% compared to those with blood pressure less than 90. On the other hand, all the other selected covariate estimates suggest an increase in the risk factor compared to their respective reference levels.

Table 5.4: Shrinkage Cox PH model selection techniques

Variable	Ridge regression Cox PH model selection HR (95% CI)	p-value	Lasso regression Cox PH model selection HR (95% CI)	p-value
Pericardiocentesis at randomisation (Yes)	-	-	0.948 (0.671 - 1.341)	0.764
NYHA Class at study entry				
II	1.630 (0.974 - 2.728)	0.063	1.598 (1.003 - 2.546)	0.048
III	2.161 (1.253 - 3.727)	0.006	2.603 (1.575 - 4.303)	0.000
IV	2.560 (1.399 - 4.682)	0.002	2.977 (1.644 - 5.389)	0.000
Creatanine (≥ 105)	1.739 (1.182 - 2.560)	0.005	1.571 (1.076 - 2.294)	0.019
Weight (kg)	0.889 (0.787 - 1.004)	0.059	-	-
Systolic blood pressure (≥ 90)	0.579 (0.385 - 0.872)	0.009	-	-
Definite TB pericarditis status (Yes)	1.324 (0.955 - 1.835)	0.092	-	-
Peripheraloedema (Yes)	1.478 (1.106 - 1.974)	0.008	-	-
Palpableparadoxus (Yes)	-	-	0.968 (0.674 - 1.392)	0.862
Effusion size				
Small ($\leq 1cm$)	-	-	1.111 (0.771 - 1.601)	0.571
Medium (1 - 2cm)	-	-	0.965 (0.519 - 1.796)	0.910

We observe that out of the five covariates selected by lasso, three are not statistically significant except NYHA class and creatanine. The risk of the composite event is increased by a factor of 1.571 for patients with creatanine $\geq$ 105 compared to patients with less than 105 creatanine. We observe that the risk increases with each level of NYHA class under lasso compared to reference level class I. We also observe that NYHA class II is statistically significant for the first time for a model selection technique.

Random survival forest

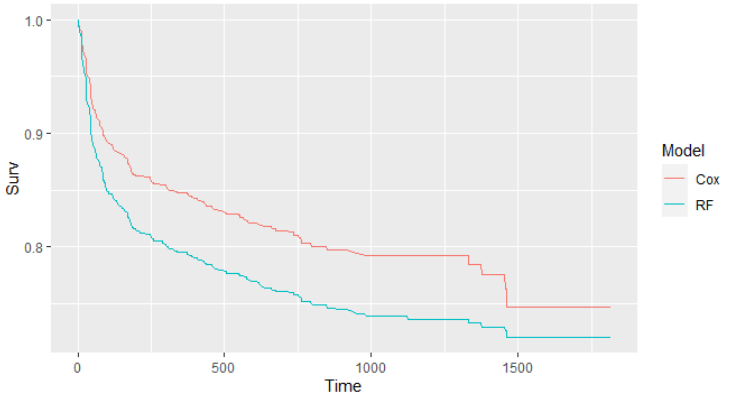


Figure 5.1: Comparison of full Cox PH model and RSF survival curves

Figure 5.1 gives a quick comparison of two survival curves. We observe that the full Cox model curve suggests a better chance of survival compared to the RSF survival curve. This might be attributed to the full Cox model taking into account more

covariates than the RSF survival curve. Thus, there is a call for further investigation on assumptions that can improve accuracy of the RSF model.

The ranking of the important variables under RSF for all events of interest are given in Figures 5.2 to 5.4.

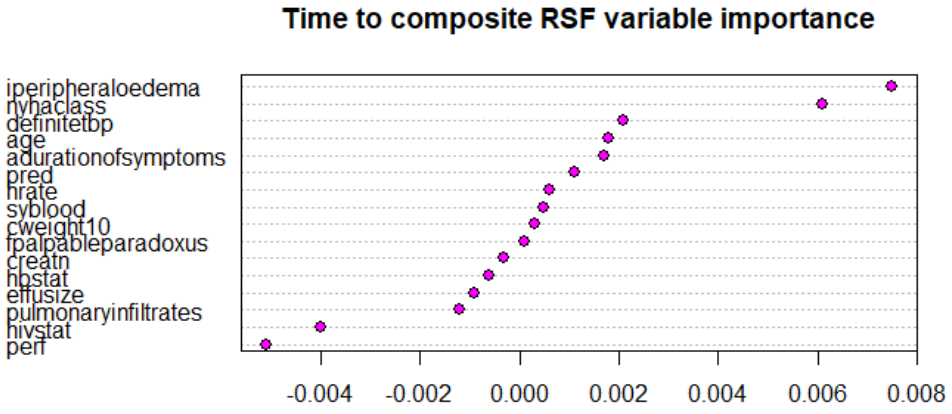


Figure 5.2: Variable importance plot for the composite event

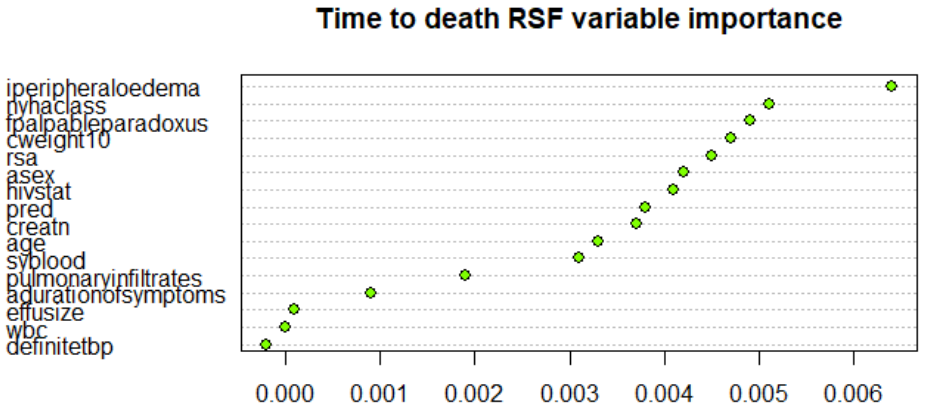


Figure 5.3: Variable importance plot for the death event

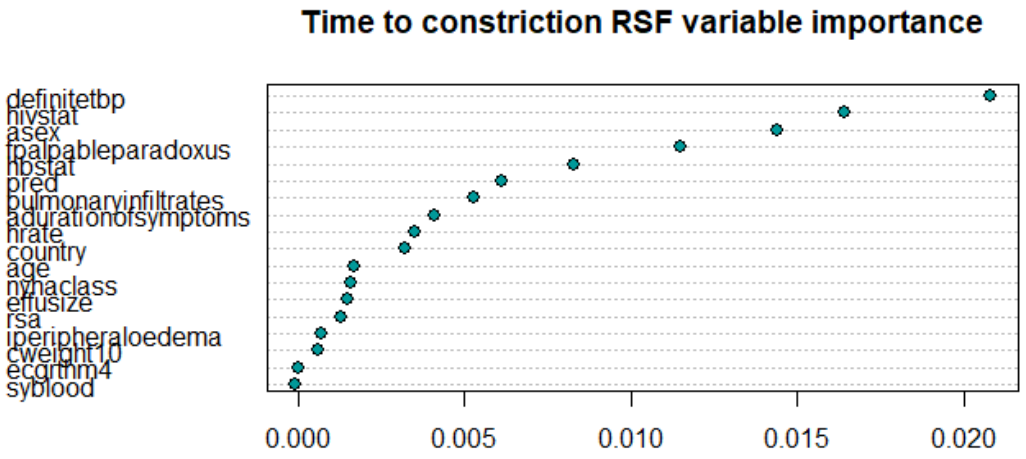


Figure 5.4: Variable importance plot for the constriction event

Table 5.5 gives the Cox PH model estimation results of the RSF covariates which were ranked by importance for the time to composite event.

Table 5.5: Random survival forest Cox PH model selection technique estimation results

Variable	**HR (95% CI)**	**p-value**
Prednisolone	1.042 (0.784 - 1.384)	0.778
Age (in years)	1.008 (0.997 - 1.019)	0.143
NYHA Class at study entry		
II	1.665 (0.992 - 2.795)	0.053
III	2.311 (1.315 - 4.061)	0.004
IV	2.945 (1.564 - 5.547)	0.001
Duration of symptoms	1.001 (0.997 - 1.005)	0.598
Weight (kg)	0.899 (0.793 - 1.019)	0.097
Palpableparadoxus (Yes)	0.855 (0.592 - 1.235)	0.405
Systolic blood pressure (≥ 90)	0.592 (0.386 - 0.908)	0.016
Heart rate (> 100)	1.031 (0.768 - 1.384)	0.840
Effusion size		
Medium (1-2cm)	1.323 (0.926 - 1.890)	0.124
Small ($< 1cm$)	0.671 (0.310 - 1.454)	0.312
Definite TB pericarditis status (Yes)	1.352 (0.970 - 1.885)	0.075
Peripheraloedema (Yes)	1.499 (1.115 - 2.014)	0.007

Of all these covariates, the 95% confidence intervals of NYHA classes III and IV

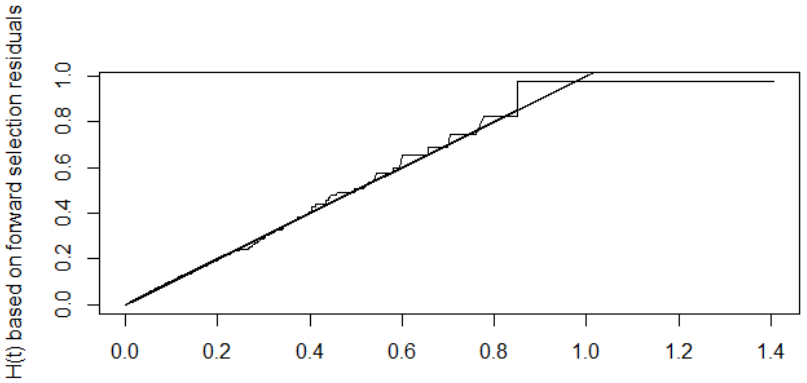
(1.315 - 4.061) and (1.564 - 5.547), Systolic blood pressure (0.386 - 0.908) and Peripheraloedema (1.115 - 2.014) show significance. We observe that the risk of the composite event is decreased by 40.8% (hazard ratio of 0.592) for patients with systolic blood pressure ≥ 90 compared to those in the < 90 category. However, the risk of the composite event is increased by 1.499-fold for patients with peripheral oedema compared to those without it.

For all selection techniques NYHA class was included in all selected models. One can also observe that in all cases, the NYHA class covariate gave the widest confidence intervals, suggesting that further information is needed for this covariate.

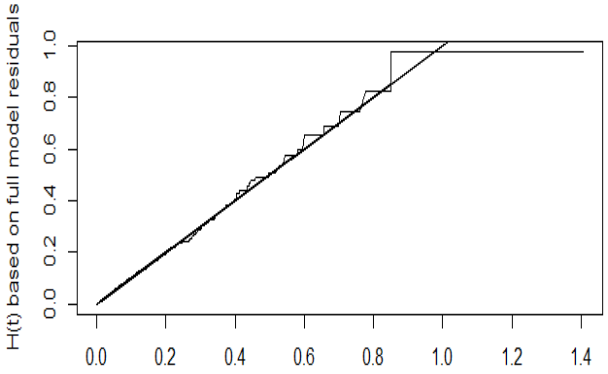
A more detailed analysis of the covariates selected by the models, which includes comprehensive model checking, is a call for future research as this is an important subject that merits more discussion than can be included in this dissertation.

5.1.2 Overall model fit

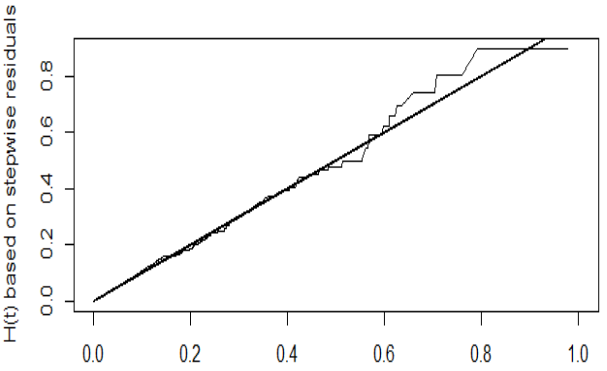
This section illustrates Cox-Snell residual plots which were used to assess the models for overall fit. These plots help in informing on whether a model is a good fit or not. Note that this dissertation does not include extensive diagnostic measures for model checking which is an area that should be explored further in future research.



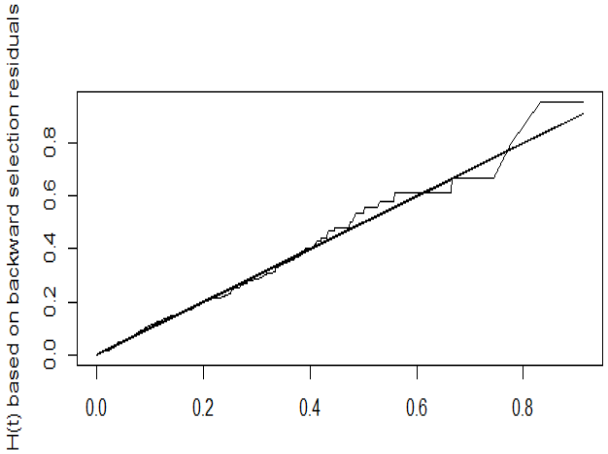
(a)


H(t) based on full model residuals
0.0 0.2 0.4 0.6 0.8 1.0
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4

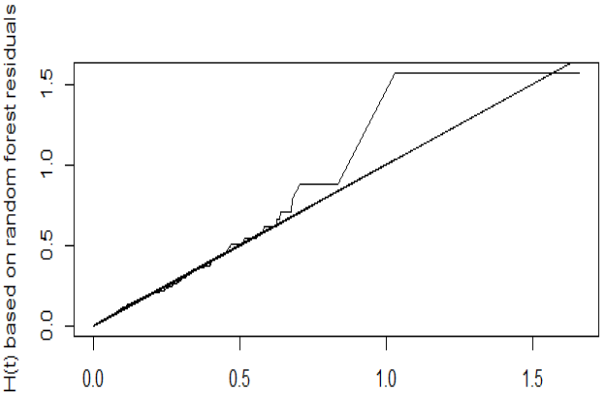
(b)


H(t) based on stepwise residuals
0.0 0.2 0.4 0.6 0.8
0.0 0.2 0.4 0.6 0.8 1.0

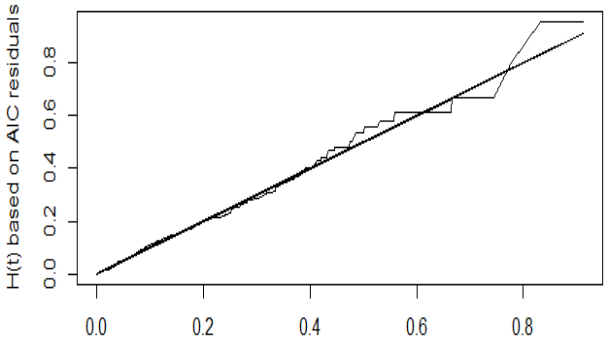
(c)


H(t) based on backward selection residuals
0.0 0.2 0.4 0.6 0.8
0.0 0.2 0.4 0.6 0.8

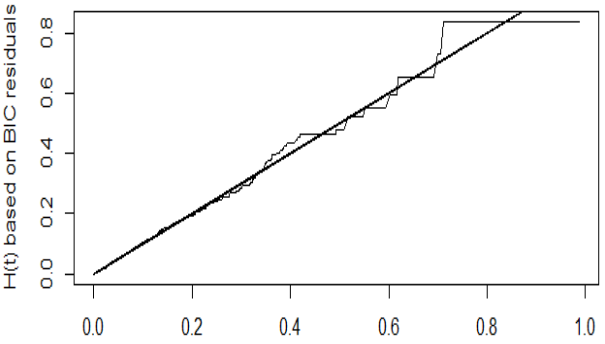
(d)


H(t) based on random forest residuals
0.0 0.5 1.0 1.5
0.0 0.5 1.0 1.5

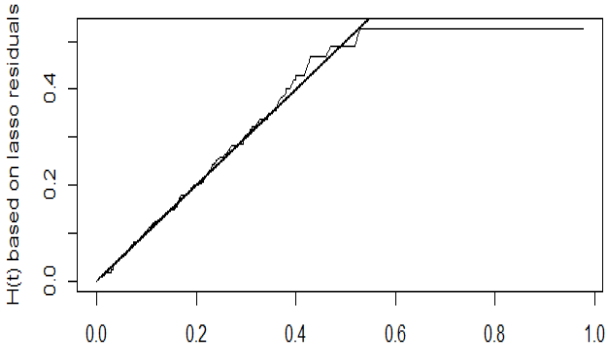
(e)



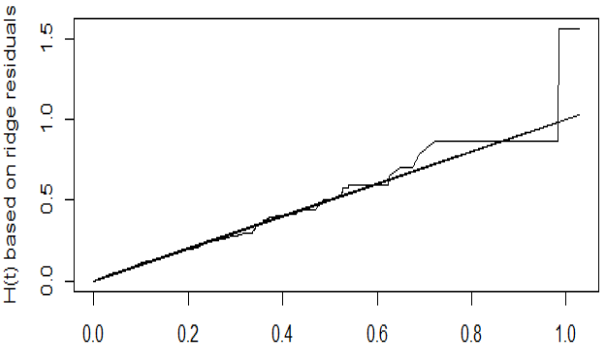
(f)



(g)



(h)



(i)

Figure 5.8: Composite event Cox-Snell residual plots for the IMPI clinical data set for each model selection technique. The residual plots correspond to (a) forward; (b) naive model; (c) stepwise; (d) backward; (e) random survival forest; (f) aic; (g) bic; (h) lasso; (i) ridge model selection.

The residuals are plotted on the x-axis, with the y-axis having the integrated hazard of the residuals, against a 45-degree line. If the model fits well, the curve of the

Cox-Snell residuals against the integrated hazard is expected to fall roughly on the 45-degree reference line.

The lines of the plots in Figure 5.8 do not deviate too much from the reference line for most models. Therefore, the Cox PH models do provide a reasonably good fit for the data. However, plot (e) of Figure 5.8 corresponding to random survival forest suggests otherwise, deviating slightly from the reference line. These results are in line with those given by the C-index where RSF had the smallest C-index of 0.417 suggesting poor performance.

5.1.3 Comparative study results

It is imperative to compare how the techniques performed, since different selection techniques were used. The C-Index, integrated AUC (iAUC) and three R^2 measures of overall performance were employed to measure the performance of the model selection techniques. These model performance methods were used purely to compare the accuracy of model selection techniques rather than assessing the predictive power of a final fitted model. This is because a clinical trial is a controlled study and researchers are generally interested in statistical inference e.g., knowing whether a treatment works or not [61]. The researchers achieve this by adjusting for covariates - assessing the covariates important in explaining the risk of an event - but not for prediction purposes.

Time to composite event

Table 5.6: Time to death, constriction or tamponade model selection techniques comparison results

	Performance Measure (n=1080)				
Selection Technique	C-index	iAUC	R^2	R^2_{mer}	R^2_{mev}
Naive	0.657	0.605	0.079	0.290	0.200
Forward selection	0.657	0.670	0.079	0.290	0.200
Backward selection	0.629	0.636	0.068	-	-
Stepwise selection	0.646	0.659	0.069	0.260	0.170
AIC	0.646	0.660	0.069	0.260	0.170
BIC	0.646	0.660	0.069	0.260	0.170
Lasso	0.614	0.618	0.038	0.150	0.100
Ridge	0.632	0.669	0.059	0.220	0.150
RSF	0.417	0.621	0.060	0.220	0.150

Table 5.6 gives the output of comparisons among the model selection techniques. The table shows that the full or naive Cox PH model without a selection technique

gave similar results to forward selection except under iAUC where forward selection gave relatively higher measure outputs than the other selection techniques. The forward selection technique performed better than the other techniques under all metrics, with an iAUC metric of 0.670. Hence, the final model for the composite endpoint is based on the iAUC metric and the estimation results are in Table 5.2.

Time to death

Table 5.7: Time to death model selection techniques comparison results

	Performance Measure (n=1080)				
Selection Technique	C-index	iAUC	R^2	R^2_{mer}	R^2_{mev}
Naive	0.667	0.584	0.065	0.310	0.210
Forward selection	0.667	0.674	0.065	0.310	0.210
Backward selection	0.617	0.619	0.046	-	-
Stepwise selection	0.657	0.656	0.058	0.280	0.190
AIC	0.657	0.656	0.058	0.280	0.190
BIC	0.657	0.656	0.058	0.280	0.190
Lasso	0.645	0.658	0.048	0.230	0.160
Ridge	0.647	0.655	0.055	0.270	0.180
RSF	0.667	0.584	0.065	0.310	0.210

Table 5.7 shows the comparative results for the model selection techniques. Based on all the comparative methods, the forward selection gave relatively high measure outputs, with the naive Cox PH model and random survival forests performing similarly. However, only the iAUC metric made a single selection of the best performing technique: the forward selection technique.

Time to constriction

Table 5.8: Time to constriction model selection techniques comparison results

	Performance Measure (n=1080)				
Selection Technique	C-index	iAUC	R^2	R^2_{mer}	R^2_{mev}
Naive	0.770	0.833	0.060	0.640	0.520
Forward selection	0.770	0.768	0.060	0.640	0.520
Backward selection	0.753	0.723	0.048	0.560	0.430
Stepwise selection	0.753	0.723	0.048	0.560	0.430
AIC	0.753	0.741	0.053	0.590	0.470
BIC	0.741	0.717	0.044	0.530	0.400
Lasso	0.733	0.738	0.044	0.530	0.400
Ridge	0.758	0.759	0.053	0.590	0.470
RSF	0.730	0.833	0.041	0.500	0.380

Table 5.8 shows that the naive Cox PH and forward selection model gave the same results with relatively high performance measure outputs under C-index and R^2 measures. For iAUC, the naive model and RSF gave similar metrics. Continuing with the naive model defeats the purpose of the study of comparing model selection techniques. Therefore, the RSF was the model selection technique that performed comparatively well under time to constriction.

Significant interactions between covariates in the model selection analysis would be worth incorporating in future research. This might help assess whether this improves the performance measures of the model selection techniques. In this dissertation, interaction was not incorporated because in clinical trial studies it is considered when a statistician has more information from a clinician.

Estimation results of models selected through the iAUC metric

Table 2 in Appendix B illustrates forward selection and random forest model estimation results for time to death and time to constriction final models selected through the iAUC comparative metric, respectively. The time to composite event selected the model given by the forward selection technique as comparatively better, which can be extracted from Table 5.2. The random plots for the hazard ratios of the selected models are shown in Figures 5.9 to 5.11 below.

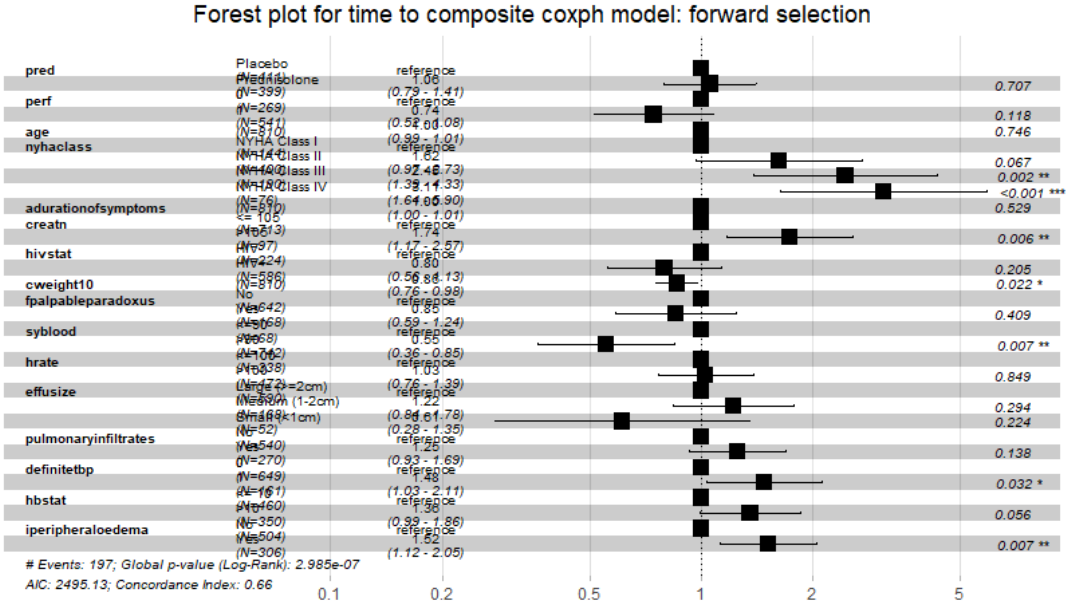


Figure 5.9: Forest plot for forward selection of time to composite event hazard ratios

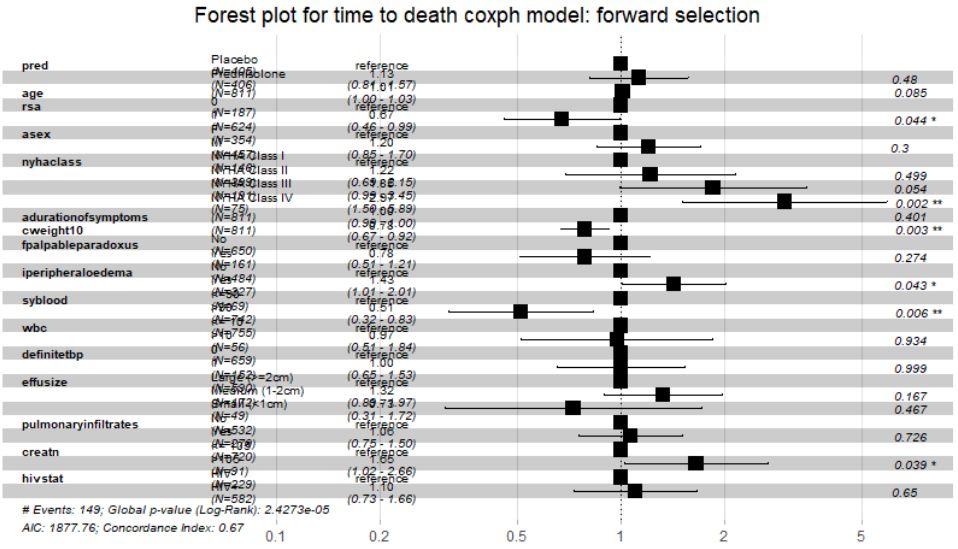


Figure 5.10: Forest plot for forward selection of time to death event hazard ratios

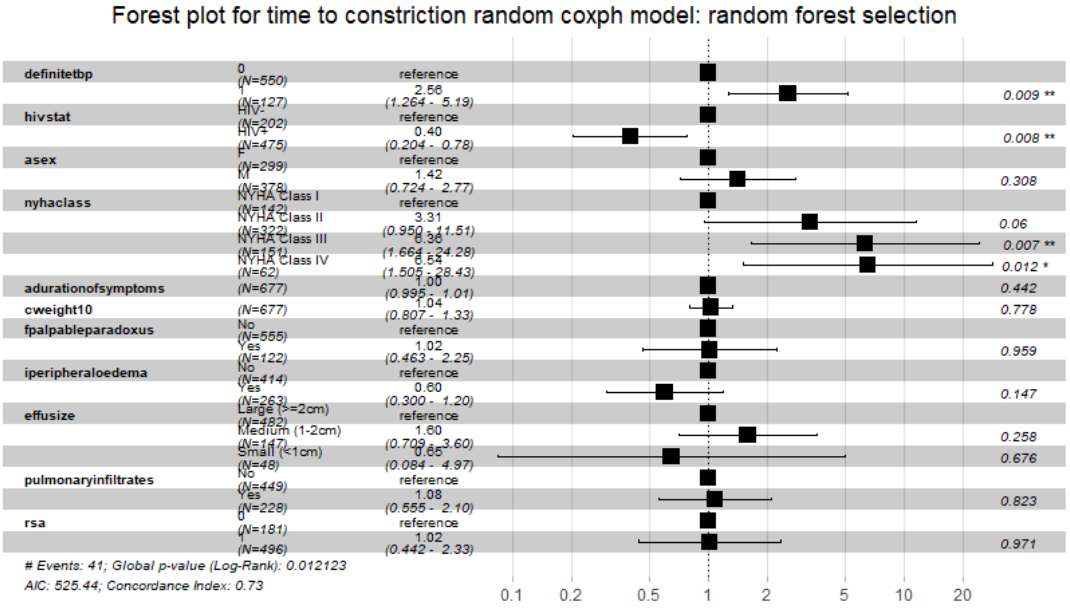


Figure 5.11: Forest plot for random forest selection of time to constriction event hazard ratios

5.1.4 Proportionality assumption test

A key assumption in a Cox PH model is that of the proportionality of the covariates. To test this assumption, the covariates included in the model are interacted with time with the expectation that the hazard ratios will be constant over time. Therefore, assuming that the baseline hazard is the same for all study participants.

To validate the proportionality assumption for the final model of the time to composite event, the Schoenfeld residuals were used. A small p-value would indicate a violation of the proportionality assumption. The p-values in Table 5.9 are not significant, except for systolic blood pressure, at the 95% confidence level. This indicates that the assumption of proportionality is appropriate for the other covariates. Overall, the assumption holds for the model with a significant global p-value of 0.081. The Schoenfeld residuals plots for each individual covariate are given in Appendix B.

Table 5.9: Proportional hazards test for the time to composite event forward selection model

	Chi-square	**Degrees of Freedom**	**p-value**
Prednisolone	2.92	1	0.088
Pericardiocentesis at randomisation	1.88	1	0.170
Age (in years)	1.03	1	0.310
NYHA Class at study entry	3.49	1	0.323
Duration of symptoms	0.833	1	0.361
Creatanine	2.53	1	0.112
HIV status	0.171	1	0.679
Weight	2.17	1	0.141
Palpableparadoxus	0.943	1	0.331
Systolic blood pressure	4.34	1	0.037
Heart rate	0.315	1	0.575
Effusion size	2.47	1	0.291
Chest Xray pulmonary infiltrate	0.000014	1	0.997
Definite TB pericarditis status	1.39	1	0.239
High blood status	3.32	1	0.069
Peripheral oedema	0.00158	1	0.900
GLOBAL	28.1	19	0.081

5.2 A simulation study

A Monte Carlo simulation is performed to provide an objective basis on which to assess the performance of the proposed model selection techniques [49].

Using the *coxed* package in R [29], three data sets with 2000, 1000 and 500 observations were generated for 150 days. The number of covariates for these data

sets were 15, 25 and 50, respectively. The choice of the number of covariates was arbitrary. However, for the 50 covariates, the code took too long to run, therefore a limitation in computational power also played a role. Two censoring levels were considered: 20% and 10%, with the observations designated as right-censored. The simulated data set had no time-varying covariates with a non-parametric hazard function. All simulation calculations were performed using the statistical software *R* on the *RStudio* interface [72].

Table 5.10: Simulation comparison results at 20% censoring level

	Performance Measure)				
Selection Technique	C-index	iAUC	R^2	R^2_{mer}	R^2_{mev}
	Time to failure with 15 covariates (n=2000)				
None	0.582	0.566	0.067	0.067	0.042
Forward selection	0.555	0.522	0.030	0.030	0.018
Backward selection	0.576	0.557	0.056	-	-
Stepwise selection	0.579	0.568	0.063	0.063	0.040
Best subset selection	0.518	0.472	0.010	0.010	0.006
AIC	0.579	0.568	0.063	0.063	0.040
BIC	0.579	0.563	0.062	0.062	0.039
Lasso	0.580	0.569	0.060	0.060	0.037
Ridge	0.577	0.568	0.060	0.060	0.037
RSF	0.579	0.562	0.062	0.062	0.039
	Time to failure with 25 covariates (n=1000)				
None	0.592	0.581	0.067	0.083	0.052
Forward selection	0.584	0.589	0.060	0.074	0.047
Backward selection	0.539	0.475	0.061	-	-
Stepwise selection	0.577	0.608	0.044	0.055	0.034
Best subset selection	0.550	0.549	0.021	0.026	0.016
AIC	0.562	0.533	0.033	0.041	0.026
BIC	0.591	0.583	0.065	0.080	0.050
Lasso	0.547	0.539	0.016	0.020	0.012
Ridge	0.582	0.584	0.057	0.071	0.044
RSF	0.572	0.562	0.045	0.056	0.035
	Time to failure with 50 covariates (n=500)				
None	0.651	0.564	0.210	0.250	0.170
Forward selection	0.619	0.511	0.140	0.170	0.110
Backward selection	0.626	0.560	0.170	-	-
Stepwise selection	0.627	0.556	0.170	0.200	0.130
Best subset selection	0.560	0.425	0.043	0.053	0.033
AIC	0.594	0.639	0.100	0.120	0.079
BIC	0.633	0.548	0.170	0.200	0.140
Lasso	0.510	0.510	0.001	0.001	0.000
Ridge	0.634	0.632	0.170	0.200	0.130
RSF	0.631	0.556	0.150	0.180	0.120

Table 5.10 gives results of the performance measures of the model selection techniques for the simulated data sets at the 20% censoring level. The performance

measures show that the naive Cox PH model without a selection technique gave relatively high metrics for this particular set of data except under iAUC. The iAUC method chose varying model selection techniques for the different simulated data sets.

Using iAUC, the set with 2000 observations selected *Lasso* as the preferred technique with a metric of 0.569, which was closely followed by Ridge, AIC and stepwise techniques at 0.568. The set with 1000 observations selected *stepwise* with a metric of 0.608, whilst for the set with 500 observations, AIC performed relatively better than all selection techniques with a metric of 0.639.

Table 5.11: Simulation comparison results at 10% censoring level

Selection Technique	Performance Measure (n=2000)				
	C-index	iAUC	R^2	R^2_{mer}	R^2_{mev}
	Time to failure with 15 covariates (n=2000)				
None	0.581	0.575	0.062	0.069	0.043
Forward selection	0.553	0.533	0.026	0.029	0.018
Backward selection	0.577	0.567	0.390	-	-
Stepwise selection	0.579	0.577	0.060	0.067	0.042
Best subset selection	0.502	0.477	0.011	0.012	0.007
AIC	0.579	0.578	0.060	0.067	0.042
BIC	0.580	0.576	0.058	0.065	0.040
Lasso	0.569	0.559	0.040	0.044	0.027
Ridge	0.575	0.569	0.053	0.059	0.037
RSF	0.577	0.569	0.053	0.060	0.037
	Time to failure with 25 covariates (n=1000)				
None	0.581	0.582	0.062	0.068	0.042
Forward selection	0.572	0.596	0.054	0.059	0.037
Backward selection	0.568	0.585	0.046	-	-
Stepwise selection	0.568	0.585	0.046	0.050	0.031
Best subset selection	0.534	0.561	0.010	0.011	0.007
AIC	0.559	0.545	0.028	0.030	0.019
BIC	0.578	0.578	0.057	0.062	0.039
Lasso	0.560	0.560	0.036	0.039	0.024
Ridge	0.562	0.558	0.037	0.040	0.025
RSF	0.512	0.582	0.038	0.041	0.026
	Time to failure with 50 covariates (n=500)				
None	0.644	0.580	0.210	0.230	0.150
Forward selection	0.617	0.574	0.160	0.180	0.110
Backward selection	0.620	0.585	0.170	-	-
Stepwise selection	0.621	0.585	0.170	0.180	0.120
Best subset selection	0.577	0.423	0.065	0.072	0.045
AIC	0.574	0.527	0.081	0.089	0.056
BIC	0.630	0.556	0.170	0.190	0.121
Lasso	0.533	0.543	0.018	0.020	0.013
Ridge	0.612	0.625	0.150	0.160	0.100
RSF	0.441	0.584	0.170	0.190	0.120

Results of the performance measures of the model selection techniques for the 10% censoring level scenario are given in Table 5.11. Using iAUC, the set with 2000 observations selected the *stepwise* model selection technique as the preferred technique with a metric of 0.577. The set with 1000 observations suggested that the *forward* model selection technique performed better than the other techniques, with a metric of 0.596, whilst the set with 500 observations, the *stepwise* model selection technique performed relatively better than all selection techniques with a metric of 0.585.

We note that, for each simulated data set under each censoring level, the performance metrics choose different model selection techniques. This suggests that the technique used for selecting a model might differ depending on sample size, number of covariates in the model and/or the censoring level. Hence, to establish a guideline for identifying an "appropriate" model selection technique based on the characteristics of a data set is a call for future research.

Chapter 6

Discussion and conclusion

The model selection techniques used in this dissertation are proposed to find a parsimonious model to simplify the Cox PH model. To test the efficiency of these techniques, a multicentre clinical trial data set and simulation illustration were used. Then, performance methods were employed to find the techniques that efficiently solve the problems of optimisation in a Cox PH regression model with many covariates.

In the original IMPI study, the use of Mycobacterium indicus pranii (Mw) was flagged as redundant. Therefore, the analysis only focused on prednisolone therapy as a treatment covariate. The results in this dissertation show that the treatment covariate of prednisolone has no significant effect on the composite outcome (see forward selection model results in Table 5.2) and the combined outcome of death from all causes (Table 2). However, the use of prednisolone is shown in Table 2 to reduce the incidence of constrictive pericarditis by 62.9%. These results are consistent with the results from the original trial study by Mayosi et al [51].

The real data set and simulations in this dissertation showed results contrary to those of Tibshirani [74] and Fan & Li [26] whose results supported the Lasso as a better selection technique than the stepwise. In the IMPI data set, the Lasso did not perform better than the stepwise selection technique for all three events of interest. Under the composite event, stepwise selection method performed comparatively better on all methods. Lasso only performed better than stepwise selection method under the iAUC performance method for both the death response variable and the constriction response but not under the C-index and R^2_{mev} metrics.

For the simulated data set, Lasso performed better than stepwise selection method under the C-index and iAUC for the data set with 2000 observations. But regarding the other data sets with comparatively fewer observations and more covariates,

Lasso did not perform better than stepwise selection method. The R code did not run for most of the selection techniques when a simulation with a small sample of 50 observations and 50 covariates was done. Another scenario that was deemed computationally infeasible for most selection techniques was when the sample size was set to 50000 with an equally large covariate base of 60.

Under the simulation study presented in this dissertation, we further observed that the Lasso technique was inefficient for the 50 covariates and 500 observations scenario. The technique on this relatively small sample case selected only two covariates, which is unreliable. Peterson [60] suggested that to improve the performance of Lasso, one might have to play around with, particularly the tuning parameter and also check the censoring scheme to match that of Fan [26] and Tibshirani [74].

The level at which censoring occurs might also have an impact on variable selection. For instance, high censoring might be suitable for gene expression profile data. For this kind of data, high censoring has been shown to improve model estimation and variable selection, unlike low censoring. On the contrary, in a heart transplant dataset, censoring levels are expected to be low as the sample size in such a study tends to be low [33].

Another observation was that the AIC and BIC techniques were computationally more challenging than the other techniques. The AIC only performed well under iAUC for the 500 small sample simulation data set. Additionally, R^2 measures performed poorly on all selection techniques with all the metrics being close to 0 rather than the desirable 1, suggesting that model accuracy was compromised. Even though the results are reported for these metrics, we decided not to use them when deciding on the technique for selecting a final model.

RSF performed consistently well alongside the naive and forward selection techniques under the IMPI death response variable but not under iAUC. However, being an area of active research means there is a need to assess the impact of noisy covariates with varying sample sizes [79]. Therefore, the best technique for model selection should be chosen for each specific data set, as there is yet to be agreement as to which one performs reasonably well under all conditions [4]. However, since the iAUC consistently gave different results for each selection technique, unlike the other methods, it was decided upon as the metric for model performance.

There is a scope for future research for handling missing data in Cox modelling in R, though attempts are made to minimise the presence of missing data in clinical trials, but hard to control in observational studies. However, missing data can still occur in clinical trials for some covariates due to a variety of reasons, including the patients not being able to make evaluation appointments [14]. For the IMPI clinical trial data set, this was observed in the analysis stage where the R code gave errors for all

selection techniques except RSF due to the presence of missing data in the covariates, not the outcome. Covariates like weight, systolic blood pressure, heart rate, high blood status, creatanine, effusion size, NYHA class and pulmonary infusion had some missing values across all three IMPI trial data set events of interest.

In conclusion, it is vital to know the objective of a study. Knowing the objective will assist the researcher identify the appropriate model selection technique which will give a well-performing model. This places importance on careful planning at the study design stage. For instance, if the objective is inference, consistency of the model is of paramount importance, and penalised techniques have typically been shown to be uncertain in high-dimensional small sample studies. But if the goal is prediction, then consistency is not much of a concern. Similarly, AIC is known to perform better when sample size is larger than number of covariates [19].

Additionally, for a potential future research direction, one may consider examining multicollinearity among model covariates before the selection process as a check to avoid having less precise results. Further, one might also want to research on techniques that deal with data sets that have a large set of covariates than the sample size like the (I)SIS of Fan et al [25]. In this dissertation, competing risks between death, constriction and cardiac tamponade were not taken into account for the IMPI trial data set. One may consider incorporating time-dependent covariates like CD4 counts for HIV positive patients and using techniques like the aLASSO using CAD [80] as an extension, as this dissertation only focused on fixed covariates.

Bibliography

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 267–281. Springer.

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

[3] Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in medicine*, 33(6):1057–1069.

[4] Bagherzadeh-Khiabani, F., Ramezankhani, A., Azizi, F., Hadaegh, F., Steyerberg, E. W., and Khalili, D. (2016). A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. *Journal of clinical epidemiology*, 71:76–85.

[5] Bělašková, S., Fišerová, E., and Krupičková, S. (2013). Study of bootstrap estimates in cox regression model with delayed entry. *Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica*, 52(2):21–30.

[6] Bou-Hamad, I., Larocque, D., Ben-Ameur, H., et al. (2011). A review of survival trees. *Statistics surveys*, 5:44–71.

[7] Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

[8] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[9] Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, 66(3):429–436.

[10] Chen, X. and Ishwaran, H. (2013). Pathway hunting by random survival forests. *Bioinformatics*, 29(1):99–105.

[11] Choi, I., Wells, B. J., Yu, C., and Kattan, M. W. (2011). An empirical approach to model selection through validation for censored survival data. *Journal of biomedical informatics*, 44(4):595–606.

[12] Clark, T. G., Bradburn, M. J., Love, S. B., and Altman, D. G. (2003). Survival analysis part i: basic concepts and first analyses. *British journal of cancer*, 89(2):232–238.

[13] Company, I.-H. S. C. (2017). *My.stepwise: Stepwise Variable Selection Procedures for Regression Analysis*. R package version 0.1.0.

[14] Council, N. R. et al. (2010). The prevention and treatment of missing data in clinical trials.

[15] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

[16] Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

[17] David, C. (2003). Modelling survival data in medical research. *Chapman Hall/CRC Texts in Statistical Science*.

[18] Dietrich, S., Floegel, A., Troll, M., Kühn, T., Rathmann, W., Peters, A., Sookthai, D., Von Bergen, M., Kaaks, R., Adamski, J., et al. (2016). Random survival forest in practice: a method for modelling complex metabolomics data in time to event analysis. *International journal of epidemiology*, 45(5):1406–1420.

[19] Ding, J., Tarokh, V., and Yang, Y. (2018). Model selection techniques: An overview. *IEEE Signal Processing Magazine*, 35(6):16–34.

[20] Efron, B. (1977). The efficiency of cox's likelihood function for censored data. *Journal of the American statistical Association*, 72(359):557–565.

[21] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

[22] Ekman, A. (2017). Variable selection for the cox proportional hazards model: A simulation study comparing the stepwise, lasso and bootstrap approach.

[23] Emmert-Streib, F. and Dehmer, M. (2019). Introduction to survival analysis in practice. *Machine Learning and Knowledge Extraction*, 1(3):1013–1038.

[Englebert et al.] Englebert, C., Quinn, T., and Bichindaritz, I. Feature selection for survival analysis in bioinformatics.

[25] Fan, J., Feng, Y., Wu, Y., et al. (2010). High-dimensional variable selection for cox's proportional hazards model. In *Borrowing strength: Theory powering applications–a Festschrift for Lawrence D. Brown*, pages 70–86. Institute of Mathematical Statistics.

[26] Fan, J. and Li, R. (2002). Variable selection for cox's proportional hazards model and frailty model. *Annals of Statistics*, pages 74–99.

[27] Gong, G. (1982). Some ideas on using the bootstrap in assessing model variability. In *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface*, pages 169–173. Springer New York.

[28] Halabi, S., Dutta, S., Wu, Y., and Liu, A. (2020). Score and deviance residuals based on the full likelihood approach in survival analysis. *Pharmaceutical statistics*, 19(6):940–954.

[29] Harden, J. J. and Kropko, J. (2019). Simulating duration data for the cox model. *Political Science Research and Methods*, 7(4):921–928.

[30] Hazra, A. (2017). Using the confidence interval confidently. *Journal of thoracic disease*, 9(10):4125.

[31] Hiller, L., Marshall, A., and Dunn, J. (2015). Assessing violations of the proportional hazards assumption in cox regression: does the chosen method matter? *Trials*, 16(S2):P134.

[32] Hosmer Jr, D. W., Lemeshow, S., and May, S. (2011). *Applied survival analysis: regression modeling of time-to-event data*, volume 618. John Wiley & Sons.

[33] Hu, S. (2007). *New Methods for Variable Selection with Applications to Survival Analysis and Statistical Redundancy Analysis Using Gene Expression Data*. PhD thesis, Case Western Reserve University.

[34] Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307.

[35] Intrator, O. and Kooperberg, C. (1995). Trees and splines in survival analysis. *Statistical methods in medical research*, 4(3):237–261.

[36] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S., et al. (2008). Random survival forests. *The annals of applied statistics*, 2(3):841–860.

[37] Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J., and Lauer, M. S. (2010). High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, 105(489):205–217.

[38] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

[39] Jin, Z., Lin, D., and Ying, Z. (2006). On least-squares regression with censored data. *Biometrika*, 93(1):147–161.

[40] Kadane, J. B. and Lazar, N. A. (2004). Methods and criteria for model selection. *Journal of the American statistical Association*, 99(465):279–290.

[41] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

[42] Kent, J. T. and O'QUIGLEY, J. (1988). Measures of dependence for censored survival data. *Biometrika*, 75(3):525–534.

[43] Koul, H., Susarla, V., Van Ryzin, J., et al. (1981). Regression analysis with randomly right-censored data. *Annals of statistics*, 9(6):1276–1288.

[44] Lai, Y., Hayashida, M., and Akutsu, T. (2013). Survival analysis by penalized regression and matrix factorization. *The Scientific World Journal*, 2013.

[45] Lang, M., Kotthaus, H., Marwedel, P., Weihs, C., Rahnenführer, J., and Bischl, B. (2015). Automatic model selection for high-dimensional survival analysis. *Journal of Statistical Computation and Simulation*, 85(1):62–76.

[46] Leung, K.-M., Elashoff, R. M., and Afifi, A. A. (1997). Censoring issues in survival analysis. *Annual review of public health*, 18(1):83–104.

[47] Liang, H. and Zou, G. (2008). Improved aic selection strategy for survival analysis. *Computational statistics & data analysis*, 52(5):2538–2548.

[48] Liu, X., Peng, Y., Tu, D., and Liang, H. (2012). Variable selection in semiparametric cure models based on penalized likelihood, with application to breast cancer clinical trials. *Statistics in medicine*, 31(24):2882–2891.

[49] Loughin, T. M. (1995). A residual bootstrap for regression parameters in proportional hazards models. *Journal of Statistical Computation and Simulation*, 52(4):367–384.

[50] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.*, 50:163–170.

[51] Mayosi, B. M., Ntsekhe, M., Bosch, J., Pandie, S., Jung, H., Gumedze, F., Pogue, J., Thabane, L., Smieja, M., Francis, V., et al. (2014). Prednisolone and mycobacterium indicus pranii in tuberculous pericarditis. *New England Journal of Medicine*, 371(12):1121–1130.

[52] Mbogning, C. and Broët, P. (2016). Bagging survival tree procedure for variable selection and prediction in the presence of nonsusceptible patients. *BMC bioinformatics*, 17(1):1–21.

[53] Miller, R. and Halpern, J. (1982). Regression with censored data. *Biometrika*, 69(3):521–531.

[54] Miller, R. G. (1976). Least squares regression with censored data. *Biometrika*, 63(3):449–464.

[55] Moore, D. F. (2016). *Applied survival analysis using R*. Springer.

[56] Nagelkerke, N. J. et al. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692.

[57] Newton, M. A., Quintana, F. A., Den Boon, J. A., Sengupta, S., Ahlquist, P., et al. (2007). Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *The Annals of Applied Statistics*, 1(1):85–106.

[58] O'Quigley, J., Xu, R., and Stare, J. (2005). Explained randomness in proportional hazards models. *Statistics in medicine*, 24(3):479–489.

[59] Pang, H., Datta, D., and Zhao, H. (2010). Pathway analysis using random forests with bivariate node-split for survival outcomes. *Bioinformatics*, 26(2):250–258.

[60] Peterson, S. and Sehlstedt, K. (2018). Variable selection techniques for the cox proportional hazards model: A comparative study.

[61] Raab, G. M., Day, S., and Sales, J. (2000). How to select covariates to include in the analysis of a clinical trial. *Controlled clinical trials*, 21(4):330–342.

[62] Rahman, M. S., Ambler, G., Choodari-Oskooei, B., and Omar, R. Z. (2017). Review and evaluation of performance measures for survival prediction models in external validation settings. *BMC medical research methodology*, 17(1):60.

[63] Ranganathan, S., Nakai, K., and Schonbach, C. (2018). *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*. Elsevier.

[64] Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C., Nussenbaum, B., and Wang, E. W. (2010). A practical guide to understanding kaplan-meier curves. *Otolaryngology—Head and Neck Surgery*, 143(3):331–336.

[65] Royston, P. (2006). Explained variation for survival models. *The Stata Journal*, 6(1):83–96.

[66] Sandri, M. and Zuccolotto, P. (2006). Variable selection using random forests. In *Data analysis, classification and the forward search*, pages 263–270. Springer.

[67] Sarkar, K., Chowdhury, R., and Dasgupta, A. (2017). Analysis of survival data: Challenges and algorithm-based model selection. *Journal of Clinical and Diagnostic Research: JCDR*, 11(6):LC14.

[68] Sauerbrei, W., Royston, P., Schumacher, M., Austin, P. C., and Tu, J. V. (2005). Austin, pc, and tu, jv (2004)," bootstrap methods for developing predictive models,"" the american statistician," 58, 131-137: Comment by sauerbrei, royston, and schumacher and reply. *The American Statistician*, 59(1):116–118.

[69] Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

[70] Stoica, P. and Selen, Y. (2004). Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, 21(4):36–47.

[71] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

[72] Team, R. et al. (2015). Rstudio: integrated development for r. *RStudio, Inc., Boston, MA URL http://www. rstudio. com*, 42:14.

[73] Therneau, T. M. (2020). A package for survival analysis in r. R package version 3.2-7.

[74] Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395.

[75] Tibshirani, R. J. and Efron, B. (1993). An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57:1–436.

[76] Valery, P. (1970). *The Collected Works of Paul Valéry: Analects; Translated by Stuart Gilbert*. Routledge & Kegan Paul.

[77] Volinsky, C. T. and Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics*, 56(1):256–262.

[78] Von Neumann, J. (1947). The mathematician. *1947*, pages 180–196.

[79] Wang, H. and Li, G. (2017). A selective review on random survival forests for high dimensional data. *Quantitative bio-science*, 36(2):85.

[80] Wang, Y., Hong, C., Palmer, N., Di, Q., Schwartz, J., Kohane, I., and Cai, T. (2021). A fast divide-and-conquer sparse cox regression. *Biostatistics*, 22(2):381–401.

[81] Xue, Y. and Schifano, E. D. (2017). Diagnostics for the cox model. *Communications for statistical Applications and Methods*, 24(6):583–604.

[82] Zhang, Y. and Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, 187(1):95–112.

[83] Zhao, X., Su, J., and Wu, X. (2014). Variable selection for cox's proportional hazards regression model based on lasso-cda. *Advances in Electrical and Electronics Engineering*, 92:199.

[84] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Appendices

Appendix A

R packages and functions used for model selection and performance techniques

Method	**Package::function**
Stepwise selection	MASS::stepAIC
Forward selection	stats::step
Backward selection	pec::selectCox
Best subset selection	BeSS:bess
AIC	bootStepAIC::boot.stepAIC
BIC	base::ifelse
Lasso	glmnet::glmnet
Ridge	glmnet::glmnet
RSF	ranger::ranger
C-index	survival::survConcordance
iAUC	survAUC::AUC.uno
R^2	survMisc::rsq
Random plots	survminer::ggforest
Proportionality test	survival::cox.zph

Covariates with missing values

Covariate	**Missing values (% of 1400)**
art-ever	1200 (86%)
aicd4count	695 (50%)
cd4cat50	624 (45%)
cd4cat	624 (45%)
perithick	error

Appendix B

IMPI data set Kaplan-Meier survival curve estimations

Figure 2 shows KM curves for the three events of interest plotted on the same graph. The composite endpoint (black curve) has, as expected, the lowest survival curve of the three endpoints as it includes observing either one of the endpoints. On the other hand, the survival curve for time to constriction (red curve) has the relatively high survival curve.

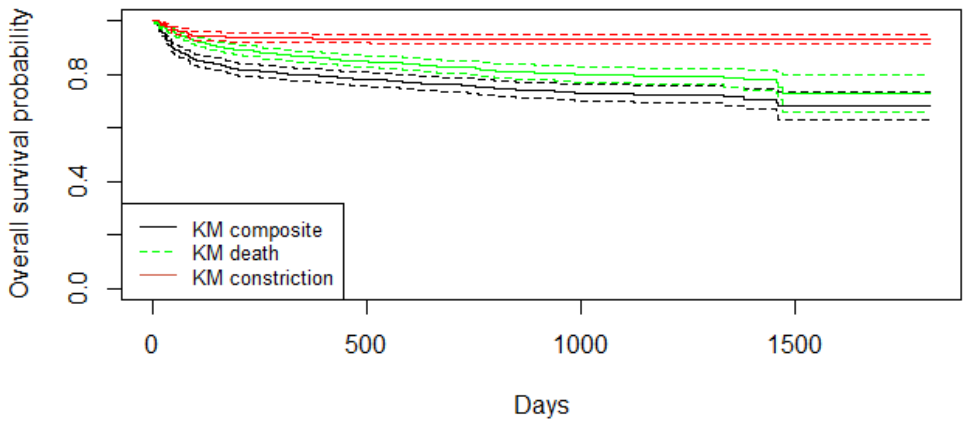


Figure 1: Comparing KM curves for time to composite (black), time to death (green), time to constriction (red).

IMPI data set AUC curves for composite endpoint

Figure 2 shows AUC curves for the model selection techniques on time to composite plotted on the same graph. Note that the AUC for the naive (black) and RSF

(coral3) models are not as constant as the other curves. The AUC curve for stepwise (green) and AIC (orange) are similar as seen by the curves.

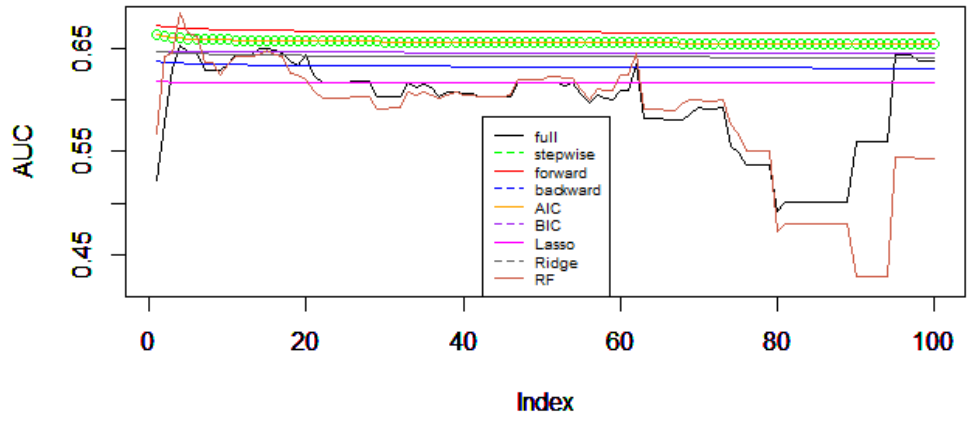


Figure 2: Comparing AUC curves for full model (black), stepwise (green), forward selection (red) , backward elimination (blue), AIC (orange), BIC (purple), Lasso (magenta), Ridge (grey) and RF (coral3)

Final IMPI data set models

Final models based on iAUC for each of the events of interest under the IMPI clinical trial data set.

Table 1: Forward Cox PH model selection technique for composite endpoint

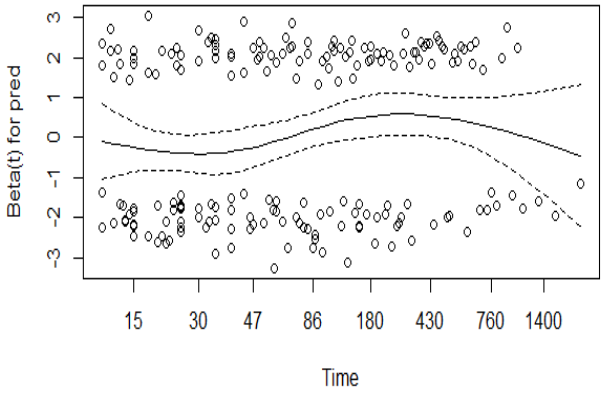
Variable	**HR (95% CI)**	**p-value**
Pericardiocentesis at randomisation (Yes)	0.756 (0.549 - 1.040)	0.086
NYHA Class at study entry		
II	1.429 (0.919 - 2.221)	0.113
III	2.416 (1.504 - 3.881)	0.000
IV	2.829 (1.651 - 4.849)	0.000
Creatanine ($\geq$ 105)	1.622 (1.149 - 2.289)	0.006
Weight (kg)	0.879 (0.789 - 0.981)	0.021
Systolic blood pressure ($\geq$ 90)	0.681 (0.459 - 1.009)	0.056
Definite TB pericarditis status (Yes)	1.523 (1.109 - 2.092)	0.009
Haemoglobin status ($\geq$ 10)	1.299 (0.988 - 1.708)	0.061
Peripheral oedema (Yes)	1.417 (1.091 - 1.840)	0.009

Table 2: Final Cox PH models selected for time to death and time to constriction events

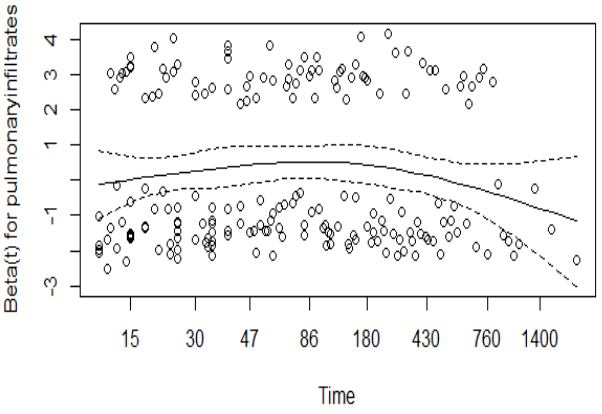
	Time to death forward Cox PH model		**Time to constriction random forest Cox PH model**	
Variable	**HR (95% CI)**	**p-value**	**HR (95% CI)**	**p-value**
Prednisolone	1.127 (0.810 - 1.568)	0.480	-	-
Age (in years)	1.012 (0.998 - 1.026)	0.085	-	-
RSA (Yes)	0.673 (0.459 - 0.989)	0.044	1.015 (0.442 - 2.333)	0.972
Sex (Male)	1.201 (0.850 - 1.696)	0.300	1.417 (0.724 - 2.774)	0.308
NYHA Class at study entry				
II	1.216 (0.689 - 2.145)	0.500	3.306 (0.950 - 11.511)	0.060
III	1.849 (0.989 - 3.454)	0.054	6.356 (1.664 - 24.275)	0.007
IV	2.974 (1.503 - 5.887)	0.002	6.542 (1.505 - 28.428)	0.012
Duration of symptoms	0.998 (1.002 - 1.003)	0.401	1.003 (0.995 - 1.012)	0.442
Weight (kg)	0.785 (0.6704 - 0.919)	0.003	1.037 (0.807 - 1.331)	0.778
Creatanine (≥ 105)	1.650 (0.606 - 2.655)	0.039	-	-
Palpable paradoxus (Yes)	0.785 (0.508 - 1.212)	0.274	1.021 (0.463 - 2.250)	0.959
Peripheral oedema (Yes)	1.426 (0.702 - 2.011)	0.043	0.599 (0.300 - 1.198)	0.147
Systolic blood pressure (≥ 90)	0.513 (0.317 - 0.830)	0.006	-	-
White blood count (≥ 10)	0.974 (0.515 - 1.842)	0.934	-	-
Definite TB pericarditis status (Yes)	0.999 (0.653 - 1.530)	0.999	2.560 (1.264 - 5.188)	0.009
Effusion size				
Medium (1-2cm)	1.323 (0.890 - 1.966)	0.167	1.598 (0.709 - 3.599)	0.258
Small ($< 1cm$)	0.728 (0.308 - 1.716)	0.467	0.648 (0.084 - 4.971)	0.676
Pulmonary infiltrates (Yes)	1.064 (0.754 - 1.501)	0.726	1.079 (0.555 - 2.098)	0.823
HIV status (positive)	1.100 (0.9090 - 1.661)	0.650	0.400 (0.204 - 0.783)	0.008

Time to composite forward selection model - Schoenfeld residual plots for proportionality assumption test

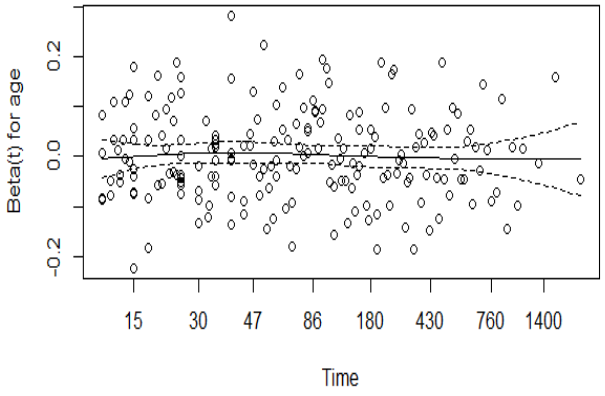
The slope in each plot should be around zero if the proportional hazards assumption holds. A slope that deviates from zero, to a large extent, suggests a violation of the proportional hazards assumption. Most of the plots suggest only a slight rise in the plotted values over time, suggesting no major problem with the proportional hazards assumption.



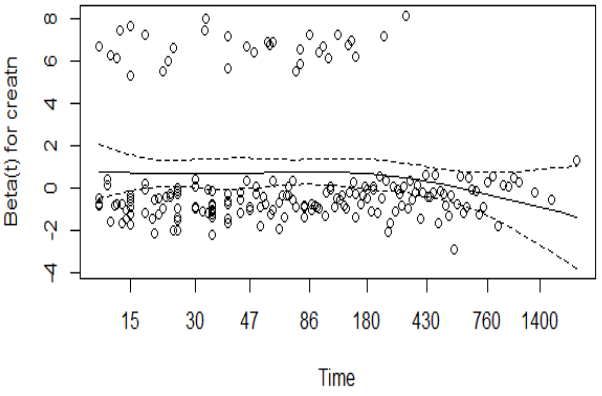
Prednisolone



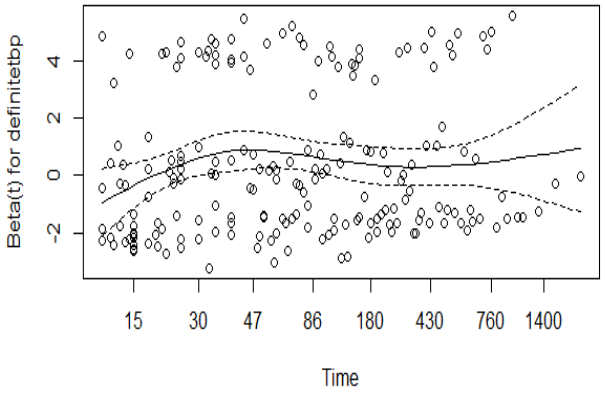
Pulmonary infiltrates



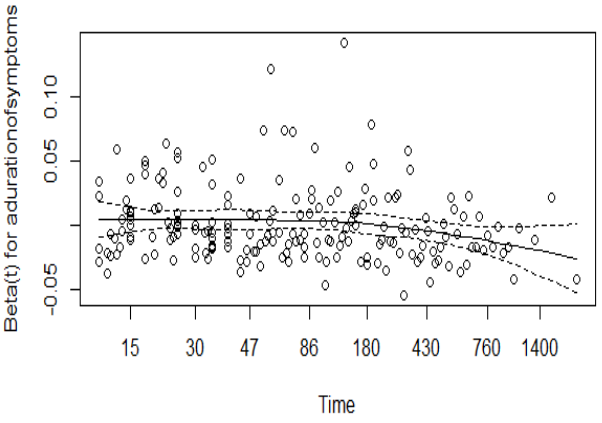
Age in years



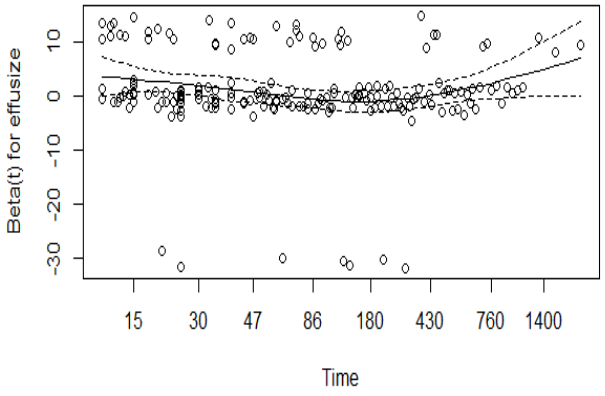
(reatanine



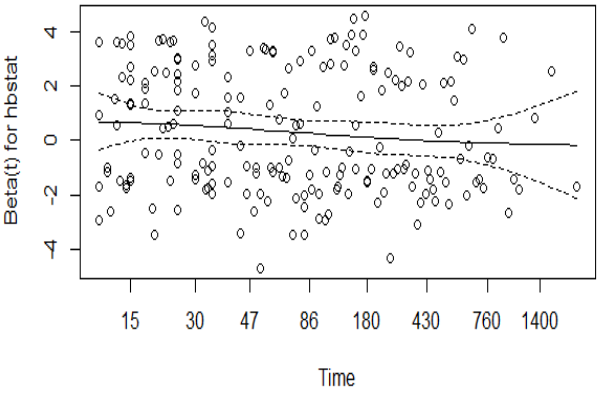
Definite TB pericarditis status



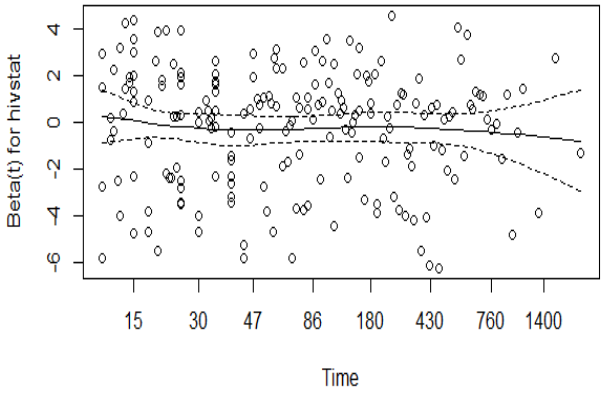
Duration of symptoms



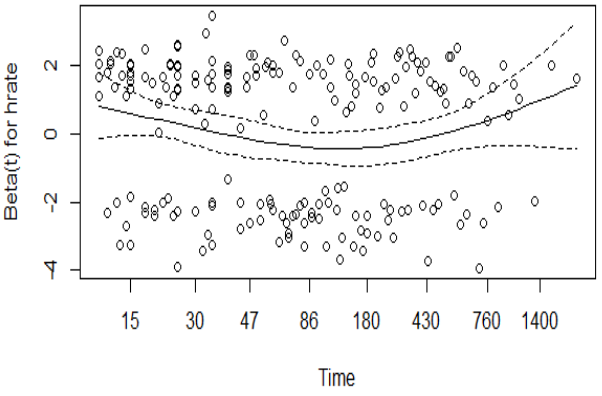
Effusion size



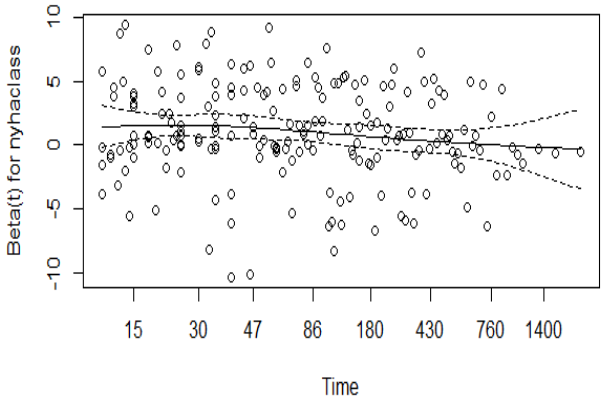
HB status



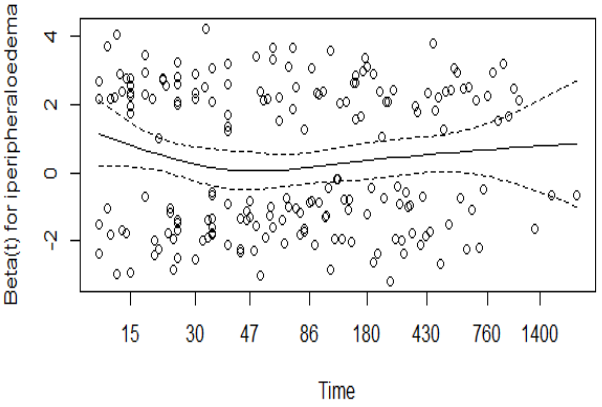
HIV status



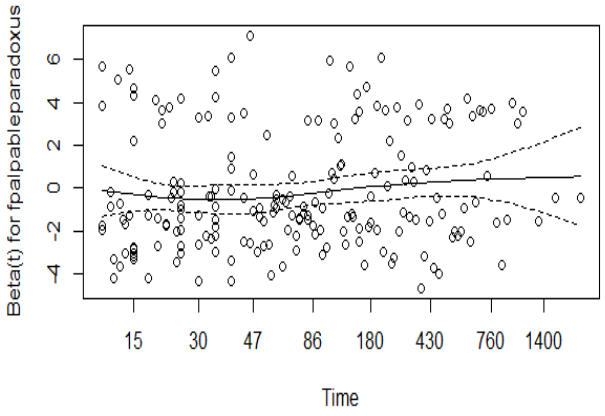
Heart rate



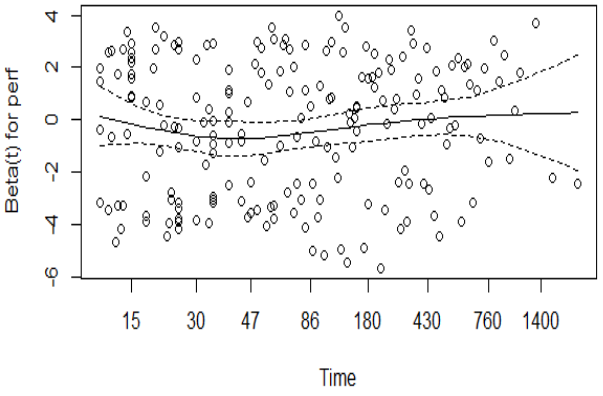
NYHA Class at study entry



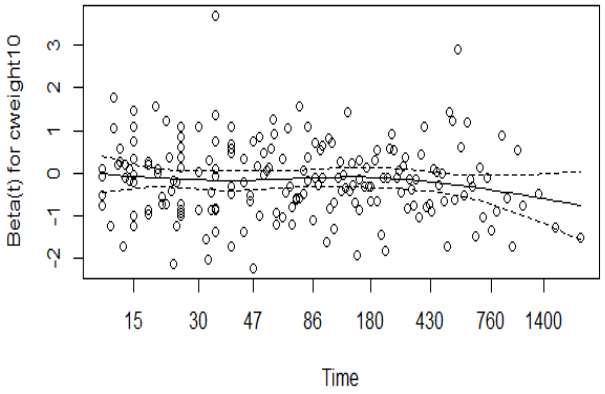
Peripheral oedema



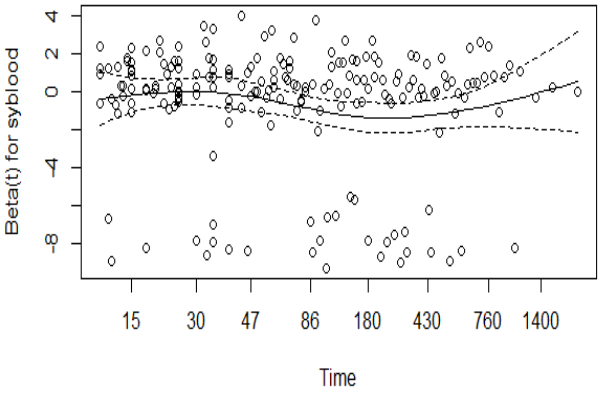
Palpable paradoxus



Pericardiocentesis at randomisation



Weight



Systolic blood pressure

Figure 10: Proportional hazard assumption test plots for the forward selection model for the time to composite event. The Null hypothesis of the test is that the residuals are a random-walk in time around a zero with no pattern.

Appendix C

RSF survival curves

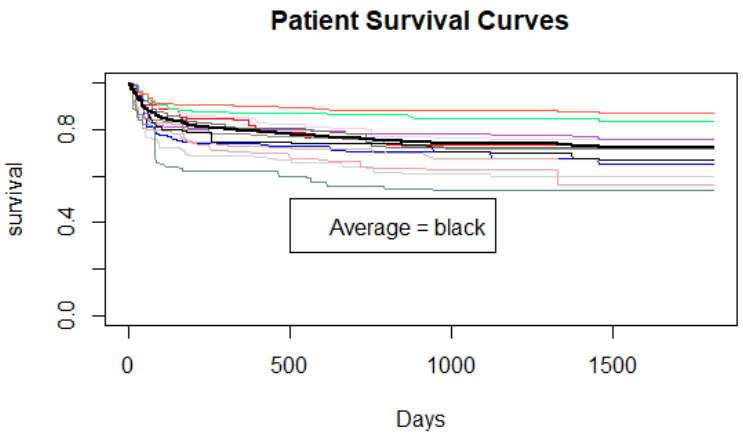


Figure 11: Patient RSF curves with global average curve for the composite event

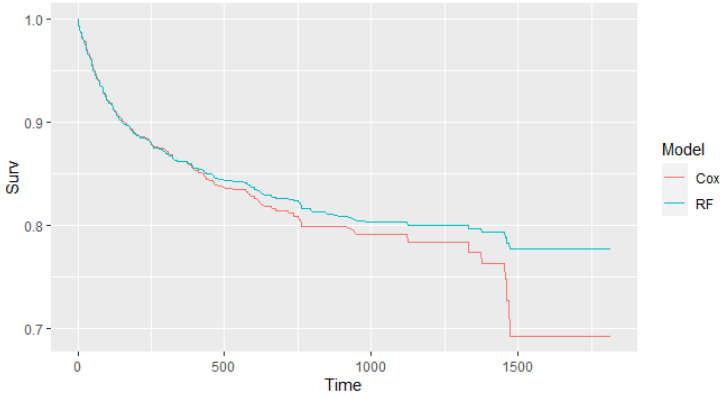


Figure 12: Survival curve of full Cox PH model vs RSF for the death event

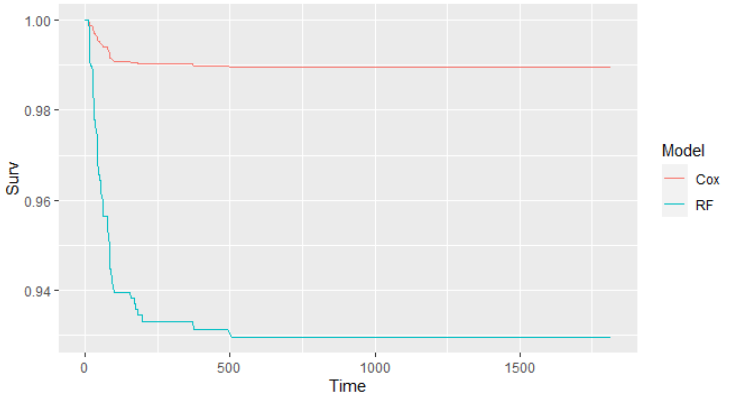


Figure 13: Survival curve of full Cox PH model vs RSF for the constriction event

RSF Variable Importance plots

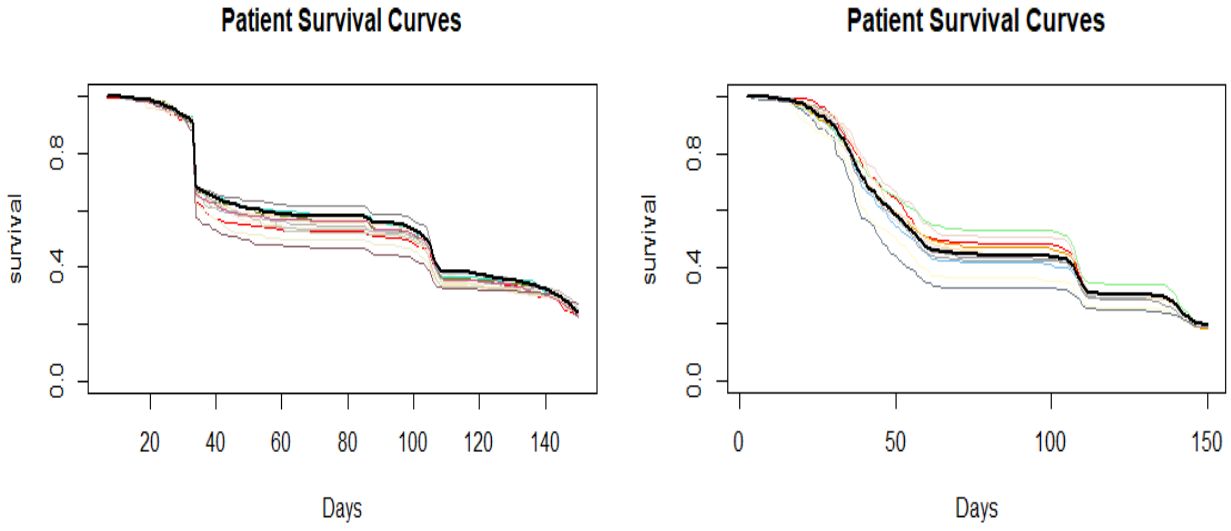
The ranking of the important variables under RSF for all events of interest is given in Table 3.

Table 3: Random survival forest variable importance ranking

Composite		Death		Constriction	
peripheral oedema	0.0075	peripheral oedema	0.0064	definite tb	0.0208
nyhaclass	0.0061	nyhaclass	0.0051	hiv status	0.0164
definite tb	0.0021	palpable paradoxus	0.0049	sex	0.0144
age	0.0018	weight	0.0047	palpable paradoxus	0.0115
duration of symptoms	0.0017	rsa	0.0045	hb status	0.0083
pred	0.0011	sex	0.0042	pred	0.0061
heart rate	0.0006	hiv status	0.0041	pulmonary infiltrates	0.0053
systolic blood	0.0005	pred	0.0038	duration of symptoms	0.0041
weight	0.0003	creatanine	0.0037	heart rate	0.0035
palpable paradoxus	0.0001	age	0.0033	country	0.0032
creatanine	-0.0003	systolic blood pressure	0.0031	age	0.0017
hb status	-0.0006	pulmonary infiltrates	0.0019	nyha class	0.0016
effusion size	-0.0009	duration of symptoms	0.0009	effusion size	0.0015
pulmonary infiltrates	-0.0012	effusion size	0.0001	rsa	0.0013
hiv status	-0.0040	wbc	0.0000	peripheral oedema	0.0007
perf	-0.0051	definite tb	-0.0002	weight	0.0006
				ECG	0.0000
				systolic blood pressure	-0.0001

Appendix D

Simulation plots



(a) (b)

Figure 14: Patient survival curves at (a) the 20% censoring level and (b) the 10% censoring level.

Appendix E

This section details the code used in R for installing all the packages that were used for processing and analysing the IMPI data set and simulations in this dissertation.

R Code

```
# Clear the environment
rm(list=ls())

pack.load<-c("BeSS",  #for best selection and simulating data
             "boot", #to bootstrap on other selection objects
             "bootStepAIC", #bootstrap on AIC
             "caret", #for easy Cross-Validation
             "caTools", #split data into train set and test set
             "coxed",#for fitting Cox PH model and simulating data
             "GGally", #extension of ggplot2
             "glmnet", # for penalised regression
             "Hmisc",
             "lattice",#plotting xyplot
             "leaps",
             "MASS", #for stepAIC function
             "My.stepwise",
             "openxlsx",#exporting/importing Excel files
             "pec", # for calibration slope
             "ranger", # for survival random forest
             "readxl",
             "rms", #for validate() used for bootstrap
             "skimr",#for summary table showin missing values
             "survAUC", #AUC
             "survival", #for fitting Cox PH model
             "survminer",
             "survMisc", #for r-sqrd measures
             "tidyverse",#data wrangling functions
             "todor", #manage R comments
             "VIM", #helps with visualising missing data
             "xtable"#write tables in tex
)

if (!require("rmsfuns")) install.packages("rmsfuns")
```

```
library(rmsfuns)
load_pkg(pack.load)
```

Descriptive IMPI data analysis

```
## Reading in the data
#_____________________________________________________________
impi.all <- read_excel("impi clean.xlsx") %>%
  as_tibble()


## Subsetting columns
#_____________________________________________________________
categ <-
  impi.all %>%
  dplyr::select(pred, perf, asex, rsa,
                country, nyhaclass,
                bonarvs, fpalpableparadoxus,
                iperipheraloedema, syblood, hrate, hbstat,
                wbc, creatn, lvef, effusize,
                tampon, constriction, pulmonaryinfiltrates,
                ecgrthm4, definitetbp)
categ <- as.data.frame(categ)


# Chi-sqr independence tests
#-------------------------------------------------------------
addmargins(table(categ$perf, categ$pred))
round(addmargins(prop.table(table(categ$perf,
                                  categ$pred))*100), 2)
chisq.test(categ$perf, categ$pred)

addmargins(table(categ$asex, categ$pred))
round(addmargins(prop.table(table(categ$asex,
                                  categ$pred))*100),2)
chisq.test(categ$asex, categ$pred)

addmargins(table(categ$nyhaclass, categ$pred))
round(addmargins(prop.table(table(categ$nyhaclass,
                                  categ$pred))*100),2)
chisq.test(categ$nyhaclass, categ$pred)

addmargins(table(categ$rsa, categ$pred))
round(addmargins(prop.table(table(categ$rsa,
                                  categ$pred))*100),2)
chisq.test(categ$rsa, categ$pred)

##
addmargins(table(categ$country, categ$pred))
round(addmargins(prop.table(table(categ$country,
                                  categ$pred))*100),2)
chisq.test(categ$country, categ$pred)
```

```
addmargins(table(categ$fpalpableparadoxus, categ$pred))
round(addmargins(prop.table(table(categ$fpalpableparadoxus,
                                  categ$pred))*100),2)
chisq.test(categ$fpalpableparadoxus, categ$pred)

addmargins(table(categ$iperipheraloedema, categ$pred))
round(addmargins(prop.table(table(categ$iperipheraloedema,
                                  categ$pred))*100),2)
chisq.test(categ$iperipheraloedema, categ$pred)

addmargins(table(categ$syblood, categ$pred))
round(addmargins(prop.table(table(categ$syblood,
                                     categ$pred))*100),2)
chisq.test(categ$syblood, categ$pred)

addmargins(table(categ$hrate, categ$pred))
round(addmargins(prop.table(table(categ$hrate,
                                     categ$pred))*100),2)
chisq.test(categ$hrate, categ$pred)

addmargins(table(categ$hbstat, categ$pred))
round(addmargins(prop.table(table(categ$hbstat,
                                     categ$pred))*100),2)
chisq.test(categ$hbstat, categ$pred)

addmargins(table(categ$wbc, categ$pred))
round(addmargins(prop.table(table(categ$wbc,
                                     categ$pred))*100),2)
chisq.test(categ$wbc, categ$pred)

addmargins(table(categ$creatn, categ$pred))
round(addmargins(prop.table(table(categ$creatn,
                                     categ$pred))*100),2)
chisq.test(categ$creatn, categ$pred)

#  Chi-squared approximation
addmargins(table(categ$lvef, categ$pred))
round(addmargins(prop.table(table(categ$lvef,
                                     categ$pred))*100),2)
chisq.test(categ$lvef, categ$pred)
fisher.test(categ$lvef, categ$pred)

addmargins(table(categ$effusize, categ$pred))
round(addmargins(prop.table(table(categ$effusize,
                                     categ$pred))*100),2)
chisq.test(categ$effusize, categ$pred)

addmargins(table(categ$tampon, categ$pred))
round(addmargins(prop.table(table(categ$tampon,
                                     categ$pred))*100),2)
chisq.test(categ$tampon, categ$pred)
```

```
addmargins(table(categ$constriction, categ$pred))
round(addmargins(prop.table(table(categ$constriction,
                                  categ$pred))*100),2)
chisq.test(categ$constriction, categ$pred)

addmargins(table(categ$pulmonaryinfiltrates, categ$pred))
round(addmargins(prop.table(table(categ$pulmonaryinfiltrates,
                                  categ$pred))*100),2)
chisq.test(categ$pulmonaryinfiltrates, categ$pred)

addmargins(table(categ$ecgrthm4, categ$pred))
round(addmargins(prop.table(table(categ$ecgrthm4,
                                     categ$pred))*100),2)
chisq.test(categ$ecgrthm4, categ$pred)

addmargins(table(categ$definitetbp, categ$pred))
round(addmargins(prop.table(table(categ$definitetbp,
                                  categ$pred))*100),2)
chisq.test(categ$definitetbp, categ$pred)

table(categ$pred)

addmargins(table(impi.all$perithick, impi.all$pred))
round(addmargins(prop.table(table(impi.all$perithick,
                                     impi.all$pred))*100),2)
chisq.test(impi.all$perithick, impi.all$pred)


# Mann-Whitney Dependence tests (no difference in treatment)
#-------------------------------------------------------------
# continuous covariates subsetting
cts <-
  impi.all %>%
  dplyr::select(age, cweight10, adurationofsymptoms)
summary(cts)

# check for normality
hist(cts$age)
hist(cts$cweight10)
hist(cts$adurationofsymptoms)

ggqqplot(cts$age)
ggqqplot(cts$cweight10)
ggqqplot(cts$adurationofsymptoms)


#p-value < 0.05 implying that the distribution
#of the data is significantly different from normal dbn
#thus, we can't assume normality for all 3.
shapiro.test(cts$age)
shapiro.test(cts$cweight10)
shapiro.test(cts$adurationofsymptoms)
```

```
impi.all$pred <- as.factor(impi.all$pred)

#get medians and IQR for covariate by pred
summary(impi.all$age)
round(tapply(impi.all$age,
       impi.all$pred,median, na.rm=T),2)
round(tapply(impi.all$age,
             impi.all$pred,IQR, na.rm=T),2)

summary(impi.all$cweight10)
round(tapply(impi.all$cweight10,
       impi.all$pred,median, na.rm=T)*10,2)
round(tapply(impi.all$cweight10,
             impi.all$pred,IQR, na.rm=T)*10,2)

summary(impi.all$adurationofsymptoms)
round(tapply(impi.all$adurationofsymptoms,
             impi.all$pred,median, na.rm=T),2)
round(tapply(impi.all$adurationofsymptoms,
             impi.all$pred,IQR, na.rm=T),2)



#H0: no difference between covariate for treatment A compared to treatment B
wilcox.test(impi.all$age~impi.all$pred)

wilcox.test(impi.all$cweight10~impi.all$pred)

wilcox.test(impi.all$adurationofsymptoms~impi.all$pred)


table(impi.all$pred)


# work with means and sd if assuming normality, use t-test
summary(cts)
apply(cts,2, sd, na.rm=TRUE) # For sd, deleting NAs
apply(cts,2, t.test, na.rm=TRUE)
#H0: true mean is equal to 0


#_____________________________________________________
#      working with survival variables
# composite
table(impi.all$event.composite)

#death
table(impi.all$event.death)

#cardiac tamponade
table(impi.all$event.tamponade)

#constriction
```

```
table(impi.all$event.constriction)

# hospitalisation
table(impi.all$event.hospitalzn)

# count HIV status
table(impi_all$hivstat)


library(janitor)
# two-way table for contriction and death
(t1 <- impi.all %>%
  tabyl(event.death, event.constriction))

# two-way table for contriction and death
(t2 <- impi.all %>%
    tabyl(rsa))



#_____________________________________________________________
#      Kaplan-Meier outputs
km_fit <- survfit(Surv(time.composite, event.composite) ~ 1,
                  data=impi.all)

kmi <- rep("KM",length(km_fit$time))
km_df <- data.frame(km_fit$time,km_fit$surv,kmi)
names(km_df) <- c("Time","Surv","Model")


#
km_fit <- survfit(Surv(time.death, event.death) ~ 1,
                  data=impi.all)

kmi <- rep("KM",length(km_fit$time))
km_df <- data.frame(km_fit$time,km_fit$surv,kmi)
names(km_df) <- c("Time","Surv","Model")


#
km_fit <- survfit(Surv(time.constriction, event.constriction) ~ 1,
                  data=impi.all)

kmi <- rep("KM",length(km_fit$time))
km_df <- data.frame(km_fit$time,km_fit$surv,kmi)
names(km_df) <- c("Time","Surv","Model")


#_____________________________________________________________
## Subsetting columns by event of interest
#_____________________________________________________________
impi1 <-
  impi.all %>%
  dplyr::select(id1,pred,perf,age
```

```
                ,country,nyhaclass,adurationofsymptoms
                ,creatn, hivstat
                ,cweight10,fpalpableparadoxus
                ,syblood,hrate
                ,effusize,pulmonaryinfiltrates,definitetbp
                ,hbstat,iperipheraloedema
                ,time.composite, event.composite)

impi1 <- as.data.frame(impi1)

# remove NA values
impi1 <- impi1[complete.cases(impi1), ]


# coercing n replacing some variables from char to factors
char <- c("nyhaclass", "country", "pulmonaryinfiltrates",
          "fpalpableparadoxus", "hrate", "creatn",
          "perf", "pred", "hbstat", "hivstat",
          "definitetbp", "iperipheraloedema",
          "syblood", "effusize")

impi1[char] <- lapply(impi1[char], as.factor)

#####################################################################

impi2 <-
  impi.all %>%
  dplyr::select(id1,pred,perf,age
                ,country,nyhaclass,adurationofsymptoms
                ,creatn, hivstat
                ,cweight10,fpalpableparadoxus
                ,syblood,hrate
                ,effusize,pulmonaryinfiltrates,definitetbp
                ,hbstat,iperipheraloedema
                ,time.death, event.death)

impi2 <- as.data.frame(impi2)

# remove NA values
impi2 <- impi2[complete.cases(impi2), ]


# coercing n replacing some variables from char to factors
char <- c("nyhaclass", "country", "pulmonaryinfiltrates",
          "fpalpableparadoxus", "hrate", "creatn",
          "perf", "pred", "hbstat", "hivstat",
          "definitetbp", "iperipheraloedema",
          "syblood", "effusize")

impi2[char] <- lapply(impi2[char], as.factor)


#             Constriction
#####################################################################
```

```
impi3 <-
  impi.all %>%
  dplyr::select(id1,pred,asex,age,rsa
                ,country,nyhaclass,ecgrthm4
                ,adurationofsymptoms,cweight10
                ,fpalpableparadoxus
                ,iperipheraloedema,syblood
                ,hrate,hbstat
                ,definitetbp,effusize
                ,pulmonaryinfiltrates
                ,hivstat
                ,time.constriction, event.constriction)

impi3 <- as.data.frame(impi3)

# remove NA values
impi3 <- impi3[complete.cases(impi3), ]


# coercing n replacing some variables from char to factors
char <- c("nyhaclass", "country", "asex", "pulmonaryinfiltrates",
          "fpalpableparadoxus", "hrate",
          "pred", "hbstat", "hivstat",
          "definitetbp", "iperipheraloedema",
          "syblood", "effusize")

impi3[char] <- lapply(impi3[char], as.factor)


#####################################################################

# Composite endpoint ----
km_fit1 <- survfit(Surv(time.composite, event.composite==1) ~ 1,
                   data=impi1)

kmi1 <- rep("KM",length(km_fit1$time))
km_comp <- data.frame(km_fit1$time,km_fit1$surv,kmi1)
names(km_comp) <- c("Time","Surv","Model")


# Death endpoint -----
km_fit2 <- survfit(Surv(time.death, event.death==1) ~ 1,
                   data=impi2)

kmi2 <- rep("KM",length(km_fit2$time))
km_death <- data.frame(km_fit2$time,km_fit2$surv,kmi2)
names(km_death) <- c("Time","Surv","Model")


# Constriction endpoint ----
km_fit3 <- survfit(Surv(time.constriction, event.constriction==1) ~ 1,
                   data=impi3)
```

```
kmi3 <- rep("KM",length(km_fit3$time))
km_const <- data.frame(km_fit3$time,km_fit3$surv,kmi3)
names(km_const) <- c("Time","Surv","Model")


# KM plots combined -----
makePlot<-function(){
  plot(km_fit1,
       xlab = "Days",
       ylab = "Overall survival probability");
  lines( km_fit2, col = "green");
  lines( km_fit3, col = "red")
}

makePlot()
legend("bottomleft", legend=c("KM composite", "KM death", "KM constriction"),
       col=c("black", "green","coral3"), lty=1:2, cex=0.8)
```

Univariate IMPI data analysis

```
## Reading in the data
#_____________________________________________________________
impi <-
  read_excel("impi clean.xlsx") %>%
  as_tibble()


#-------------------------------------------------------------
#                          response_composite
#-------------------------------------------------------------
surv_object1 <- Surv(time = impi$response_composite,
                     impi$var1 ==1)

fit <- coxph(surv_object1 ~ hivstat, impi)

# making formulas
univ_formulas1 <- sapply(c("pred","asex","age",
                           "mw", "myco", "perf",
                          "rsa" , "country",
                          "nyhaclass", "ecgrthm4",
                           "adurationofsymptoms", "bonarvs",
                           "oppinf", "cweight10",
                          "fpalpableparadoxus",
                           "iperipheraloedema",
                          "syblood", "hrate", "hivstat",
                           "hbstat", "wbc", "definitetbp",
                          "constriction",
                           "effusize", "pulmonaryinfiltrates",
                           "art_ever" , "lvef" , "creatn" ,
                          "tampon", "creatn"),
                         function(x)
```

```
                          as.formula(paste('surv_object1~',x)))

#"perithick"  Error in `contrasts:contrasts be applied only
# to factors with 2 or more levels
levels(as.factor(impi$perithick))


#making a list of models
univ_models1 <- lapply(univ_formulas1,
                       function(x)
                         {coxph(x,data=impi)})

#extract data (here I've gone for HR and confint)
univ_results1 <- lapply(univ_models1,
                        function(x)
                          {return(summary(x))})
                  #{return(exp(cbind(coef(x),confint(x))))})

univ_results1

save(univ_results1, file = "univariate_composite.RData")


#------------------------------------------------------------
#                          response_death
#------------------------------------------------------------
surv_object2 <- Surv(time = impi$response_death,
                     impi$var2 ==1)

fit2 <- coxph(surv_object2 ~ hivstat, impi)


# making formulas
univ_formulas2 <- sapply(c("pred","asex","age", "mw", "myco",
                           "perf","rsa" , "country",
                           "nyhaclass",
                           "ecgrthm4","adurationofsymptoms",
                           "bonarvs",
                           "oppinf", "cweight10",
                           "fpalpableparadoxus",
                           "iperipheraloedema", "syblood",
                           "hrate", "hivstat",
                           "hbstat", "wbc", "definitetbp",
                           "constriction",
                           "effusize", "pulmonaryinfiltrates",
                           "art_ever" , "lvef" , "creatn" ,
                           "tampon"),
                         function(x)
                        as.formula(paste('surv_object1~',x)))


#making a list of models
univ_models2 <- lapply(univ_formulas2,
                       function(x)
```

```
                        {coxph(x,data=impi)})

#extract data (here I've gone for HR and confint)
univ_results2 <- lapply(univ_models2,
                        function(x)
                        {return(summary(x))})
#{return(exp(cbind(coef(x),confint(x))))})

univ_results2
save(univ_results2, file = "univariate_death.RData")


#-------------------------------------------------------------
#                          response_constriction
#-------------------------------------------------------------
surv_object3 <- Surv(time = impi$response_constriction,
                     impi$var4)
(fit3 <- coxph(surv_object3 ~ country, impi))


# making formulas
univ_formulas3 <- sapply(c("pred","asex","age", "mw", "myco",
                           "perf", "rsa" , "country",
                           "nyhaclass",
                           "ecgrthm4","adurationofsymptoms",
                           "bonarvs",
                           "oppinf", "cweight10",
                           "fpalpableparadoxus",
                           "iperipheraloedema", "syblood",
                           "hrate",
                           "hbstat", "wbc", "definitetbp",
                           "constriction",
                           "effusize", "pulmonaryinfiltrates",
                           "hivstat",
                           "art_ever" , "lvef" , "creatn" ,
                           "tampon"),
                         function(x)
                         as.formula(paste('surv_object1~',x)))


#making a list of models
univ_models3 <- lapply(univ_formulas3,
                       function(x)
                       {coxph(x,data=impi)})

#extract data (here I've gone for HR and confint)
univ_results3 <- lapply(univ_models3,
                        function(x)
                        {return(summary(x))})
#{return(exp(cbind(coef(x),confint(x))))})

univ_results3
save(univ_results3, file = "univariate_constriction.RData")
```

Multicentre clinical trial Cox PH data analysis

Presented below is the code that was used in the Cox PH model analysis for the time to composite event, the same code was appended for all other events including the simulated data sets.

```
rm(list=ls())

# Installing all the packages I will need for data processing
pack.load<-c("BeSS",  #for best selection and gen.data
             "boot", #to bootstrap on other selection objects
             "bootStepAIC", #bootstrap AIC
             "caret", #for easy Cross-Validation
             "caTools", #split data into train set n test set
             "coxed",#for fitting Cox PH model and simulation
             "GGally", #extension of ggplot2
             "glmnet", # for penalised regression
             "Hmisc",
             "lattice",#plotting xyplot
             "leaps",
             "MASS", #for stepAIC function
             "My.stepwise",
             "openxlsx",#exporting/importing Excel files
             "pec", # for calibration slope
             "ranger", # for survival random forest
             "readxl",
             "rms", #for validate() used for bootstrap
             "SIS", #for datasets wit many covars than sample
             "skimr",#for summary table showin missing values
             "survAUC", #AUC
             "survival", #for fitting Cox PH model
             "survminer",
             "survMisc", #for r-sqrd measures
             "tidyverse",#data wrangling functions
             "todor", #manage R comments
             "VIM", #helps with visualising missing data
             "xtable"#write tables in tex
)

if (!require("rmsfuns")) install.packages("rmsfuns")
library(rmsfuns)
load_pkg(pack.load)

## Reading in the data
#-------------------------------------------------------------
impi.all <-
  read_excel("impi clean.xlsx") %>%
  as_tibble()


#-------------------------------------------------------------
## Subsetting columns
#-------------------------------------------------------------
```

```
impi <-
  impi.all %>%
  dplyr::select(id1,pred,perf,age
                ,country,nyhaclass,adurationofsymptoms
                ,creatn, hivstat
                ,cweight10,fpalpableparadoxus
                ,syblood,hrate
                ,effusize,pulmonaryinfiltrates,definitetbp
                ,hbstat,iperipheraloedema
                ,time.composite, event.composite)

impi <- as.data.frame(impi)


#checking for missing values
mice.plot1 <- aggr(impi, col=c('navyblue','yellow'),
                   numbers=TRUE, sortVars=TRUE,
                   labels=names(impi), cex.axis=.7,
                   gap=3, ylab=c("Missing data","Pattern"))


# remove NA values
impi <- impi[complete.cases(impi), ]
str(impi)


plot(sort(impi$time.composite),
     pch =".",
     type = "o",
     color = "blue",
     lwd = 2)


# coercing n replacing some variables from char to factors
char <- c("nyhaclass", "country", "pulmonaryinfiltrates",
          "fpalpableparadoxus", "hrate", "creatn",
          "perf", "pred", "hbstat", "hivstat",
          "definitetbp", "iperipheraloedema",
          "syblood", "effusize")

impi[char] <- lapply(impi[char], as.factor)
sapply(impi, class)


# split data into train set and test set
set.seed(123)
sample <- sample.int(n = nrow(impi),
                     size = floor(.75*nrow(impi)),
                     replace = F)
impi. <- impi[sample, ]
test  <- impi[-sample, ]

#-------------------------------------------------------------#
#-------------------------------------------------------------#
```

```
#             Object creation and naive cox ph
#-------------------------------------------------------------
surv.composite <- Surv(time = impi.$time.composite,
                       event = impi.$event.composite ==1)



fit.composite <- coxph(surv.composite ~ pred+perf+age
                       +nyhaclass+adurationofsymptoms
                       +creatn+hivstat
                       +cweight10+fpalpableparadoxus
                       +syblood+hrate
                       +effusize+pulmonaryinfiltrates
                       +definitetbp
                       +hbstat+iperipheraloedema,
                       x=T, y=T,
                       data = impi.)

summary(fit.composite)$conf.int


# extract C-index
survConcordance(surv.composite ~ predict(fit.composite))

# get r-sqrd measures
rsq(fit.composite)

# AUC estimator proposed by Song and Zhou
lp <- predict(fit.composite)
lpnew <- predict(fit.composite,
                 newdata = test)
Surv.rsp <- Surv(impi.$time.composite,
                 impi.$event.composite==1)
Surv.rsp.new <- Surv(test$time.composite,
                     test$event.composite==1)
times <- seq(10, 1000, 10)

AUC <- AUC.uno(Surv.rsp, Surv.rsp.new,  lpnew, times)
names(AUC)
AUC$iauc

plot(AUC)

#-------#
#Overall model fit using cox-snell residuals
# First calculate Martingale residuals
comp.resid <- resid(fit.composite, type="martingale")

# We subtract these residuals from actual values of the event to get
# Cox-snell residuals.
comp.res <- impi.$event.composite - comp.resid

# Compute S(t)
comp.surv <- survfit(Surv(comp.res, impi.$event.composite ==1)~1)
```

```
# Plot integrated hazard function, H(t)=-log(S(t)), on the y-axis
plot(comp.surv$time,
     -log(comp.surv$surv),
     type = "l",
     xlab="Time",
     ylab = "H(t) based on full model residuals"
)
lines(comp.res, comp.res, type = "l")


#-------------------------------------------------------------#
#-------------------------------------------------------------
#                      Classical techniques
#-------------------------------------------------------------

#                      Stepwise procedure
#-------------------------------------------------------------#

# factor variables # fix "not found error"
impi.stepdata <- impi.
impi.stepdata$iperipheraloedemaYes <- ifelse(impi.stepdata$iperipheraloedema
                                             == "Yes", 1, 0)
impi.stepdata$iperipheraloedemaNo <- ifelse(impi.stepdata$iperipheraloedema
                                            == "No", 1, 0)

impi.stepdata$predPlacebo <- ifelse(impi.stepdata$pred
                                    == "Placebo", 1, 0)
impi.stepdata$predPrednisolone <- ifelse(impi.stepdata$pred
                                         == "Prednisolone", 1, 0)

impi.stepdata$perf1 <- ifelse(impi.stepdata$perf == "1", 1, 0)
impi.stepdata$perf0 <- ifelse(impi.stepdata$perf == "0", 1, 0)


impi.stepdata$nyhaclass1 <- ifelse(impi.stepdata$nyhaclass == "NYHA Class I",
                                   1, 0)
impi.stepdata$nyhaclass2 <- ifelse(impi.stepdata$nyhaclass == "NYHA Class II",
                                   1, 0)
impi.stepdata$nyhaclass3 <- ifelse(impi.stepdata$nyhaclass == "NYHA Class III",
                                   1, 0)
impi.stepdata$nyhaclass4 <- ifelse(impi.stepdata$nyhaclass == "NYHA Class IV",
                                   1, 0)

# stepwise cox modelling
impi.step <- stepAIC(fit.composite,
                     direction
                     = c("both"))

summary(impi.step)

# performance measures
survConcordance(surv.composite ~ predict(impi.step))
```

```
rsq(impi.step)


# AUC estimator proposed by Song and Zhou
lp <- predict(impi.step)
lpnew <- predict(impi.step,
                 newdata = test)
Surv.rsp <- Surv(impi.$time.composite,
                 impi.$event.composite==1)
Surv.rsp.new <- Surv(test$time.composite,
                     test$event.composite==1)
times <- seq(10, 1000, 10)

AUCstep <- AUC.sh(Surv.rsp, Surv.rsp.new, lp, lpnew, times)
names(AUC)
AUCstep$iauc

#--------#
#Overall model fit using cox-snell residuals
# First calculate Martingale residuals
step.resid <- resid(impi.step, type="martingale")

# We subtract these residuals from actual values of the event to get
# Cox-snell residuals.
step.res <- impi.$event.composite - step.resid

# Compute S(t)
step.surv <- survfit(Surv(step.res, impi.$event.composite ==1)~1)

# Plot integrated hazard function, H(t)=-log(S(t)), on the y-axis
plot(step.surv$time,
     -log(step.surv$surv),
     type = "l",
     xlab="Time",
     ylab = "H(t) based on stepwise residuals"
)
lines(step.res, step.res, type = "l")



#                     Forward selection
#-------------------------------------------------------------#

impi.fore <- step(fit.composite, scale = 0,
                  direction = c("forward"),
                  trace = 1, keep = NULL, steps = 1000, k = 2)

summary(impi.fore)$conf.int

# performance measures
survConcordance(surv.composite ~ predict(impi.fore))

rsq(impi.fore)
```

```
# AUC estimator proposed by Song and Zhou
lp <- predict(impi.fore)
lpnew <- predict(impi.fore,
                 newdata = test)
Surv.rsp <- Surv(impi.$time.composite,
                 impi.$event.composite==1)
Surv.rsp.new <- Surv(test$time.composite,
                     test$event.composite==1)
times <- seq(10, 1000, 10)

AUCfore <- AUC.sh(Surv.rsp, Surv.rsp.new, lp, lpnew, times)
names(AUC)
AUCfore$iauc

#-------#
#Overall model fit using cox-snell residuals
# First calculate Martingale residuals
fore.resid <- resid(impi.fore, type="martingale")

# We subtract these residuals from actual values of the event to get
# Cox-snell residuals.
fore.res <- impi.$event.composite - fore.resid

# Compute S(t)
fore.surv <- survfit(Surv(fore.res, impi.$event.composite ==1)~1)

# Plot integrated hazard function, H(t)=-log(S(t)), on the y-axis
plot(fore.surv$time,
     -log(fore.surv$surv),
     type = "l",
     xlab="Time",
     ylab = "H(t) based on forward selection residuals"
)
lines(fore.res, fore.res, type = "l")


#                     Backward selection
#-------------------------------------------------------------#
# wrapper fn which 1st selects variables in Cox regression model
# using fastbw from the rms package and then
# returns a fitted Cox regression model with selected variables.
impi.back <- selectCox(surv.composite ~
                        pred+perf+age
                       +nyhaclass+adurationofsymptoms
                       +creatn+hivstat
                       +cweight10+fpalpableparadoxus
                       +iperipheraloedema+syblood+hrate
                       +effusize+pulmonaryinfiltrates
                       +definitetbp
                       +hbstat,
                       data = impi.)

backward <- impi.back$fit
```

```
# fit Cox PH to get confident intervals
fit.back <- coxph(surv.composite ~ pred+
                    +nyhaclass
                  +creatn
                  +syblood+iperipheraloedema,
                  x=T, y=T,
                  data = impi.)

summary(fit.back)$conf.int

# performance measures
survConcordance(surv.composite ~ predict(backward))

rsq(backward)

# AUC estimator proposed by Song and Zhou
lp <- predict(backward)
lpnew <- predict(backward,
                 newdata = test)
Surv.rsp <- Surv(impi.$time.composite,
                 impi.$event.composite==1)
Surv.rsp.new <- Surv(test$time.composite,
                     test$event.composite==1)
times <- seq(10, 1000, 10)

AUCback <- AUC.sh(Surv.rsp, Surv.rsp.new, lp, lpnew, times)
names(AUCback)
AUCback$iauc


#Overall model fit using cox-snell residuals
#-------#
# First calculate Martingale residuals
back.resid <- resid(backward, type="martingale")

# We subtract these residuals from actual values of the event to get
# Cox-snell residuals.
back.res <- impi.$event.composite - back.resid

# Compute S(t)
back.surv <- survfit(Surv(back.res, impi.$event.composite ==1)~1)

# Plot integrated hazard function, H(t)=-log(S(t)), on the y-axis
plot(back.surv$time,
     -log(back.surv$surv),
     type = "l",
     xlab="Time",
     ylab = "H(t) based on backward selection residuals"
)
lines(back.res, back.res, type = "l")
```

```
#                  Best subset selection
#-------------------------------------------------------------#
library(BeSS)

impi.new <-
  impi. %>%
  dplyr::select(pred,perf,age
                ,country,nyhaclass,adurationofsymptoms
                ,creatn, hivstat
                ,cweight10,fpalpableparadoxus
                ,syblood,hrate
                ,effusize,pulmonaryinfiltrates,definitetbp
                ,hbstat,iperipheraloedema
                ,time.composite, event.composite)



X <- as.matrix(impi.new[,1:17])
Y <- as.matrix(impi.new[,18:19])


# mode(x) shows character matrix, change to numeric
X[[1]] <- as.numeric(X[[1]])
Y[[1]] <- as.numeric(Y[[1]])


# get the matrix
x <- data.matrix(X)
y <- data.matrix(Y)



impi.bess <- bess(y, x,
                  method = "sequential",family = "cox")
print(impi.bess)
coef(impi.bess, sparse = TRUE)
bestmodel <- impi.bess$bestmodel

bess <- summary(bestmodel)


# extract C-index
survConcordance(simobject ~ predict(bestmodel))

# get r-sqrd measures
rsq(bestmodel)

# AUC estimator proposed by Song and Zhou
lp <- predict(bestmodel)
lpnew <- predict(bestmodel,
                 newdata = test)
Surv.rsp <- Surv(dat2$y,
                 dat2$failed==1)
Surv.rsp.new <- Surv(test2$y,
```

```
                    test2$failed==1)
times <- seq(10, 1000, 10)

AUC_fore <- AUC.uno(Surv.rsp, Surv.rsp.new,  lpnew, times)
names(AUC_fore)
AUC_fore$iauc



#-------------------------------------------------------------#
#-------------------------------------------------------------
#                    Probability techniques
#-------------------------------------------------------------


#                  Akaike Information Criteria
#-------------------------------------------------------------#
# with bootstrap

bootAIC <- boot.stepAIC(fit.composite,
                        impi.,
                        B = 500,
                        k = 2)

summary(bootAIC$OrigStepAIC)

fit.aic1 <- bootAIC$OrigStepAIC

summary(fit.aic1)$conf.int

# performance measures
survConcordance(surv.composite ~ predict(fit.aic1))

rsq(fit.aic1)

# AUC estimator proposed by Song and Zhou
lp <- predict(fit.aic1)
lpnew <- predict(fit.aic1,
                 newdata = test)
Surv.rsp <- Surv(impi.$time.composite,
                 impi.$event.composite==1)
Surv.rsp.new <- Surv(test$time.composite,
                     test$event.composite==1)
times <- seq(10, 1000, 10)

AUC_aic1 <- AUC.sh(Surv.rsp, Surv.rsp.new, lp, lpnew, times)
names(AUC_aic1)
AUC_aic1$iauc

#Overall model fit using cox-snell residuals
#-------#
# First calculate Martingale residuals
aic.resid <- resid(backward, type="martingale")
```

```
# We subtract these residuals from actual values of the event to get
# Cox-snell residuals.
aic.res <- impi.$event.composite - aic.resid

# Compute S(t)
aic.surv <- survfit(Surv(aic.res, impi.$event.composite ==1)~1)

# Plot integrated hazard function, H(t)=-log(S(t)), on the y-axis

plot(aic.surv$time,
     -log(aic.surv$surv),
     type = "l",
     xlab="Time",
     ylab = "H(t) based on AIC residuals"
)
lines(aic.res, aic.res, type = "l")


#               Bayesian Information Criteria
#-------------------------------------------------------------#
# Create vectors for outcome and predictors
outcome    <- c("surv.composite")

predictors <- c("pred",
                "perf",
                "age",
                "nyhaclass",
                "adurationofsymptoms",
                "creatn",
                "cweight10",
                "fpalpableparadoxus",
                "syblood",
                "hrate",
                "effusize",
                "pulmonaryinfiltrates",
                "definitetbp",
                "hbstat",
                "hivstat",
                "iperipheraloedemaYes",
                "iperipheraloedemaNo")

dataset    <- impi.


## Create list of models
list.of.models <- lapply(seq_along((predictors)),
                         function(n) {

                           left.hand.side  <- outcome
                           right.hand.side <- apply(X = combn(predictors, n),
                                                    MARGIN = 2, paste, collapse = "+")

                           paste(left.hand.side, right.hand.side, sep = "  ~  ")
                         })
```

```
## Convert to a vector
vector.of.models <- unlist(list.of.models)

## Fit coxph to all models
list.of.fits <- lapply(vector.of.models, function(x) {

  formula    <- as.formula(x)
  fit        <- coxph(formula, data = dataset)
  result.BIC <- extractAIC(fit, k=log(16))

  data.frame(num.predictors = result.BIC[1],
             BIC            = result.BIC[2],
             model          = x)
})

## Collapse to a data frame
result1 <- do.call(rbind, list.of.fits)


## Sort and print
library(doBy)
BICorder <- orderBy(~ BIC, result1)


#fit cox PH on selected model giving lowest BIC
fit.bic <- coxph(surv.composite  ~  nyhaclass + creatn
                 + cweight10 + pulmonaryinfiltrates
                 + definitetbp + hbstat
                 + iperipheraloedema,
                 data = impi.)

confint(fit.bic)

summary(fit.bic)$conf.int

survConcordance(surv.composite ~ predict(fit.bic))
rsq(fit.bic)


# AUC estimator proposed by Song and Zhou
lp <- predict(fit.bic)
lpnew <- predict(fit.bic,
                 newdata = test)
Surv.rsp <- Surv(impi.$time.composite,
                 impi.$event.composite==1)
Surv.rsp.new <- Surv(test$time.composite,
                     test$event.composite==1)
times <- seq(10, 1000, 10)

AUC_bic <- AUC.sh(Surv.rsp, Surv.rsp.new, lp, lpnew, times)
names(AUC_bic)
AUC_bic$iauc
```

```
#Overall model fit using cox-snell residuals
#-------#
# First calculate Martingale residuals
bic.resid <- resid(fit.bic, type="martingale")

# We subtract these residuals from actual values of the event to get
# Cox-snell residuals.
bic.res <- impi.$event.composite - bic.resid

# Compute S(t)
bic.surv <- survfit(Surv(bic.res, impi.$event.composite ==1)~1)

# Plot integrated hazard function, H(t)=-log(S(t)), on the y-axis
plot(bic.surv$time,
     -log(bic.surv$surv),
     type = "l",
     xlab="Time",
     ylab = "H(t) based on BIC residuals"
)
lines(bic.res, bic.res, type = "l")


#-------------------------------------------------------------#
#-----------------------------------------------------------------
#                      Penalised Techniques
#-----------------------------------------------------------------

# http://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html#cox

#                        Lasso regression
#-------------------------------------------------------------#

x <- model.matrix( ~pred+perf+age
                   +nyhaclass+adurationofsymptoms
                   +creatn+hivstat
                   +cweight10+fpalpableparadoxus
                   +syblood+hrate
                   +effusize+pulmonaryinfiltrates
                   +definitetbp
                   +hbstat+iperipheraloedema-1, impi)


y <- Surv(impi$time.composite,
          1-(impi$event.composite-1))


# Split data into train and test sets
n <- nrow(impi)
train_rows <- sample(1:n, .66*n)
x.train <- x[train_rows, ]
x.test <- x[-train_rows, ]

y.train <- y[train_rows]
```

```
y.test <- y[-train_rows]


# fit the models
fit.lasso <- glmnet(x.train, y.train,
                    family="cox",
                    alpha=1)
summary(fit.lasso)
# Plot variable coefficients vs. shrinkage parameter lambda
plot(fit.lasso,
     label=TRUE,
     xvar = "lambda")

# n extract the coefficients at certain values of
coef(fit.lasso,
     s = 0.00120739)

#get the exact lambda using cross validation
set.seed(1)
cv.lasso <- cv.glmnet(x.train, y.train,
                      type.measure="C",
                      family="cox",
                      alpha=1)
plot(cv.lasso$glmnet.fit,
     label = TRUE,
     "norm")


# 10-fold CV for each alpha = 0, 0.1, ....,10
set.seed(2)
for (i in 0:10) {
  assign(paste("fit", i, sep=""),
         cv.glmnet(x.train, y.train,
                   type.measure="C",
                   alpha=i/10,family="cox"))
}

# Plot solution paths:
par(mfrow=c(1,2))
plot(fit.lasso, xvar="lambda")
plot(fit10)
mtext("LASSO",
      side = 3, line = -2, outer = TRUE)
par(mfrow=c(1,1))


opt.lambda <- cv.lasso$lambda.min
opt.lambda

# Get the non-zero variables
coef.lasso = as.matrix(coef(cv.lasso))
ixl = which(abs(coef.lasso[,1]) > 0)
length(ixl)
```

```
coef.lasso[ixl,1, drop=FALSE]
# seems like Lasso's penalty too strict; only few vars selected


# fit cox model on covariates selected
fit.lasso <- coxph(surv.composite ~ nyhaclass + creatn
                   + fpalpableparadoxus + effusize,
                   data = impi.)

summary(fit.lasso)$conf.int

survConcordance(surv.composite ~ predict(fit.lasso))
rsq(fit.lasso)

# AUC estimator proposed by Song and Zhou
lp <- predict(fit.lasso)
lpnew <- predict(fit.lasso,
                 newdata = test)
Surv.rsp <- Surv(impi.$time.composite,
                 impi.$event.composite==1)
Surv.rsp.new <- Surv(test$time.composite,
                     test$event.composite==1)
times <- seq(10, 1000, 10)

AUC_lasso <- AUC.sh(Surv.rsp, Surv.rsp.new, lp, lpnew, times)
names(AUC_lasso)
AUC_lasso$iauc

#Overall model fit using cox-snell residuals
#-------#
# First calculate Martingale residuals
lasso.resid <- resid(fit.lasso, type="martingale")

# We subtract these residuals from actual values of the event to get
# Cox-snell residuals.
lasso.res <- impi.$event.composite - lasso.resid

# Compute S(t)
lasso.surv <- survfit(Surv(lasso.res, impi.$event.composite ==1)~1)

# Plot integrated hazard function, H(t)=-log(S(t)), on the y-axis
plot(lasso.surv$time,
     -log(lasso.surv$surv),
     type = "l",
     xlab="Time",
     ylab = "H(t) based on lasso residuals"
)
lines(lasso.res, lasso.res, type = "l")



#                          Ridge regression
#-------------------------------------------------------------#
lambdas <- 10^seq(3, -2, by = -.1)
```

```
fit.ridge <- glmnet(x, y,
                    family="cox",
                    alpha=0,
                    lambda = lambdas)
plot(fit.ridge,
     label=TRUE,
     xvar = "lambda")

coef(fit.ridge,
     s = 0.5)



# get C-index
pred = predict(fit.ridge, newx = x)
apply(pred, 2, Cindex, y=y)

cv.ridge <- cv.glmnet(x, y,
                      family="cox",
                      alpha=0,
                      type.measure = "C",
                      lambda = lambdas)

plot(cv.ridge$glmnet.fit,
     label = TRUE,
     "norm")

plot(cv.ridge)

# Get the non-zero variables
#convert to matrix (618x1)
coef.ridge = as.matrix(coef(cv.ridge))
ix = which(abs(coef.ridge[,1]) > 0)
length(ix)

coef.ridge[ix,1, drop=FALSE]


##
surv.ridge <- coxph(surv.composite ~  nyhaclass
                    +creatn
                    +cweight10
                    +syblood
                    +definitetbp
                    +iperipheraloedema,
                    data = impi.)

confint(surv.ridge)

survConcordance(surv.composite ~ predict(surv.ridge))
rsq(surv.ridge)
```

```
# AUC estimator proposed by Song and Zhou
lp <- predict(surv.ridge)
lpnew <- predict(surv.ridge,
                 newdata = test)
Surv.rsp <- Surv(impi.$time.composite,
                 impi.$event.composite==1)
Surv.rsp.new <- Surv(test$time.composite,
                     test$event.composite==1)
times <- seq(10, 1000, 10)

AUC_ridge <- AUC.sh(Surv.rsp, Surv.rsp.new, lp, lpnew, times)
names(AUC_ridge)

AUC_ridge$iauc


#Overall model fit using cox-snell residuals
#-------#
# First calculate Martingale residuals
ridge.resid <- resid(surv.ridge, type="martingale")

# We subtract these residuals from actual values of the event to get
# Cox-snell residuals.
ridge.res <- impi.$event.composite - ridge.resid

# Compute S(t)
ridge.surv <- survfit(Surv(ridge.res, impi.$event.composite ==1)~1)

# Plot integrated hazard function, H(t)=-log(S(t)), on the y-axis
plot(ridge.surv$time,
     -log(ridge.surv$surv),
     type = "l",
     xlab="Time",
     ylab = "H(t) based on ridge residuals"
)
lines(ridge.res, ridge.res, type = "l")


#https://rstudio-pubs-static.s3.amazonaws.com
#/248427_03e0cd90980c4404b86ebcf666e84be6.html


#-----------------------------------------------------------------#
#-----------------------------------------------------------------
#                    Survival Random Forest
#-----------------------------------------------------------------

# ranger model
r_fit <- ranger(Surv(time.composite, event.composite) ~ pred
                +perf+age
                +nyhaclass+adurationofsymptoms
                +creatn+hivstat
                +cweight10+fpalpableparadoxus
                +syblood+hrate
```

```
                +effusize+pulmonaryinfiltrates
                +definitetbp
                +hbstat+iperipheraloedema,
                data = impi,
                mtry = 4,
                num.trees = 1000,
                importance = "permutation",
                splitrule = "extratrees",
                verbose = TRUE,
                seed = 123)


# Average the survival models
event_times <- r_fit$unique.death.times
surv_prob <- data.frame(r_fit$survival)
avg_prob <- sapply(surv_prob, mean)


# Plot the survival models for each patient
plot(event_times,surv_prob[1,],
     type = "l",
     ylim = c(0,1),
     col = "red",
     xlab = "Days",
     ylab = "survival",
     main = "Patient Survival Curves")
#
cols <- colors()

for (n in sample(c(2:dim(impi)[1]), 20))
{
  lines(event_times,
        surv_prob[n,],
        type = "l",
        col = cols[n])
}
lines(event_times,
      avg_prob,
      lwd = 2,
      label = TRUE)
legend(500, 0.5, legend = c('Average = black'))


# ranking by variable importance
vi <- data.frame(sort(round(r_fit$variable.importance, 4),
                      decreasing = TRUE))
names(vi) <- "importance"
vi

# plot vi vars dots
vim.p <- vi
vim.p <- cbind(variables = rownames(vim.p),
               vim.p)
rownames(vim.p) <- NULL
```

```
dotchart(vim.p$importance, labels = vim.p$variables, pch = 21,
         bg = "magenta", pt.cex = 1.1,
         sort(vim.p$importance, decreasing = F),
         main = "Time to composite RSF variable importance")

# plot vi vars on a barplot
barplot(r_fit$variable.importance,
        horiz = TRUE,
        las = 1)


cat("Prediction Error = 1 - Harrell's c-index = ",
    r_fit$prediction.error)


fit.rf <- coxph(surv.composite ~ pred+age
                +nyhaclass+adurationofsymptoms
                +cweight10+fpalpableparadoxus
                +syblood+hrate
                +effusize
                +definitetbp
                +iperipheraloedema,
                x=T, y=T,
                data = impi.)

summary(fit.rf)$conf.int


# extract C-index
survConcordance(surv.composite ~ predict(fit.rf))

# get r-sqrd measures
rsq(fit.rf)

# AUC estimator proposed by Song and Zhou
lp <- predict(fit.rf)
lpnew <- predict(fit.rf,
                 newdata = test)
Surv.rsp <- Surv(impi.$time.composite,
                 impi.$event.composite==1)
Surv.rsp.new <- Surv(test$time.composite,
                     test$event.composite==1)
times <- seq(10, 1000, 10)

AUC_rf <- AUC.uno(Surv.rsp, Surv.rsp.new,  lpnew, times)
names(AUC_rf)
AUC_rf$iauc


#Overall model fit using cox-snell residuals
#-------#
# First calculate Martingale residuals
rf.resid <- resid(fit.rf, type="martingale")
```

```
# We subtract these residuals from actual values of the event to get
# Cox-snell residuals.
rf.res <- impi.$event.composite - rf.resid

# Compute S(t)
rf.surv <- survfit(Surv(rf.res, impi.$event.composite ==1)~1)

# Plot integrated hazard function, H(t)=-log(S(t)), on the y-axis
plot(rf.surv$time,
     -log(rf.surv$surv),
     type = "l",
     xlab="Time",
     ylab = "H(t) based on random forest residuals"
)
lines(rf.res, rf.res, type = "l")


#-------------------------------------------------------------#
#-------------------------------------------------------------
#                  OTHER PLOTS
#-------------------------------------------------------------
# pull out the survival data from the cox and rsf model
# objects and puts them into a data frame for ggplot()

cox_fit <- survfit(fit.composite)

coxi <- rep("Cox",length(cox_fit$time))
cox_df <- data.frame(cox_fit$time,cox_fit$surv,coxi)
names(cox_df) <- c("Time","Surv","Model")

#
rfi <- rep("RF",length(r_fit$unique.death.times))
rf_df <- data.frame(r_fit$unique.death.times,avg_prob,rfi)
names(rf_df) <- c("Time","Surv","Model")

plot_df <- rbind(cox_df, rf_df)


p <- ggplot(plot_df, aes(x = Time, y = Surv, color = Model))
p + geom_line()

# rviews.rstudio.com/2017/09/25/survival-analysis-with-r/

#-------------------------------------------------------------
#-------------------------------------------------------------

# AUC plots for all techniques
plot.auc <- rbind(AUC, AUCstep, AUCfore,
                  AUCback, AUC_aic1, AUC_bic,
                  AUC_lasso, AUC_ridge, AUC_rf)

plot(AUC$auc, type = "l", ylim=c(0.42, 0.68), ylab = "AUC");
par(new=T); plot( AUCstep$auc, col = "green", type = "b",
```

```
                  ylim=c(0.42, 0.68), ylab = "AUC");
par(new=T); plot(AUCfore$auc, col = "red", type = "l",
                  ylim=c(0.42, 0.68), ylab = "AUC");
par(new=T); plot(AUCback$auc, col = "blue", type = "l",
                  ylim=c(0.42, 0.68), ylab = "AUC");
par(new=T); plot(AUC_aic1$auc, col = "orange", type = "l",
                  ylim=c(0.42, 0.68), ylab = "AUC");
par(new=T); plot(AUC_bic$auc, col = "purple", type = "l",
                  ylim=c(0.42, 0.68), ylab = "AUC");
par(new=T); plot(AUC_lasso$auc, col = "magenta", type = "l",
                  ylim=c(0.42, 0.68), ylab = "AUC");
par(new=T); plot(AUC_ridge$auc, col = "grey39", type = "l",
                  ylim=c(0.42, 0.68), ylab = "AUC");
par(new=T); plot(AUC_rf$auc, col = "coral3", type = "l",
                  ylim=c(0.42, 0.68), ylab = "AUC")
legend("bottom",
       legend=c("full", "stepwise", "forward",
                "backward", "AIC", "BIC", "Lasso",
                "Ridge", "RF"),
       col=c("black", "green","red", "blue", "orange",
             "purple", "magenta", "grey39", "coral3"),
       lty=1:2, cex=0.6)

#-----------------------------------------------------------------
# forest plots for final models

ggforest(impi.fore,
         main = "Forest plot for time to composite forward selection coxph model",
         cpositions = c(0.02, 0.22, 0.4),
         fontsize = 0.6,
         refLabel = "reference")


#-----------------------------------------------------------------
# Hazard proportionality assumption test

prop <- cox.zph(impi.fore, transform="km", global=TRUE)

plot(cox.zph(impi.fore, transform="km", global=TRUE))
```

Simulated data

Presented is the code that was used to generate the simulated data set.

```
## Generate simulated data
set.seed(1)
simdata <- sim.survdata(N=2000,
                        T=150,
                        xvars=15,
                        censor=0.2,
                        num.data.frames=1)
attributes(simdata)
```

```
#require(ggplot2)
#check if plot is close to what you know as truth for survival
ggplot(data.frame(simdata), aes(simdata)) +
  geom_histogram(aes(y=..density..)) +
  stat_function(fun=function(x)dgamma(x, shape=nexps,
                                      scale=1/rate),
                color="red", size=2)


# simulated data plots
(lineplots <-  survsim.plot(simdata,
                            df=1,
                            type="baseline"))
(histplots <- survsim.plot(simdata,
                           df=1,
                           type="hist",
                           bins = 50))


# writing the output in Excel
write.xlsx(x = simdata$data,
           file = "simulate.xlsx")
openXL("simulate.xlsx")


# put the data in a dataframe
datt <- as.data.frame(simdata$data)
head(datt)



# split data into train set and test set
set.seed(123)
sample <- sample.int(n = nrow(datt),
                     size = floor(.75*nrow(datt)),
                     replace = F)
dat <- datt[sample, ]
test  <- datt[-sample, ]
```