

https://dspace.org/

A study of risk factors for acute myeloid leukemia using parametric and semi parametric models

Muchatibaya, M C

2022

https://hdl.handle.net/10539/33920

Downloaded from a DSpace repository. https://dspace.org/

SCHOOL OF STATISTICS AND ACTUARIAL SCIENCES

MASTERS RESEACH REPORT IN MATHEMATICAL STATISTICS

A study of risk factors for acute myeloid leukemia using parametric and semi parametric models



Student:
MC MUCHATIBAYA
(1259754)

Supervisor:
Dr Jacob Majakwara

A research report submitted to the Faculty of Science, University of the Witwatersrand, in partial fulfilment of the requirements for the degree of Master of Science

February 9, 2022

Dedication

I would like to dedicate this research report to God Almighty,
my parents, Maxwell and Itayi Muchatibaya;
and siblings, David, Desmond and Michelle Muchatibaya.
This research report would not have been possible without your support and inspiration.
I love you all!

Acknowledgement

I would like to firstly thank my God Almighty who got me this far in my life. I want to give all praise to Him for bringing all the right people into my life in order for this degree to be possible.

To my uncle, Dr Gift Muchatibaya, thank you for believing in me and gaining this opportunity for me at the University of Witwatersrand. I was able to study and work amongst great statisticians because you saw it fit that I be granted the opportunity. May the dear Lord bless you.

I am grateful to my supervisor, Dr Jacob Majakwara. Thank you for pushing me to always aim for better and for helping me see beyond my own capabilities – your support and guidance is truly appreciated. I am also grateful for the support I received from Dr Honest Chipoyera and Dr Charles Chimedza throughout my time at the University.

Lastly, to my parents, siblings and friends, thank you for always being there for me. The love you have for me pushed me forward. I am truly blessed to have such a wonderful support system. I am deeply indebted to God for giving me such an amazing family and friends. Thank you for every cheer and for every little form of encouragement and support you have offered me over the years. May my Lord Almighty bless you all.

Declaration

I, Maxeen Muchatibaya, declare that this research report is my own, unaided work. It is being submitted for the degree of Master of Science by coursework and research report at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other university.

Signature: 
Full name: Maxeen Chiedza Muchatibaya
Date: 10 February 2022

Abstract

This research report involves the study of a dataset compiled at a single cancer centre of patients with the chronic disease known as Acute Myeloid Leukemia (AML). A semi-parametric model (i.e., the Cox Proportional Hazard (PH)) and four parametric models, namely: exponential, Weibull, lognormal, and the log-logistic were fitted to the data. In fitting the survival models, variables such as patient age at diagnosis, sex, hemoglobin levels, cytogenic categories, and infection status, as well as whether or not the patient had chemotherapy before treatment, were found to be significant in the models. Based on information criteria and forecast error metrics, the Cox PH model, the semi-parametric model performed best in comparison to the parametric models. The Cox PH model had the smallest Akaike's information criterion (AIC) and Bayesian information criterion (BIC) values and Integrated Brier Score (IBS). The Cox PH model gave the best predictions.

List of Acronyms

AIC	Akaike's information criterion
ALL	acute lymphoblastic leukemia
AML	acute myeloid leukaemia
ara-C	Cytosine Arabinoside
BIC	Bayesian information criterion
cdf	cummulative density function
CRF	chronic renal failure
DNA	deoxyribonucleic acid
HLA	human leukocyte antigen
IBS	Integrated Brier Score
IgA-GN	IgA glomerulonephritis
IQR	Interquartile Range
MAD	Mean Absolute Deviation
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MDA	Mean Decrease Accuracy
MDG	Mean Decrease Gini
MDS	myelodsyplastic syndrome
MLE	maximum likelihood estimate
pdf	probability density function
PH	proportional hazard
RMSE	Root Mean Square Error
RNA	ribonucleic acid
RPPA	Reverse Phase Protein Array
sd	standard deviation
se	standard error
USA	United States of America

Contents

1 Introduction **1**

1.1 Statement of the problem 5

1.2 Aim and Objectives 6

1.3 Significance of the Study 6

2 Literature Review **8**

2.1 Introduction 8

2.2 Related Studies 10

3 Methodology **14**

3.1 Introduction 14

3.1.1 Survival function 14

3.1.2 Hazard function 15

3.2 Survival models 15

3.2.1 Non-parametric model: Kaplan-Meier estimator 16

3.2.2 Semi-parametric model: Cox PH model 17

3.2.3 Parametric models 18

3.3 Parameter estimation 21

3.3.1 Censoring 21

3.3.2 Partial likelihood 22

3.3.3 Maximum likelihood estimate 22

3.4 Model development 23

3.4.1 Variable selection 23

3.4.2 Information based criteria 26

3.4.3 Model performance evaluation techniques 27
3.4.4 Forecast measures 28
3.4.5 Model diagnostics 29
3.5 Data source 30
3.5.1 Risk factors for Acute Myeloid Leukemia with missing values 32
3.5.2 Risk factors for Acute Myeloid Leukemia without missing values 33
3.6 Variable Selection 36
3.6.1 Random forests 36
3.7 Descriptive Statistics 37
3.7.1 Gender and Age at death 37
3.7.2 Overall Survival and Survival Status 37
3.7.3 Stepwise regression 40

4 Data Analysis and Results **41**
4.1 Introduction 41
4.2 Kaplan Meier estimator 42
4.3 Predictive statistical models 43
4.4 Cox PH model fit 46
4.4.1 Cox PH model residuals and assumptions analysis 48
4.5 Significant variables 53

5 Conclusion and Discussion **55**

6 Limitations and Recommendations for future **59**

A Extra results **65**

B R and STATA code **76**

List of Figures

3.1 Survival status. 37
3.2 Distribution of the overall survival of patients in weeks. 39

4.1 Kaplan-Meier survival Estimate. 42
4.2 Kaplan-Meier failure Estimate. 43
4.3 Log-logistic model fit survival curve. 45
4.4 Cox-Snell residuals for the log-logistc model fit. 46
4.5 Cox PH survival curve. 48
4.6 Cox PH Cox-Snell residuals. 49

A.1 Distribution for PB.MONO. 65
A.2 Distribution for BM.BLAST. 66
A.3 Distribution for CD19. 66
A.4 SEX harzard plot. 69
A.5 Schoenfeld residuals (1). 69
A.6 Schoenfeld residuals (2). 70
A.7 Schoenfeld residuals (3). 70
A.8 Schoenfeld residuals (4). 71
A.9 Schoenfeld residuals (5). 71
A.10 Schoenfeld residuals (5). 75

List of Tables

3.1 Clinical Covariates. 31
3.2 Variables with missing values 33
3.3 Variables without missing values 34
3.4 Risk factors frequency 35
3.5 Variable importance for the first six variables. 36
3.6 Gender and age of participants. 37
3.7 Survival status by gender. 38
3.8 Survival status by age. 39
3.9 Number of variables chosen by each model. 40

4.1 Cox PH model fit for all clinical covariates. 47
4.2 Scaled Schoenfeld Residuals of Significant Covariates on the PH. . . 50
4.3 A summary of the results from training and testing the semi-parametric and parametric models. 51
4.4 Significant variables at 5 percent level of significance. 54

A.1 Exponential model fit for all clinical covariates. 68
A.2 Weibull model fit for all clinical covariates. 72
A.3 Lognormal model fit for all clinical covariates. 73
A.4 Loglogistic model fit for all clinical covariates. 74
A.5 A summary of the forecast results from training and testing the semi-parametric and parametric models. 75

Chapter 1

Introduction

Hematologists and oncologists consider a vast number of prognostic factors before recommending a treatment plan to a patient suffering from any kind of blood-related disease or cancer. In most cases, clinical trials determine the treatment plan to be used. If the type of cancer or disease is rare, however, it makes it difficult to conduct large clinical trials. In such instances, the preferred approach is to use multivariate survival models that are built from large observational databases (e.g., the Acute Myeloid Leukemia (AML) dataset). Models derived from large databases can offer better guidance for assessing each patient's different outcomes, as influenced by different factors. In this study, different models, such as the Cox Proportional Hazard (PH), exponential, Weibull, log-logistic, and lognormal models were employed. These models were used to evaluate the effects of covariates on the overall survival times of respective patients.

AML (also known as Acute Myelogenous Leukemia or Acute Myelocytic Leukemia), is one of many types of cancers of the blood and bone marrow; it is, thus, also one of the most rapidly-killing types of diseases (Lowenberg et al., 1999). This type of cancer is acquired or inherited through genetic alteration and is characterised by the multiplication of the number of myeloid cells found in the bone marrow of human beings (Lowenberg et al., 1999). These multiplied myeloid cells, in turn, interfere with the production of normal blood cells, which can affect the process of the bone marrow producing normal blood cells. As a result, the disease ultimately leads to hematopoietic

stem cells insufficiency. The condition of having insufficient hematopoietic stem cells is medically identified as anemia (Lowenberg et al., 1999).

The majority of anemic patients have also been diagnosed with primary AML, which has no risk factors or exposures to account for its development. A few of these patients have, however, been diagnosed with secondary AML, which is known as 'myeloproliferative disease' (Oran et al., 2007). Secondary AML develops in patients with disorders that affect the blood or blood-forming organs, or through other inherited diseases. Patients with secondary AML would, thus, have had a disease known as myelodysplastic syndrome (MDS) for 3 months or more before being diagnosed with AML. As a result, such patients would most likely have been exposed to leukemogenic agents during therapy sessions for various unrelated diseases (Oran et al., 2007).

AML is described as a clonal hematopoietic blood-related disease (Mardis et al., 2009). This disease has mutations in malignancy-associated genes. Due to the nature of the disease, most clinicians study the chromosomes of AML cells to predict whether or not a patient has the disease (Mardis et al., 2009). However, it has come to researchers' attention that AML can easily be mistaken for closely related diseases, such as MDS and/or acute lymphoblastic leukemia (ALL) (Xu et al., 2009) and (Mardis et al., 2009). AML can, still, be distinguished from ALL by committing to the myeloid lineage, which is a method of distinguishing blood cells. Such distinction is conducted through the insightful use of biological methods that distinguish abnormal cells. In comparison, it is more difficult to distinguish AML from MDS (Lowenberg et al., 1999), which requires a more mindful clinical, structural, and genetic analysis.

Statistics published in the United States of America (USA) show that AML is one of the rarest types of all cancers. Studies done in the USA claim that in every year, 24 people in a million are diagnosed with AML, which is less than 1% of the American population (Lowenberg et al., 1999). Data analysis conducted by Gill et al. (2020) reveals that more men have been diagnosed with the disease than women although the average lifetime risk of being diagnosed for both sexes is equal. The annual incidence for adults who are 65 years of age and older is 126 per 1 million adults (Lowenberg et al., 1999). Both the young (45 years and below) and old (65 years and older) can be

diagnosed with AML, but it is more commonly detected in adults. It is very uncommon to diagnose the disease in individuals aged 45 years and below. Studies show that the average age of those diagnosed with the disease is 65 years in the USA (Hassan and Smith, 2014).

The AML Outcome Prediction Challenge survey presented by the M.D. Anderson Cancer Center predicted that by the end of 2014, there would be at least 18860 cases of AML. Both Oran et al. (2007) and Mardis et al. (2009) had conducted studies that further supported a similar prediction. Due to the diverse prognoses of AML patients, however, there were 10460 deaths, which was more than half of the patients that were diagnosed that year. Moreover, it was discovered that of the patients who are diagnosed with AML, less than a quarter survive for more than 5 years (Hassan and Smith, 2014).

It should be noted that the diagnostic methods used in the early 1970s were solely dependent on blood tests and the cytologic/pathological examination of bone marrow (Lowenberg et al., 1999). These diagnosis methods yielded survival rates of no more than 15% (Lowenberg et al., 1999). A survival rate associated with a disease relates to the fraction of people who remain alive from out of the group diagnosed over a particular time span. As time progressed beyond the 1970s, methods of diagnosing AML subtypes and advanced therapeutic approaches were incorporated in dealing with the disease. However, these improved diagnostic methods still did not increase patients with AML's survival rates. Indeed, the survival rates of patients below the age of 65 continue to remain low at 40% (Hassan and Smith, 2014).

The treatment of various types of cancer is influenced by a number of factors, including patient age, general health status, and any possible coexisting conditions. Age is a particularly important factor, as some treatment options may be too harsh for the elderly (Mardis et al., 2009). Treatment for AML has, thus, constituted a combination of different methods, with some failing and leading to patient relapse. Such failure is problematic, as the main objective of any treatment is to increase remission and prevent relapse. In terms of AML, specifically, remission is indicated by the patient reporting less than 5% blast cells in his or her bone marrow, along with the recovery

of peripheral blood counts (Mardis et al., 2009).

AML treatment occurs in two stages. The first stage involves the induction of remission, where the medication daunorubicin is administered to a patient thrice daily. Daunorubicin is given in doses of 40-60 milligrams per square metre of the body-surface area (Rowe and Tallman, 2010). Oftentimes, this drug is administered in conjunction with courses of chemotherapy (Mardis et al., 2009). The second stage is post-induction therapy, which works to prevent relapse after remission. It should be noted that, in general, post-induction therapy tends to be more efficient for younger patients (Mardis et al., 2009).

Another option for treatment is allogeneic bone marrow transplantation, whereby diseased blood stem cells are replaced by healthy ones (Sengsayadeth et al., 2018). The recipient of these healthy blood stem cells can, however, only receive such from a Human Leukocyte Antigen (HLA)-matched donor – the donor may or may not be a relative. An alternative option is the autologous bone marrow transplantation (Sengsayadeth et al., 2018). While this option does not significantly change the risk of relapse for adults, it has been relatively successful in children (Oran et al., 2007). According to Oran et al. (2007) and Heuser et al. (2020), a single high dose of cytarabine is another effective treatment that can enhance the survival rate for patients.

There have also been several clinical prognostic factors noted by Grimwade (2012). The most common and significant ones used to predict complete remission and overall survival have been found to be: 1) a lack of myeloproliferative disease, 2) a young age, 3) a high white blood cell count, and 4) a low degree of adverse cytogenic abnormalities. Myeloproliferative disease (defined as secondary AML) is considered a much worse prognosis than AML, as complete remission is close to impossible (Sengsayadeth et al., 2018). The presence of myeloproliferative disease is, furthermore, an indication of the presence of adverse cytogenics (Thirman and Larson, 1996). Patients with a high degree of adverse cytogenic abnormalities are, thus, considered ‘high risk’ (Hassan and Smith, 2014).

There are many methods/techniques discovered that can assist in diagnosing medical patients. In this study, the main challenge noted was the diagnosis of patients and the identification of variables that will assist in treating AML patients. Survival analysis is best suited in this case because it allows for the analysis of the study of time until the event occurs (Klein and Moeschberger, 2006). Survival analysis is known for having the statistical power to detect the significance of a treatment plan.

Just like most type of modeling, survival models also uses non-parametric, semi-parametric and parametric models. The non-parametric model is known as the Kaplan-Meier (K-M) estimator which estimates the chance of a patient surviving for a specific given time (Guo, 2010). The semi-parametric model, also known as the Cox PH model, is widely used to detect epidemiologic covariates that contribute to the risk a patient is facing. The parametric models consist of accelerated failure time models for example the Weibull and exponential, loglogistic and lognormal models (Guo, 2010). These models are used as an alternative to the semi-parametric model. They are used where the proportional hazard assumption is not held constant (Guo, 2010).

In this study, the dataset being used has a distinct starting time and ending time (Klein and Moeschberger, 2006) and hence, survival analysis will be able to sufficiently analyse the data.

1.1 Statement of the problem

As the previous section highlighted, it is imperative to find an efficient diagnosis method. A study done by Levis (2011) identifies the grieve consequences of misdiagnosing AML in a patient that has any type of cancer especially breast cancer. Previous cancers increase the relative risk for AML patients by 22% (Levis, 2011). AML is a disease that unfortunately generally has a poor prognosis (Valentini et al., 2011). A majority of the patients diagnosed with AML do not survive the disease. The survival rates are lower than 50% for patients aged 65 years and below, which shows that the risk factors are not being identified in time for the most effective treatment to

be administered.

The foregoing section indicates that obtaining an efficient diagnostic method has been a challenge; however, it needs to be done.

1.2 Aim and Objectives

The main aim of this study was to identify the risk factors that affect the survival rates (time to death) of the AML patients using parametric and semi parametric models.

This was achieved by:

- Imputing the missing values using mean/median.
- Fitting the parametric models namely the Weibull, exponential, log-logistic, and lognormal models.
- Fitting the Cox PH model.
- Compare the performance of the parametric and semi parametric models using AIC, AICc, BIC, IBS, Concordance index, MAP, MAD and RMSE.
- Identifying and evaluating the effects of covariates on overall survival.

1.3 Significance of the Study

As mentioned in the closing remark in Section 1.1, finding the best diagnosis for AML is critically important. Determining factors that lead to either death or survival at a later stage for patients afflicted by the disease need, therefore, to be investigated. It is hoped that the model determined by the study could be best used for such predictive purposes, particularly with regard to being used as a supporting tool for predicting the overall survival of a patient diagnosed with AML. Prior knowledge is also important for determining and predicting the overall survival of patients. Therefore, this current study aimed to assist researchers in distinguishing different methods of treatments

needed in particular cases.

The rest of the research report is explained as follows: Chapter 2 presents the literature review, which offers a brief overview of articles where survival analysis was used in various types of cancer studies. Chapter 3 outlines the methodologies used in this study (i.e., the models employed and the statistical techniques conducted). Chapter 4 presents the results obtained from the analysis and the interpretation thereof. Chapter 5, presents the conclusion of the study. The appendices, which appear at the end of this research report, comprise the R and STATA syntax as well as some of the related results obtained.

Chapter 2

Literature Review

2.1 Introduction

This chapter describes survival analysis and provides a literature review relevant to the case under investigation. In addition, the chapter details comparative and related studies conducted by previous researchers. The chapter also reviews the procedures followed, models used, tests done, and conclusions drawn from related research. Furthermore, the review of related literature covers the concepts and theories used in and which frame this current study.

Survival analysis, also known as 'time to event' analysis, is used to estimate the time for an event of interest to occur. Survival analysis is, thus, a statistical technique used to analyse data that experience a failure at a certain point in time. Allison (2010) defines survival analysis as "an assembled set of statistical procedures used to estimate the chance and time in which an event may occur".

There are many examples of end events, depending on the context and field of study. In the medical field, for example, an 'event' could be death, the re-occurrence of a disease, or the failure of a drug being tested for patients in a clinical trial. By comparison, in economics, an event could be related to the duration of a strike or a period of unemployment. Based on the nature of survival analysis, this approach's techniques

have been deemed useful for studying cancers, including AML, since it allows for the measurement of patients' average survival time along with the hazards that they face.

Of further note, the response for survival analysis is often referred to as survival time, event time, or failure time. The random variable for survival analysis, namely time, is strictly positive. In the case of this current study, the response was related to overall survival time, measured in weeks. However, survival time may also be measured in months or years, depending on the end of a given study or event occurrence.

As mentioned in the previous chapter, this current study aimed to identify the influence of different factors on the time of the study's determined event. In investigating the dependence of survival time on one or more predictor variables, most researchers in the medical field tend to prefer using the Cox PH model, which is a semi-parametric model, as opposed to a parametric model (Pourhoseingholi et al., 2007), (Schlichting et al., 1983) and Sephton et al. (2000). This preference is based on how the Cox PH model is considered to be more appropriate for survival analysis as it reduces the amount of assumptions that need to be made.

An obstacle normally encountered by researchers when analysing survival data is, however, the likelihood that some of the individuals involved in a study may not be fully observed until the 'end-time' of the experiment (Schober and Vetter, 2018). In such cases, researchers are not able to obtain the accurate, full, or complete recovery time data, or data pertaining to the time until the death of the relevant individuals (Schober and Vetter, 2018). Rather, researchers only have on record some of the time before the occurrence of the event. Such incomplete observation of the survival time is known as 'censoring'. Allison (2010), thus, emphasises the importance of survival analysis, since censoring and time-dependent covariates are not easy to use when employed in conventional statistics. In survival analysis, however, since time to event is strictly positive and has a skewed distribution (Klein and Moeschberger, 2006), censoring has to be done.

When censoring is not employed, issues may be encountered during data analysis. There are various types of censoring that can occur, including right censoring, which

can be classified as either Type I or Type II; left censoring; and interval censoring. Regardless of the type of censoring, an assumption is enforced that it should be non-informative about the event (Klein and Moeschberger, 2006). As such, usual analysis techniques cannot be used. In alignment with this understanding, and as previously mentioned, the techniques explored in this current study were the Cox PH model as well as various parametric models, namely the Weibull, exponential, lognormal, and log-logistic models. Parametric models provide complementary statistics for clinicians and researchers about how risks differ over time.

The next section details some of the more relevant results obtained by previous researchers. In addition, a comparison of the procedures used in each reviewed study in relation to this current study is provided.

2.2 Related Studies

The following previously conducted studies hold similarities to the one described in this research report:

Pourhoseingholi et al. (2007) conducted a study on the survival of patients with gastric carcinoma. The Pourhoseingholi et al. (2007) study shares the same main objective as this current study in that it compares survival regression methods. Specifically, (Pourhoseingholi et al., 2007) used a dataset comprising 746 individuals, with five selected prognostic covariables, and compared the efficiency of the selected models using Akaike's information criterion (AIC). This gastric carcinoma study shows that patients diagnosed with AML aged over 45 years were at a much greater risk of death and other complications than other patients (Pourhoseingholi et al., 2007).

The study conducted by Pourhoseingholi et al. (2007) found a correlation between age and the degree of risk severity in gastric carcinoma, which aligns with similar findings established in this current study – namely that patients diagnosed with AML aged 65 years or older tend to be at a very high risk of death or relapse. Pourhoseingholi et al. (2007) concluded that both the Cox and exponential models in the multivariate

analysis presented with similar results. In their univariate analysis, the authors found that lognormal regression was strongly supported by the data; moreover, its results were more precise (Pourhoseingholi et al., 2007).

Clinical trials on 488 patients with chronic liver disease were performed in 1983 by Schlichting et al. (1983). These trials included a placebo treatment, and the survival dataset collected was analysed using the Cox PH model. All 51 variables were used to build a model. However, the variables were later reduced using a stepwise procedure so as to build a final model with only 12 variables. The stepwise procedure was used for the same reasons as it was used in this current study to select variables that had a significant prognostic effect and could, thus, be used as effective treatment indicators (Schlichting et al., 1983).

Both Pourhoseingholi et al. (2007) and Schlichting et al. (1983) found age to be highly significant in their disease studies, which aligns to findings regarding AML. In the chronic liver disease study conducted by Schlichting et al. (1983), the aim was to develop a prognostic index based on the final model, which consisted of significant prognostic effect variables. The index that was ultimately developed, however, allowed only for the calculation of survival probabilities for the period of 5 years. Cross-validation methods were, thus, used to test the usefulness of the prognostic index. The results obtained showed that the difference between the estimated function and the observed survival function was statistically insignificant (Schlichting et al., 1983).

Both Alamartine et al. (1991) and Sephton et al. (2000) studied the multivariate Cox regression model as well as standard univariate models (e.g., the lognormal, exponential, and Weibull models). These sets of authors specifically compared the semi-parametric and parametric models by clarifying risk factors for patients. In the study conducted by Alamartine et al. (1991), the risk was that patients might experience chronic renal failure (CRF) due to severe kidney inflammation, while the Sephton et al. (2000) study considered risk in terms of patients with metastatic breast cancer experiencing diurnal variation of salivary cortisol.

For the study of patients with IgA glomerulonephritis (IgA-GN), the univariate comparison between those with CRF and those without indicated multiple risk factors, whereas the multivariate study indicated only four risk factors. The noted four risk factors were the only ones that had a statistically significant effect on patient survival rates (Alamartine et al., 1991). From the results obtained using the data on IgA-GN, the multivariate study was deemed more useful in finding risk factors for CRF.

For the study conducted on patients with metastatic breast cancer by (Sephton et al., 2000), patients' salivary cortisol levels were assessed four times daily for 3 days. The study included 104 patients. The Cox PH model was used as the basis of the analysis. Patients with low survival rates were identified as having abnormal diurnal variations within the levels. It was concluded, therefore, that patients with the type of breast cancer involving low or abnormal diurnal cortisol rhythms experienced an early failure rate (i.e., such patients were pronounced dead earlier than those with normal diurnal variations) (Sephton et al., 2000).

A study conducted by Ravangard et al. (2011) was found to also have the same aim as this current study, namely to compare the semi-parametric Cox PH model to parametric models whilst also determining the factors that affect patients' length of stay (i.e., the survival variable). The Ravangard et al. (2011) study involved 3,421 cases from across different hospital units. The authors used only eight variables to fit the models, with the AIC and Cox-Snell residuals being used to compare the fit of the models. A probability-value of 5% was used as an assisting tool to measure the significance of the models. The graphs generated by the Cox-Snell and the AIC values showed that the gamma model projected the factors affecting length of stay most accurately. This finding, thus, highlighted that the gamma model had a better fit to the data than either the other parametric models or the Cox PH (Ravangard et al., 2011).

Wang et al. (2010) did a survival analysis study in oncology, with special reference to gallbladder cancer. The main objective of the study was to compare different types of accelerated failure time parametric survival techniques. These techniques were used to model the benefit of adjuvant chemotherapy. The semi-parametric model was found to not be of good use when the proportionality assumption is not met. Indeed, in

the case of the Wang et al. (2010) study, the assumption was not met; hence, the semi-parametric model was not used. The researchers further compared the parametric models using AIC, which was similarly employed as a method of comparison in the current study as well. Wang et al. (2010) found the lognormal survival model to be most favorable, as it was able to predict which patients would benefit most from chemotherapy.

Xihui Lin and Hunter (2014) performed a similar study as this research using the same dataset. The aim of their study was to predict overall survival time in AML using a bagged semi-parametric model. They used the benchmark variables as instructed by the M.D. Anderson Cancer Center. The benchmark variables included age at diagnosis, hemoglobin levels, cytogenic category, specific anthra based treatment administered and lastly albumin levels. Xihui Lin and Hunter (2014) found that adding more variables to the clinical model did not improve the performance of the model. Addition of new variables reduced the performance of the clinical model. The study by Xihui Lin and Hunter (2014) concluded that the protein information did not provide additional predictability to the clinical model. The 5 clinical benchmark variables gave sufficient information to predict the overall survival time for the AML patients (Xihui Lin and Hunter, 2014). This research will use parametric survival models to include other variables that will help improve the predictability of the clinical model.

In all, the majority of articles presented have used the Cox PH model as the basis for their survival analyses. Hence, Cox PH model is the most commonly used model in medical research for survival analysis.

Chapter 3

Methodology

3.1 Introduction

In this chapter, the models mentioned in Section 1.2, as well as the methods used in the analysis of the survival data gleaned in this study, are presented. This study employed the same terminology as that used by Klein and Moeschberger (2006):

- The i^{th} survival time is denoted by T_i;
- The fixed right censoring time is denoted by C_i; and
- Observations are assumed to be independent and identically distributed.

Let $\delta_i = I(T_i \leq C_i)$ be the event indicator where if $T_i \leq C_i$ then $\delta_i = 1$, otherwise $\delta_i = 0$.

3.1.1 Survival function

The survival function $S(t)$ is an essential function in survival analysis. This function is defined as the probability that a patient or subject will survive past any specified time t (Klein and Moeschberger, 2006).

The expression for the survival function is given as:

$$S(t) = P(T > t) = 1 - F(t) = \int_t^{\infty} f(x)dx. \tag{3.1}$$

The density function is given as $f(t) = -\frac{\partial S(t)}{\partial t}$. Furthermore, the survival function is a function with domain $\mathbb{R}^+$ and counter-domain $[0, 1]$, which satisfies the following three properties:

- It is a non-increasing function i.e $S(t) \to 0$, as $t \to \infty$; and
- $S(0) = 1$ at $t = 0$.

3.1.2 Hazard function

The hazard function is the instantaneous rate at which an event occurs, and is presented as:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}, \tag{3.2}$$

where $\Delta t > 0$ represents a small time interval. The hazard function can also be expressed as $h(t) = -\frac{\partial \log(S(t))}{\partial t}$, and the survival function as:

$$S(t) = \exp(-h(t)). \tag{3.3}$$

Hence, if one of the two is known, then the other can be calculated:

$$S(t) = \exp\left[-\int_0^t h(u)du \right] = \exp(-H(t)), t \geq 0. \tag{3.4}$$

3.2 Survival models

There are many ways of estimating $S(t)$, using either non-parametric estimators (e.g., the K-M estimator), or by making some parametric assumptions (Miller Jr, 2011). By making use of the assumption that every subject follows the same survival function, it is possible to simplify the estimation of the survival function.

3.2.1 Non-parametric model: Kaplan-Meier estimator

The K-M estimate is classified by Goel et al. (2010) as a better option to use when measuring the fraction of myeloid cells within a patient after remission for a certain period of time t_i. Furthermore, the K-M estimator can be categorised as a non-parametric method because it makes no assumptions about the shape of the underlying survival curve. The K-M curve is, thus, equivalent to the empirical distribution function when no censoring is observed.

An advantage of the K-M method is that it includes all available information about a researcher's observations. All uncensored and censored observations that are included cause a non-parametric analysis to generate wider confidence intervals, compared to those generated by parametric analyses (Goel et al., 2010). The K-M method is, thus, the simplest method of computing the survival function until time t_i. As a result, this method is most often used to estimate an individual's probability of surviving for a given period of time. The K-M method can also take into account different types of censored data, particularly right censoring, where the patient does not complete the experiment (Goel et al., 2010).

The survival function, as presented in the K-M method, can be expressed as a product of the probabilities of patients surviving a disease in k or more periods. The function is mathematically represented as:

$$S_k = \prod_{i=1}^{k} p_i, \tag{3.5}$$

where $p_i = \frac{r_i - d_i}{r_i}$, with r_i representing the number of surviving patients at the start of each period and d_i the number of deaths observed in each period.

In addition, the K-M estimator is mathematically represented as:

$$\widehat{S}_t = \prod_{i=1}^{k} \left(\frac{n_i - d_i}{n_i} \right), \tag{3.6}$$

where n_i is the number of subjects facing the risk of death by a disease (e.g., AML) just before time t_i and d_i is the number of deaths observed at time t_i (Kaplan and Meier,

1958).

It is important to consider the number of censored subjects and their distributions when evaluating the K-M curve, as the number of participating patients is seen in the number of steps that the curve has. For example, if there are many participating patients, the curve will have many small steps. Conversely, the curve will have large steps if there is a limited number of participating patients.

3.2.2 Semi-parametric model: Cox PH model

The Cox PH model is a semi-parametric model that was developed by Cox (1972). This model is a multivariate method that is widely used in survival analysis in order to analyse the effect of several risk factors relating to survival. The Cox PH model is most generally used to detect clinical variables that contribute to risk. Non-parametric methods, however, do not allow for easy control of covariates, which leads to the requirement of categorical predictors. In this current study, the Cox PH model was used to examine the relationship between the survival function for each patient and exploratory variables.

It should be noted that there are a few assumptions to be considered in order to use the model well. The first indicates the independence of the individual survival times of the patients in a study. The second indicates that there must be a multiplicative relationship between the hazard and the predictor(s). The final indicates that the hazard ratio should be constant over time. The Cox PH has been explicitly designed to estimate the hazard ratio that estimates the hazard rate, which rate represents the probability of an event occurring (be it death or any other relevant event).

Furthermore, the exploratory variables considered should affect the patient prognosis. Hence, analysis using the Cox PH is conducted in order to examine the effect of specified covariates in causing an event such as death to occur in a certain period. The Cox PH function is given as:

$$h(t) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p) = h_0(t)\exp\left(\boldsymbol{\beta}'\mathbf{X}\right). \tag{3.7}$$

If the natural logarithm is taken on both sides of the equation, the Cox PH can be given as:

$$\log(h(t)) = h_0(t) + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p, \tag{3.8}$$

where $\mathbf{X}$ is a vector of predictor variables $x_1, x_2, x_3, ..., x_p$ and $h_0(t)$ is the baseline hazard. The coefficients, in turn, measure the effect size of the covariates.

In the case where the predictor variables x_i are zero then $h(t) = h_0(t)$. The representation of the survival function under the Cox PH model is then given as:

$$S(t) = S_0(t) \exp(\boldsymbol{\beta}'\mathbf{X}), \tag{3.9}$$

where $S_0(t)$ is the baseline survival function.

In the current study, the Cox PH model was fitted in R in order to investigate the effect of the chosen risk factors as the most important prognostic features that influence the survival time of AML patients according to the random forests model. In order to best fit the Cox PH model for this study, it was necessary to ensure that all variables were first converted to numeric form so that the model would fit using the command $coxph()$.

3.2.3 Parametric models

Parametric survival analysis models typically require a non-negative distribution, such as those mentioned in Section 1.2 (i.e., the lognormal, log-logistic, exponential, and Weibull models). A non-negative distribution is defined for non-negative real numbers, and the choice of distribution affects the shape of the model's hazard function, as previously explained. The alternative method that requires no distributional assumptions to be made estimates the hazard function from the data.

Miller Jr (2011) states that any probability distribution that satisfies $T \in [0, \infty)$ can be used in a survival model. A parametric survival model is, thus, based on its survival time following a specified probability distribution. In this current study, the following were considered, the:

- Exponential model.
- Weibull model.
- Log-normal model.
- Log-logistic model.

These parametric models were compared to the Cox PH model in this study, and are further detailed in the following subsections.

The exponential model

The exponential model is considered to be one of the most vital models in survival analysis. Specifically, this model is the simplest form of distribution in lifetime studies. If the time to event T follows an exponential distribution with parameter $\lambda > 0$, then the probability density function (pdf) will be given as:

$$f(t) = \lambda \exp(-\lambda t). \tag{3.10}$$

The survival function, in turn, is given as:

$$S(t) = \exp(-\lambda t), \tag{3.11}$$

and the corresponding hazard function as:

$$h(t) = \frac{f(t)}{S(t)} = \lambda. \tag{3.12}$$

The hazard function for the exponential distribution is constant, while the exponential distribution possesses a 'memoryless' property (i.e., the failure rate remains constant and does not change with time).

The Weibull model

The Weibull model is also known as a generalised exponential model. Amongst the parametric models used in this current study, only the Weibull distribution can be presented as both a proportional hazard model and as an accelerated failure time model.

The distribution model is further noted as being vital in medical sciences, as it is helpful for analysing the uniform increase or decrease observed in patient mortality rates.

If the time to failure follows a two-parameter Weibull distribution, with $\gamma > 0$ as its shape parameter and $\lambda > 0$ as its scale parameter, then the pdf will be given as

$$f(t) = \gamma\lambda t^{\gamma-1} \exp\left(-\lambda t^{\gamma}\right). \tag{3.13}$$

The survival function, in turn, is given as

$$S(t) = \exp(-\lambda t^{\gamma}) \tag{3.14}$$

and a corresponding hazard function as:

$$h(t) = \gamma\lambda t^{\gamma-1}. \tag{3.15}$$

The hazard function for the Weibull distribution is a monotonic increasing function of t for $\gamma > 1$. When $\gamma < 1$, it is decreasing and when $\gamma = 1$, the Weibull model has a constant hazard function (i.e., it becomes an exponential distribution). Therefore, the exponential distribution is a special case of the Weibull distribution.

The lognormal model

If a random variable T follows a lognormal distribution, then the $\log$ of T follows a normal distribution,

$$Y = \log(T) = \alpha + \sigma Z,$$

where Z has a standard normal distribution. The lognormal is a two-parameter distribution where the mean is denoted as μ and the standard deviation (sd) as $\sigma > 0$.

Its pdf is given by

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log t - \mu)^2}{2\sigma^2}\right) = \phi\left(\frac{\log t - \mu}{\sigma}\right)/t, \tag{3.16}$$

where $\phi(\frac{\log t - \mu}{\sigma})$ is the density function of the standard normal variable.

The survival and hazard functions of the lognormal are given respectively as:

$$S(t) = 1 - \phi\left(\frac{\log t - \mu}{\sigma}\right) \tag{3.17}$$

and

$$h(t) = \frac{\phi(\frac{\log t}{\sigma})}{\sigma t[1 - \phi(\frac{\log t}{\sigma})]}, \tag{3.18}$$

where $\phi(\frac{\log t}{\sigma})$ is the cumulative distribution function of the standard normal.

The hazard function for the lognormal distribution increases from 0 to a maximum value and then it declines monotonically to reach zero as $T \to \infty$.

The log-logistic model

The log-logistic distribution is a continuous probability distribution for a non-negative random variable. This model's pdf is given as:

$$f(t) = \frac{(\frac{\nu}{\alpha})(\frac{x}{\alpha})^{\nu-1}}{(1 + (\frac{x}{\alpha})^{-\nu})}, \tag{3.19}$$

where ν is the shape parameter and α is the scale parameter. It should be noted that the log-logistic distribution, unlike other distributions, has a nonmonotonic hazard function when the shape parameter is greater than 1.

The survival function, in turn, is given as:

$$S(t) = \frac{1}{1 + (\frac{t}{\alpha})^{\nu}}, \tag{3.20}$$

and the hazard function as:

$$h(t) = \frac{(\frac{\nu}{a})(\frac{t}{\alpha})^{\nu-1}}{1 + (\frac{t}{\alpha})^{\nu}}. \tag{3.21}$$

3.3 Parameter estimation

3.3.1 Censoring

As mentioned previously, the lifetimes of all subjects may not be completely observed due to censoring. Right censoring occurs when exact lifetimes are known for only

a portion of individuals under investigation, and the other censoring forms are only known to exceed certain values. While there are various types of censoring, this current study only considered right censoring, as it is the most common form of censoring encountered in survival analysis and in the AML dataset. The fixed censoring time is denoted by C_i and the survival time by T_i for the i^{th} individual. The observed response will be denoted as:

$$X_i = \min(T_i, C_i). \tag{3.22}$$

3.3.2 Partial likelihood

An advantage of the Cox PH model is that partial likelihood estimates the parameters, which eliminates the need to estimate the baseline hazard. While partial-likelihood estimates are not as efficient as maximum-likelihood estimates for a correctly specified parametric hazard regression model, after estimating a Cox PH model, it is possible to recover a non-parametric estimate of the baseline hazard function.

In such a case, it is necessary to let the probability density for an event that occurred at time t be represented by $p(t|\mathbf{X}, \boldsymbol{\beta})$. Hence, if a patient i experiences an event at time t_i they will contribute $p(t|X_i, \beta)$ to the likelihood. By comparison, if a patient is censored, they will contribute $S(t|X_i, \beta)$ to the likelihood, where $S(t)$ is the probability of survival at time t. The likelihood function is given as:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} [f(t|X_i, \beta)]^{\delta_i} [S(t|X_i, \beta)]^{1-\delta_i}.$$

It should be noted that the partial likelihood is determined at each failure time, and is expressed as the product of likelihoods:

$$L(\boldsymbol{\beta};\ \mathbf{X}) = \prod_{i=1}^{n} \left(\frac{\exp\left(\boldsymbol{\beta}'\ \mathbf{X}\right)}{\sum_{l \in R(X_i)}^{\max} \exp\left(\boldsymbol{\beta}'\ \mathbf{X}\right)} \right) \tag{3.23}$$

3.3.3 Maximum likelihood estimate

The Maximum Likelihood Estimation (MLE) method is commonly used on parametric models. If a patient relapses before the expected time t_i, the i^{th} patient contributes

$S(t_i)$ to the likelihood function, where the likelihood is given as:

$$L(\boldsymbol{\beta};\ \mathbf{X}) = S(t_i). \tag{3.24}$$

A censored patient, whose event of relapse did not occur before the expected time t_i contributes just $S(t_i)$ to the likelihood function. Such contribution is then interpreted as the probability of the failure occurring:

$$L(\boldsymbol{\beta};\ \mathbf{X}) = S(t_i)h(t_i).$$

The joint likelihood function is, in turn, given as:

$$L(\boldsymbol{\beta};\ \mathbf{X}) \propto \prod_{i=1}^{n} f(X_i;\beta_i)^{\delta_i} S(X_i;\beta_i)^{1-\delta_i}. \tag{3.25}$$

3.4 Model development

A model has to be developed in order to summarise and analyse data efficiently. An efficient model helps identify and reduce the number of covariates, and accomplishes a high level of explanation and prediction with the least number of predictor variables.

3.4.1 Variable selection

Variable selection is classified as an essential procedure for both data analysis and the model building process. The main aim for a variable selection procedure is to end up with a subset of covariates that best explain the data (Kazemitabar et al., 2017). Considering the ratio of the number of variables and cases in the AML dataset, it is best to make use of a non-parametric variable selection method. In particular, using a non-parametric variable selection method that is unrelated to the models compared in the study helps to reduce bias.

Decision Trees

Decision trees are well known for predicting the performance of a model as well as for providing important variable information. The decision trees algorithm ranks

predictors according to importance scores, with the high-ranking predictors being chosen to model the data. This procedure prunes the variables; thereby allowing for a more parsimonious model to be constructed (Kazemitabar et al., 2017).

Random Forests

Another well-used method of variable selection is random forests, which are made up of a multitude of decision trees. In essence, random forests are a machine learning classification algorithm with similar hyper-parameters as those of decision trees (Kazemitabar et al., 2017). The benefit of using random forests over decision trees is, however, that they provide higher levels of accuracy, flexibility, and ease of output interpretation (Ali et al., 2012). Han et al. (2016) and Ali et al. (2012) further maintain that random forests improve the performance of decision trees and reduce the possibility of over-fitting.

The packages **randomForest** and **ranger** are often used to execute the variable selection algorithm in R. The algorithm uses the following control parameters to find the best model:

1. ntree - number of trees in the forest;
2. importance - set to TRUE, allows for the importance of variables to be assessed;
3. mtry - number of predictors to be randomly drawn as a sample and choose best split; and
4. maxnodes - number of terminal nodes in the forest.

Random forests also provide extra useful information, such as variable importance and proximity measures (Ali et al., 2012). Variable importance offers a measure of the importance of each variable in the building of a model (Genuer et al., 2010). The process of choosing the covariates is, thus, simplified by the variable importance plot. The variables are then ordered from most important to least important variable on the plot. The random forests algorithm also provides an importance table, which gives a summary of the Mean Decrease Accuracy (MDA) and Mean Decrease Gini (MDG)

(Han et al., 2016).

The MDA represents the mean decrease in the accuracy of the model, as the algorithm classifies the variables whilst excluding the variable. Hence, the more inaccurate the model is, the higher the importance of the excluded variable. The MDG coefficient is measured on a scale from 0 to 1, and measures the homogeneity of node, which depends on the contribution of a certain variable (Han et al., 2016). An MDA graph is also used to choose the variables by selecting the first few variables separated by a jump between them. It is necessary, however, to be careful and not select too few variables, as this could mean that there is insufficient information to create a significant model. There is also the problem of choosing too many variables, which would mean that there is more information than the researcher needs (Genuer et al., 2010).

Stepwise Regression

Stepwise regression is a method that iteratively assesses the statistical significance of each independent variable in a linear regression model. The technique is a combination of both the forward and backward selection techniques, and is, thus, a modified form of the forward selection technique (Olusegun et al., 2015). In addition, stepwise regression assesses each step in which a variable is added to a model and then determines whether or not the variable's significance has been reduced below the preset significance level. Any variable found to be insignificant is removed from the model.

It should be noted that stepwise regression makes use of two significance levels. The first significance level is used for adding variables and the second is used for removing insignificant variables from the model. The cutoff probability used to add variables to the model must also be less than that used to add the variables to the given model (Olusegun et al., 2015). In so doing, the difference between the cutoff probabilities can help the procedure to avoid getting into an infinite loop.

For this current study, the best approach to the creation of a parsimonious model was deemed to be the combination of random forests and the stepwise p-value approach. The variables were kept in the model when its associated significance level was less

than a specified p-value. In contrast, if the associated significance level was greater than the stipulated p-value, then the variable was removed from the model.

3.4.2 Information based criteria

The parametric models and semi-parametric model used in this study were compared using both the Bayesian information criterion (BIC) and the AIC. These two criteria have been described as "penalized-likelihood" criteria by (Vrieze, 2012). Penalised likelihood estimation is a method used to take into account model complexity during the estimation of parameters from different models. The two information based criteria are generally used to compare non-nested models, with each of the criterion attempting to resolve overfitting by making use of a penalty term. A penalty term accounts for the number of parameters used in the model(s) (Kuha, 2004).

The AIC is defined as a quality estimator of statistical models, and it estimates the relative amount of information that a model loses. The best model always loses the least amount of information (Burnham and Anderson, 2004). The AIC is given as:

$$\text{AIC} = -2\log(L) + 2k, \tag{3.26}$$

where L is the maximised value of the likelihood function of the model, and the number of parameters estimated in the model being evaluated is represented by k.

The BIC, in turn, was formulated by Schwarz (1978), which is why it is also known as the Schwarz information criterion. The BIC is well known for penalising model complexity more severely when compared to the AIC, and is given as:

$$\text{BIC} = -2\log(L) + k\log(n), \tag{3.27}$$

where n is the sample size. The BIC of any model is defined as a true estimation of the function of posterior probability, and its value represents how true the model is (Schwarz, 1978). As a result, low values are preferred for this particular criterion. The criterion is also only valid when the sample size is greater than the number of parameters in the model (Burnham and Anderson, 2004) which is true for the AML dataset.

3.4.3 Model performance evaluation techniques

The model's overall performance were tested using the Integrated Brier Score and the concordance index (c-index) techniques as well. According to Graf et al. (1999) IBS was initially developed for the purposes of predicting inaccuracies encountered when forecasting weather. Graf et al. (1999) later extended it to measure the performance of survival models. IBS was developed in such a way that it can provide an overall value that quantifies the performance of a model at times $t_1 \leq t \leq t_{\max}$ (Graf et al., 1999).

IBS for a survival model over the interval $[0; \max(t_i)]$ is given as:

$$\text{IBS} = \frac{1}{\max(t_i)} \int_0^{\max(t_i)} BS(t)dt,$$

where $BS(t)$ is the Brier score of the model at time t. Perfect accurate predictions will give a Brier score of 0 and perfect inaccurate predictions will have a Brier score of 1 (Graf et al., 1999).

The c-index is widely used to assess the predictions made by a survival algorithm (Steck et al., 2008). The index represents the area under the ROC curve which also includes the censored data (Brentnall and Cuzick, 2018). In this study the c-index was used to evaluate the predictive abilities of both the parametric and semi-parametric models which allowed for better comparison of the models.

The c-index assisted by ranking the survival times based on the individual risk scores. The index is calculated as follows:

$$\text{C-index} = \frac{\sum_{i,j} 1_{T_j<T_i} \cdot 1_{\eta_j>\eta_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j<T_j} \cdot \delta_j},$$

where

- η_i represents the risk score of i;
- $1_{T_j<T_i}$ = 1 if $T_j < T_i$ else 0; and
- $1_{\eta_j<\eta_i}$ = 1 if $\eta_j < \eta_i$ else 0.

A good model prediction would give a c-index value of 1 and a random prediction would give a c-index of 0.5 (Brentnall and Cuzick, 2018).

3.4.4 Forecast measures

The main purpose of using forecast measures is to examine and evaluate the difference between the actual values and the predicted values (Chase Jr, 1995). They are used to define the average model accuracy and performance (Willmott and Matsuura, 2005). The forecast measures are used as follows:

- Mean absolute deviation (MAD): MAD is the avaerage distance between each data point and the mean of the data. It is given as:

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|,$$

  where n is the number of data values, $\bar{x}$ is the average of the dataset and x_i is the data values in the data set.

- Mean absolute error (MAE): MAE gives the average absolute error value expected by obtaining average of the differences between the predicted values and the actual values. It is given by:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|e_i|,$$

  where n is the number of errors and e_i are the error values for each data value in the data set.

- Mean absolute percentage error (MAPE): MAPE measures the accuracy of the forecast system used. It gives the forescast measure as a percentage.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{a_i - f_i}{a_i}\right|,$$

  where a_i is the actual value obtained and f_i is the forecasted value.

- Root mean square error (RMSE): RMSE gives the standard deviation of the prediction errors. It is given by:

$$\text{RMSE} = \left[\frac{1}{n}\sum_{i=1}^{n}|e_i|^2\right]^{\frac{1}{2}}.$$

3.4.5 Model diagnostics

Many assumptions are made when developing models; however, measures to investigate the validity of these assumptions must also be taken. In this current study, residuals were used to verify the validity or overall fit of the parametric and semi-parametric models that were evaluated. In particular, Cox-Snell residuals, which are defined as a type of standardised residuals, were used (Nardi and Schemper, 2003).

Cox-Snell residuals are widely used for lifetime models. When these residuals are plotted on an exponential probability plot, they allow easy assessment of the following:

- Extreme data points that require additional attention;
- The appropriateness of the failure time distribution; and
- The appropriateness of the relationship between the failure time and the risk factors.

In this current study, the Cox-Snell residuals are given in two forms for the Cox PH model and the parametric models, respectively as follows:

$$r_i = \hat{H}_0(t) \exp\left(\hat{\boldsymbol{\beta}}'\mathbf{X}\right)$$

and

$$r_i = -\log \hat{S}(t),$$

where $\hat{H}_0$ represents the fitted Cox PH model's estimated value of the baseline cumulative hazard rate, and the estimate of the hazard rate is given by $\hat{S}(t)$. The plot of the cumulative hazard rate versus the residuals helps to give a clear indication of whether the models fit well. Specifically, the plot of a good fit should have a straight line passing through the origin with a slope of 1 (Nardi and Schemper, 2003).

Furthermore, the martingale residual is another useful diagnosis method. The primary use of martingale residuals is to determine the functional forms of covariates in a model. In other words, the residuals assess nonlinearity (Therneau et al., 1990). Since the martingale residuals take any value in the range $-\infty < r_j < +1$ values closest to 1 in this current study represent patients who experienced relapse 'too early' (i.e.,

a failure time was observed before the censored time t). Negative martingale residual values, in turn, show that no failures were observed before censored time t (Therneau et al., 1990).

The martingale residuals are given as:

$$r_j = \delta - \hat{H}_0(t) \exp(\boldsymbol{\beta}' \mathbf{X}).$$

The study makes use of the Schoenfeld residuals to test the proportionality assumptions of the Cox PH model. The Schoenfeld residuals represent the difference between the observed covariate and the expected covariate X_i given the risk set R_i at that time t_i (Xue et al., 2013). The residuals are calculated for each individual covariate to check if the covariates individually satisfy the assumptions of the Cox model. Since the residuals are calculated for each covariate both the number of predictor variables in the model and Schoenfeld residual variables are same (Xue et al., 2013). The residuals are not defined for censored individuals.

The Schoenfeld residuals for an individual predictor variable are given as:

$$r_{ik} = X_{ik} - E(X_{ik}|R_i).$$

3.5 Data source

The study described in this research report used the patient dataset provided by the M.D. Anderson Cancer Center. The data were collected over a period of approximately 11 years. This AML dataset was then presented to the public as a challenge, and it is currently freely available at: https://www.synapse.org/#!Synapse:syn2488690 website.

The data captured 271 measurements for each patient, and included 191 patients who were diagnosed and treated at the M.D. Anderson Cancer Center. These patients were treated using cytosine arabinoside (ara-C)-based chemotherapy. The 271 measurements for each patient comprised 40 clinical correlates and 231 protein and phosphoprotein levels. The clinical correlates describe patient demographics,

cytogenics, and mutation statuses. These correlates also include the results of several standard blood tests administered during the course of the treatment.

There are many data imputation techniques, such as partial imputation, mean or median imputation, regression imputation, and hot and cold imputation. Since the percentage of missingness of the dataset available on the internet is low at 0.895288%, and the missingness was found only in numeric variables, it was appropriate for this study to resort to the use of the mean/median imputation technique.

Among the clinical covariates within the collected data, there were variables that were deemed clinically important by many prior researchers, including the M.D. Anderson Cancer Center, to include in the models to be built in this study. Such variables included: 'age.at.dx', 'SEX', 'HGB', 'ALBUMIN', 'Chemo.Simplest', 'Infection' and 'cyto.cat'. These variables were, thus, included in all the included models in this study, whether they are were found to be significant or not by the variable selection methods. The clinical covariates are as follows:

Table 3.1: Clinical Covariates.

Clinical Covariate	Values	Desciption
SEX	M,F	Patient gender
AGE	numeric	Patient age at the time of diagnosis
AHD	numeric	Prior antecedent hematologic disorder
PRIOR.CANCER	YES,NO	Whether the patient has been diagnosed with a prior cancer.
PRIOR.CHEMO	YES,NO	Whether the patient has had prior chemotherapy
PRIOR.RAD	YES,NO	Whether the patient has had prior radiation therapy
INFECTION	YES,NO	Whether the patient was diagnosed with an infection
CYTO.CAT	"-5","-5,-7","-5,-7, +8","-7,+8","11q23", "21","8","diploid", "IM","inv9","Misc", "t6;9","t8;21","t9;22"	The cytogenic category of the patient
ITD	NEG , POS , ND	Whether the patient was found to have a ITD FLT3 mutation
D385	NEG , POS , ND	Whether the patient was found to have a D835 FLT3 mutatuion
RAS.STAT	NEG , POS , NotDone	Whether the patient was found to have a Ras Stat mutation
CHEMO.SIMPLEST	Anthra-HDAC,Anthra-Plus, Flu-HDAC,HDAC-Plus non Anthra,StdAraC-Plus	The specific Anthra based treatment administered
RESP.SIMPLE	CR , RESISTANT	Patients were categorized as having a complete response or to be restistant to treatment
RELAPSE	Yes , No , NA	Whether a patient with complete response later relapsed
VITAL.STATUS	A , D	The final outcome of each patient at the end of study, either alive or deceased
OVERALL.SURVIVAL	numeric	Patient's overall survival time measured in weeks from diagnosis to exiting the study
REMISSION.DURATION	numeric or NA	The duration of time spent in the remission measured in weeks
WBC	numeric	The white blood cell count

Table 3.1 continued from previous page

Clinical Covariate	Values	Desciption
ABS.BLST	numeric	The total number of myeloid blast cells measured in blood samples
BM.BLAST	numeric	The number of myeloid blast cells measured in bone marrow samples
BM.MONO	numeric or NA	The number of monocytes measured in bone marrow samples
BM.PROM	numeric or NA	The number of promegakaryocytes measured in bone marrow samples
PB.BLAST	numeric or NA	The number of myeloid blast cells measured in blood samples
PB.MONOCYTES	numeric or NA	The number of monocytes measured in blood samples
PB.PROM	numeric or NA	The number of promegakaryocytes measured in blood samples
HGB	numeric or NA	Hemoglobin count measured in blood samples
PLT	numeric or NA	Platelet count measured in blood samples
LDH	numeric or NA	Lactate dehydrogenase levels measured in blood samples
ALBUMIN	numeric	Albumin levels measured in blood samples
BILIRUBIN	numeric or NA	Bilirubin levels measured in blood samples
CREATININE	numeric	Creatinine levels measured in blood samples
FIBRINOGEN	numeric or NA	Fibrinogen levels measured in blood samples
CD13	numeric or NA	Levels of cell surface marker CD13 detected
CD33	numeric or NA	Levels of cell surface marker CD33 detected
CD34	numeric or NA	Levels of cell surface marker CD34 detected
CD7	numeric or NA	Levels of cell surface marker CD7 detected
CD10	numeric or NA	Levels of cell surface marker CD10 detected
CD20	numeric or NA	Levels of cell surface marker CD20 detected
HLA.DR	numeric or NA	Levels of cell surface marker human leukocyte antigen detected
CD19	numeric or NA	Levels of cell surface marker CD19 detected

3.5.1 Risk factors for Acute Myeloid Leukemia with missing values

The variable reduction methods used in this study (i.e., random forests and stepwise regression) reduced the missing values percentage significantly. Table 3.2 provides summary statistics of the variables used that had missing values. The number of myeloid blast cells measured in the bone marrow samples of patients was found to be roughly symmetrical according to the values presented in Table 3.2, as the mean equals the median. The method used to impute the missing value was the mean method. The distribution of values for CD19 as well as the number of monocytes measured in the blood samples were both positively skewed, as seen in the figures presented in Appendix A. Hence, the suitable method to impute the missing values for the two variables was the median method. Based on the information presented in Table 3.2, more than three quarters of the patients had above-average levels of CD19 and monocytes in their blood samples. The kurtosis of Bilirubin, CD10, CD13, CD19, and CD33 variables, were also all found to be above 3, which indicates that they have

heavier tails than the regular normal distribution. The median method of imputation was therefore used.

Table 3.2: Variables with missing values

Variable	Mean	Std. Dev.	Median	Min	Max	Skewness	Kurtosis	% of missingness
Bilirubin	0.6047	0.4838	0.5	0.1	5	4.9773	42.3546	0.010
Myeloid Blast	52.75	23.1455	52	5	94	0.0491	1.8224	0.005
CD10	4.011	8.396	2	0	91	7.3035	70.6242	0.021
CD13	79.83	23.3033	90	1	100	-1.599	4.8442	0.021
CD19	7.624	15.5228	1.2	0	94	3.1914	13.6573	0.021
CD33	76.77	30.2431	92	0	100	-1.3473	3.4052	0.021
Fibrogen	442.7	147.7018	421	0	701	0.1203	2.5488	0.073
Hemoglobin	9.591	1.6441	9.4	5.4	13.7	0.2028	2.8352	0.005

3.5.2 Risk factors for Acute Myeloid Leukemia without missing values

Table 3.3 tabulates a few of the key risk factors of AML that did not have missing values, while in Table 3.4 there are details noting how the majority of AML patients (56.84%) had above-average (mean=3.41; sd=0.70) albumin levels in comparison to the remainder (43.16%) who had below-average levels. Also according to Table 3.3, the values observed for albumin levels were found to be somewhat symmetrical (i.e., mean $\approx$ median). In the case of the levels of cell surface marker CD19, more than three quarters (78.69%) of the patients had below-average readings (mean=8.18; sd=15.95), while less than a quarter (21.31%) of the patients had above-average readings. The cytogenic categories of the patients were, in turn, regrouped as follows: 'High', 'Intermediate', 'Intermediate-low', and 'Lower'. These groups were represented as categories 1, 2, 3, and 4, respectively in Table 3.4. The same table also indicates that four to five out of 10 (47.54%) patients fell into Category 3 (Intermediate-low), while

three out of 10 (32.24%) fell into Category 2 (Intermediate); thereby making these two groups the more highly populated groups. In comparison, Category 1 (High) had the fewest patients (8.74%) and Category 4 (Lower) had slightly more than Category 1 at (11.48%).

Table 3.3: Variables without missing values

Variable	**Mean**	**sd**	**Median**	**Min**	**Max**	**Skewness**
ALBUMIN	3.411475	0.69664	3.5	0.7	5	-0.47534
ChemoSimplest	2.202186	1.543364	1	1	5	0.738157
cytocat	2.617486	0.802551	3	1	4	-0.22637
Infection	1.234973	0.425145	1	1	2	1.250183
ITD	2.20765	0.445424	2	1	3	0.86454
PRIORCHEMO	1.103825	0.30587	1	1	2	2.597582

The variable 'PRIORCHEMO' presented in Table 3.4 represented whether a patient had prior chemotherapy.

Table 3.4: Risk factors frequency

Variable	Category	Freq.	Percent
ALBUMIN	Below avg	79	43.16%
	Above avg	104	56.84%
CD19	Below avg	144	78.69%
	Above avg	39	21.31%
cytocat	1	16	8.74%
	2	59	32.24%
	3	87	47.54%
	4	21	11.48%
PRIORCHEMO	1	164	89.62%
	2	19	10.38%
Infection	1	140	76.50%
	2	43	23.50%
ITD	1	3	1.64%
	2	139	75.96%
	3	41	22.40%
ChemoSimplest	1 and 2	109	59.56%
	3	29	15.85%
	4	20	10.93%
	5	25	13.66%

As seen in Table 3.4, almost 9 out of 10 (89.62%) patients had prior chemotherapy (mean=1.10; sd=0.31), while 10.38% of the patients did not. About three quarters (76.50%) of the participants had also been diagnosed with an infection, while (23.50%) had not. Furthermore, three quarters (75.96%) of the patients did not experience ITD FLTS mutation, while a fifth (22.40%) fell into the 'unknown' category. Only 1.64% of patients actually experienced ITD FLTS mutation. In comparison, the variable

'Chemo.Simplest' presented in Table 3.4 was related to the specific Anthra treatment administered to patients. There were five types administered, of which almost six out of 10 (58.47%) patients had received Anthra-HDAC chemotherapy treatment. Of the remaining four types, a few patients each had received, in descending order of frequency, Flu-HDAC (15.85%), StdAraC-Plus (13.66%), HDAC-Plus non-Antra (10.93%) and Antra-Plus (1.09%).

3.6 Variable Selection

3.6.1 Random forests

In Section 3.4, a discussion was presented regarding the random forests technique and how it has the ability to rank variables in the order of most to least importance. Table 3.5 presents the variable importance for the first five variables out of the 142 variables chosen using random forests. It should be noted that some of the clinically important variables mentioned in Section 3.5 were not included in the 54% of the variables identified as important by the random forests algorithm (e.g., 'SEX' and 'Infection').

Table 3.5: Variable importance for the first six variables.

Variables	**Importance**
Cytogenic category	0.0067
Age	0.0058
CD19	0.0033
CCND3	0.0019
ALBUMIN	0.0018
Myeloid blast	0.0017
...	...

3.7 Descriptive Statistics

3.7.1 Gender and Age at death

Table 3.6: Gender and age of participants.

	Variable	Freq.	Percent
Gender	Male	98	51.31
	Female	93	48.69
Age group	4-25 years	10	5.24
	25-49 years	49	25.65
	50-64 years	67	35.08

As noted previously, the database used in this study consisted of participants who all had AML. Table 3.6 depicts how approximately half of these patients were male (51.31%), while 48.69% were female. The median patient age at diagnosis was 58 years. The Interquartile Range (IQR) was 46-67 years old. The maximum age at diagnosis was 87 years, while the minimum was 4 years.

3.7.2 Overall Survival and Survival Status

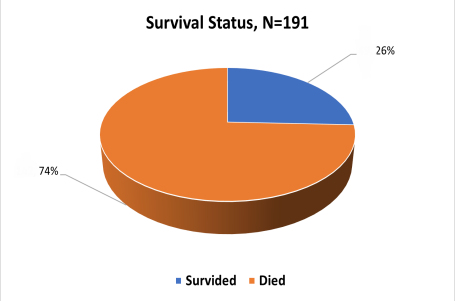


Figure 3.1: Survival status.

The survival status variable indicated whether a participant was alive or deceased at the end of the experiment. During the course of treatment, 142 (74.35%) patients were unsuccessful in their treatment (i.e., experienced death) while 49 (25.65%) survived, as shown in Figure 3.1.

Table 3.7 indicates that of those who survived, under half were female (49%) while the rest (51%) were male. There was, thus, little to no difference between males and females in terms of survival rate, which indicated that individuals, regardless of gender, have an almost equal chance of dying from AML. The hazard plot shown in Figure A.4 in Appendix A further indicates, however, that males tend to be slightly less likely to die when compared to females.

Table 3.7: Survival status by gender.

SEX	**Survived**		**Died**		**Total**
	Freq.	Percent	Freq.	Percent	Freq.
Male	25	51.02	73	51.41	98
Female	24	48.98	69	48.59	93

Table 3.8 indicates that those patients aged 50-64 years were almost equally likely to survive (34.7%), while those aged 25-49 years were more than twice as likely to survive (44.9%). In comparison, those aged younger than 25 years were five times more likely to survive (12.2%) than die, and those aged 65 years and older were 5 times more likely to have died than survive (42.9%).

Table 3.8: Survival status by age.

	Variable	Freq.	Percent
Age group	0=4-25 years	10	5.24
	25-49 years	49	25.65
	50-64 years	67	35.08
	65+ years	65	34.03
Total		191	100

The overall survival of patients was measured in weeks from diagnosis to exiting the study. The maximum number of weeks was 734 and the minimum was 3, as shown in Figure 3.2. The median survival period was 69 weeks (IQR: 29-331 weeks), the mean was 168 weeks, and the sd was 188 weeks. Slightly more than half of the patients survived up to 90 weeks (54.5%), and the survival period for about one tenth of the patients was between 91 and 180 weeks (13%). Very few patients survived during the 181-352 weeks period, but approximately a quarter (23.6%) survived for over 352 weeks. The longest survivor reached 734 weeks, while the shortest survivor reached only 4 weeks.

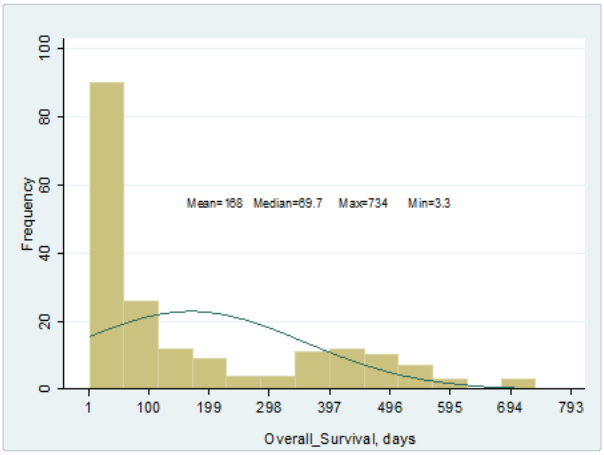


Figure 3.2: Distribution of the overall survival of patients in weeks.

3.7.3 Stepwise regression

From the 142 variables chosen by random forests, stepwise regression was used in fitting the models. Table 3.9 details the number of variables each model selected using this technique.

Table 3.9: Number of variables chosen by each model.

Model	No. of variables
Exponential	23
Weibull	48
Lognormal	36
Log-logistic	33
Cox PH	32

Chapter 4

Data Analysis and Results

4.1 Introduction

This chapter provides the results of the study of risk factors for AML using parametric and semi-parametric models. The variables with missing values were imputed in R using the mean and median imputation method, as the percentage of missingness was very low. Furthermore, the variables with missing values were only numeric not categorical. The data were analysed using Stata software (version 16) and R software (version 4.1.1).

The chapter first provides results from the variable selection procedure conducted using random forests. Thereafter, the chapter presents descriptive statistics in the form of frequencies and mean scores for the variables selected as important by the random forests and stepwise regression methods, and which were used in the model building. This chapter ends with details regarding the fitting of non-parametric, semi-parametric and parametric models and a comparison of the models using information based criteria and forecasting error measures.

4.2 Kaplan Meier estimator

This section presents curves that were constructed to assess the population failure and survival rates of patients using the non-parametric method. Figure 4.1 illustrates the trends in the survival rate of patients. The median survival rate is approximately 53% which is relatively low.

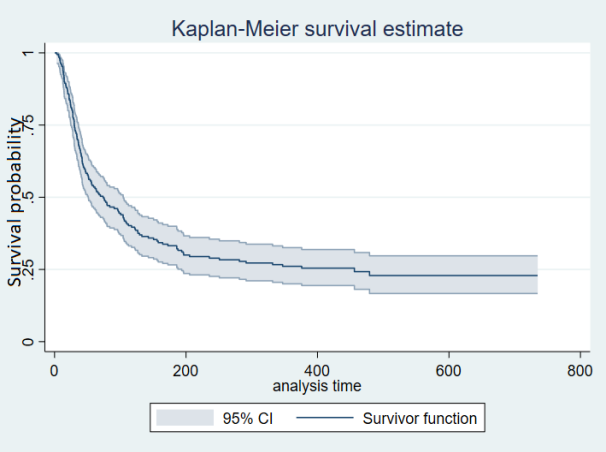


Figure 4.1: Kaplan-Meier survival Estimate.

Based on the information presented in Figure 4.1, the survival and failure rates were found to be inversely related. At 200 weeks of the study, the survival rate dropped from 100% to 30% and the failure rate increased from 0% to 70%. The number of patients dying increased to approximately 70% in 200 weeks. The steepest decrease in the survival rate (i.e., converse increase in failure rate) was encountered in the first 200 weeks. Then for the next 300 weeks, the rate of increase in failure (i.e., converse decrease in survival) decreased and approached a constant rate in the last 250 weeks.

In reference to Table A in Appendix A, at 3.29 weeks, there were 191 patients alive and only 1 patient experienced death. The probability of death at the serial time of 3.29 weeks was given as the survival rate of 0.9948. The 95% confidence interval for this event to occur at the serial time of 3.29 weeks was [0.9634; 0.993]. At 9 weeks into

the study, there were 184 patients alive and still participating in the study. One patient did, however, experience death at 9 weeks; hence, the probability of failure became 0.0367.

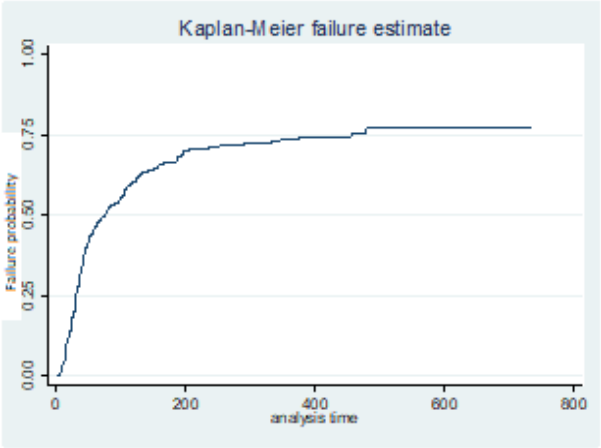


Figure 4.2: Kaplan-Meier failure Estimate.

The 95% confidence interval of this event was [0.9246; 0.9823]. In respect to week 205.4 (Table A), the probability of survival significantly decreased to 0.2946, which indicated a 70% decrease from the beginning of the study. At this point in the study, it became evident that the chance of survival for patients was very low. Furthermore, as shown by the two graphs for the K-M estimates, Table A supports that from week 479.1, the rate of patient death stopped decreasing and remained constant at 0.2290.

4.3 Predictive statistical models

This section presents the results of fitting the respective models: exponential, Weibull, log-normal, loglogistic and Cox PH. The results of the comparisons conducted between the predictive statistical models to find the one that best modelled the AML dataset using information-based criteria and forecast measures are also disccussed in this section. The other objective of the study was to evaluate and compare a variety

of parametric models (i.e., Weibull, exponential, log-logistic, and lognormal) to a semi-parametric model (Cox PH).

All models were trained using 80% of the data and tested using 20% of the data. The parametric estimates for the exponential, Weibull and lognormal models are in Appendix A. The hazard rate for age at diagnosis was 1.0025 for the exponential model fit shown in Table A.1 strongly suggests that older patients were more likely to die compared to younger patients. The variable had a very high level of significance close to 0. The lognormal results, in Table A.3, show that the age at diagnosis had a hazard rate of 1.0042 which supports the same conclusion drawn by the results from the exponential model that older patients were more likely to die compared to younger patients. The variable had a high level of significance which approximates to 0 as well. However, the Weibull model results in Table A.2 and the loglogistic model results in Table A.4 oppose those found by the exponential and lognormal model fits. According to the results found by the Weibull (HR$= 0.9988$, p-value < 0.05) and loglogistic (HR$= 0.9992$, p-value < 0.05) model fits older patients were less likely to die when compared to younger patients.

For the cytogenic categories all four model fit results had hazard rates less than 1 with very high levels of significance. For the parametric models, the patients with high cytogenic categories were found to be less likely to die compared to the patients with low cytogenic categories (exponential (HR$= 0.9074$, p-value < 0.05), Weibull (HR$= 0.9266$, p-value < 0.05), lognormal (HR$= 0.9489$, p-value < 0.05) and loglogistic (HR$= 0.9758$, p-value < 0.05)). Hence, the chance of death due to AML was seen to be reduced by having a high cytocat category.

The results show that there are no statistically significant associations between hemoglobin, 'SEX', 'Chemo.Simplest' and infections to all-cause mortality. The variables were found not statistically significant at the 5% level of significance. All 4 parametric models identified the variable protein ARC as highly significant to the outcome of a patient in the study. The lognormal model fit results suggested that patients with positive readings for the ARC protein were less likely to die when compared to those with negative levels (HR$= 0.8901$, p-value < 0.05). The lognormal

model fit results suggested that the probability of death by AML was reduced by ARC protein positive. The exponential, Weibull and loglostic model fit results, however strongly suggested that patients with negative ARC protein levels were less likely to die when copared to the patients with positive ARC protein levels. The hazard rates and p-values for the exponential, Weibull and loglogistic models were as follows: HR$= 1.1000$, p-value < 0.05, HR$= 1.3691$, p-value < 0.05 and HR$= 1.0062$, p-value < 0.05, respectively.

The black survival curve with the 95% confidence interval represents the general plot of the loglogistic survival model whereas the red survival curve gives the survival curve produced by the covariates chosen in Figure 4.3. This figure specifically indicates that the survival rates of patients decrease at a fast rate for the first 250 weeks but slows thereafter. The full summary of the model fit can be found in Table A.4 (Appendix A).

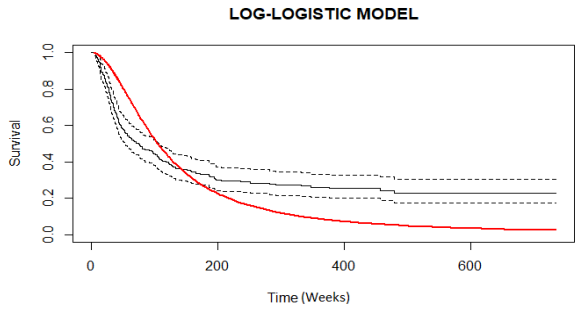


Figure 4.3: Log-logistic model fit survival curve.

Figure 4.4, furthermore, projects the results of testing the overall fit of the log-logistic model using Cox-Snell residuals. As can be seen in this figure, the hazard function for the log-logistic model successfully followed a straight line from the origin. These results indicate that the hazard function approximated the exponential function with Hazard Rate 1, which reflected how well the model fit the data.

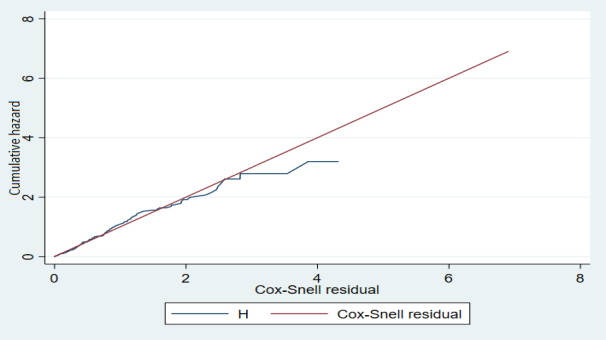


Figure 4.4: Cox-Snell residuals for the log-logistc model fit.

4.4 Cox PH model fit

Time to event curves analysed by Cox PH regression are commonly used to describe the outcome of clinical studies. The Cox PH model in this study was fitted using 80% of the data and tested using 20% of the data. The split presented a ratio of 152:39, respectively.

The results obtained from fitting the Cox PH model with the variables that were found to be significant via the variable selection methods, including the clinically significant variables, are presented in Table 4.1. The results show that even for the Cox PH model there are no statistically significant associations between hemoglobin, 'SEX', 'Chemo.Simplest' and infections to all-cause mortality just as found for the parametric models. The variables were also found to not be statistically significant to the 5% level of significance. The rest of the variables were, however, found to be statistically significant.

The Cox PH model fit results strongly indicate that older patients were more likely to die when compared to younger patients, as the hazard ratio was 1.042, with a very high level of significance (HR$= 1.042$, p-value < 0.05). The time to death due to AML was also found to increase for older patients. Those with high cyto.cat were, in turn, found to be less likely to die when compared to those with low cytogenic categories (HR$= 0.4430$, p-value < 0.05). Chances of dying due to AML were, thus, seen to be reduced by having a high cyto.cat.

Table 4.1: Cox PH model fit for all clinical covariates.

Variable	coef	exp(coef)	se(coef)	z-value	P-value
Age	0.0412	1.042061	0.008905	4.627	3.72E-06
cyto.cat	-0.81429	0.442953	0.172519	-4.72	2.36E-06
TGM2	0.313835	1.368664	0.12491	2.512	0.011988
CCNE2	0.391174	1.478716	0.14617	2.676	0.007447
PRIOR.CHEMO	1.227749	3.413538	0.474747	2.586	0.009706
PARP1.cl214	-0.39937	0.670746	0.159319	-2.507	0.012187
PRIOR.CANCER	-0.86144	0.422554	0.342115	-2.518	0.011803
GATA3	0.409482	1.506037	0.14999	2.73	0.006332
H3K27Me3	-0.47821	0.61989	0.170595	-2.803	0.005060
ITD	0.99566	2.706509	0.257678	3.864	0.000112
EGLN1	-0.46676	0.627031	0.159332	-2.929	0.003395
Hemoglobin	-0.11297	0.893181	0.0732	-1.543	0.122770
SEX	-0.1632	0.849424	0.221519	-0.737	0.461296
ALBUMIN	-0.48011	0.618716	0.164387	-2.921	0.003494
Chemo.Simplest	-0.11274	0.893384	0.082403	-1.368	0.171268
Infection	0.031828	1.03234	0.288892	0.11	0.912272

One highly significant result was that patients who had prior chemotherapy had an increased chance of death when compared to patients who had not had chemotherapy prior to being diagnosed with AML (HR$= 3.414$, p-value < 0.05). Such a result indicates that prior cancers significantly affect patients' chance of death from AML. Patients without prior cancers were also found to be less likely to die when compared to those who had been diagnosed with prior cancers (HR$= 0.4226$, p-value < 0.05). Chances of dying due to AML were found, therefore, to be reduced by not having any form of prior cancer.

AML patients with high levels of TGM2 and CCNE2, and who were found to have positive results of the ITD mutation, were found to be more likely to die when compared to those with lower levels of these genes and who were also found to not have the ITD mutation. Indeed, these patients' hazard ratios were all above 1, and the variables were found to be highly significant, as they reported p-values below 5%. A further highly significant result related to this finding was that patients with an above-average level of albumin tended to be less likely to die when compared to those who had lower levels (HR$= 0.6187$, p-value < 0.05). Thus, a strong relationship was found between patients with above-average levels of albumin and decreased risk of death from AML.

The Cox PH model fit can be seen in Figure 4.5. The median survival rate is approximately 60%. This figure includes all the significant variables as well as the clinically significant variables.

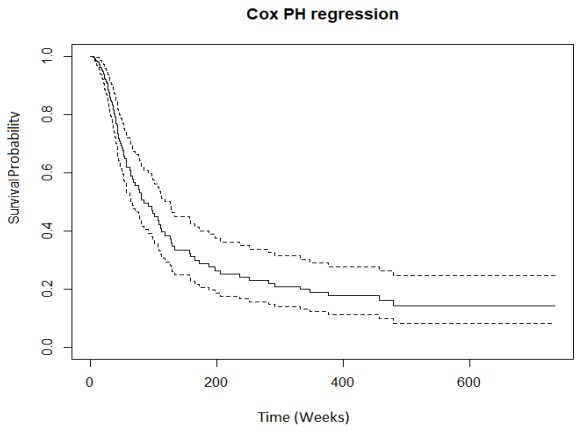


Figure 4.5: Cox PH survival curve.

4.4.1 Cox PH model residuals and assumptions analysis

This section presents the residuals mentioned previously in Section 3.4.5, which were assessed to check the overall fit of the model using Cox-Snell residuals as well as to

test the functional form of the predictors selected by the Cox PH model. In addition, the section summarises the results from testing the proportional hazards assumption of the Cox PH model. Specifically, Figure 4.6 shows that the hazard function does follows the 45-degree line but not too closely for very large values of time. This finding reveals that the Cox PH model fit the data well for mostly small values of time.

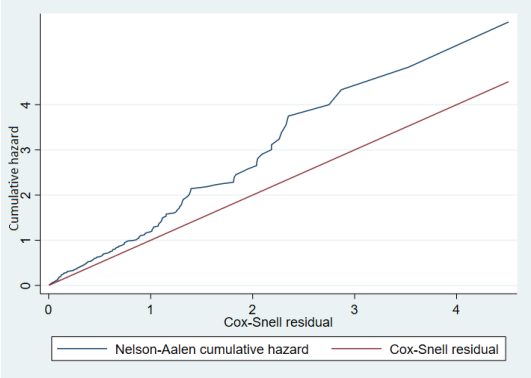


Figure 4.6: Cox PH Cox-Snell residuals.

Furthermore, in Table 4.2, the Schoenfeld residuals are computed to display the proportionality of all the clinically and other significant predictors according to the Cox PH model fit.

Table 4.2: Scaled Schoenfeld Residuals of Significant Covariates on the PH.

Variable	Chisq	DF	P-value
Age	1.78e-01	1	0.6732
cyto.cat	5.08e-02	1	0.8217
ARC	2.20e-01	1	0.6388
TGM2	3.70e+00	1	0.0543
PRKCD.pT507	8.53e-03	1	0.9264
CD74	8.96e-02	1	0.7647
CCNE2	2.95e-01	1	0.5869
PRIOR.CHEMO	8.92e-01	1	0.3448
PARP1.cl214	3.37e-01	1	0.5614
PRIOR.Cancer	3.26e-02	1	0.8566
GATA3	1.03e-02	1	0.9192
H3K27Me3	5.44e-01	1	0.4607
ITD	1.81e-03	1	0.9660
EGLN1	1.67e+00	1	0.1960
EGFR.pY992	1.89e-04	1	0.9890
Hemoglobin	3.30e-02	1	0.8558
SEX	4.70e-04	1	0.9827
ALBUMIN	6.74e+00	1	0.0094
Chemo.Simplest	1.21e+00	1	0.2706
Infection	6.44e+00	1	0.0111
GLOBAL	2.06e+01	20	0.4190

As evident in Table 4.2, there was strong evidence of proportionality which was supported by the small global test statistic and/or the large p-value. All the predictors in Table 4.2, thus, satisfy the proportional hazards assumption. Since their p-values are greater than the significance level of 10%, it was not possible to reject the null

hypotheses. Hence, it can be stated that there is no correlation between the covariates in Table 4.2 and time. The residual plots related to these findings can be seen in Appendix A.

The martingale residuals of the Cox PH model predictors, in reference to the functional form of the predictors classified as clinically significant to fit the Cox PH model, are all presented in Appendix A. Of particular note is that the relevant figures presented in that appendix show that 'SEX', 'Infection', 'cyto.cat', 'Chemo.Simplest', 'ALBUMIN' were all linear variables. In comparison, 'Age' and 'Hemoglobin' were nonlinear variables.

Table 4.3 shows the model performance evaluation techniques used, along with the aforementioned information-based criteria. The results of the forecasting measures used are presented in Table A.5 in Appendix A. The Chi-square test statistic and log likelihood for each model as well as the p-values are also displayed in Table 4.3.

According to the p-values for the models presented in Table 4.3, all were statistically significant to the 5% level of significance. The p-values for all models fitted, including the Cox PH model with a p-value ≈ 0. These results indicate that all the models were good at explaining the factors affecting AML patients.

Table 4.3: A summary of the results from training and testing the semi-parametric and parametric models.

Model	**IBS**	**C-index**	**AIC**	**BIC**	**AICc**	**Chisq**	**Loglik**	**P-value**
Exponential	0.145	0.7689	1363.6960	1424.1730	1297.917	183.72	-625.8	5.70E-29
Weibull	0.196	0.8051	1337.7200	1431.4600	1287.704	189.61	-604.6	1.60E-25
Lognormal	0.123	0.812	1332.3890	1423.1050	1283.892	157.61	-604.3	5.20E-20
Loglogistic	0.087	0.8347	1328.1810	1430.9930	1275.801	183.45	-593.7	3.60E-23
Cox PH	0.064	0.7834	950.6567	1005.5550	917.73			≈ 0

The median survival time obtained from the K-M was 69.71 weeks and the probability of survival at that time was approximately 53%. The survival probability from

the parametric models, at 69 weeks were 68.5%, 61.6%, 66.2% and 68.5% for the exponential, Weibull, lognormal and loglogistic models respectively. All model fits gave higher chances of survival than that from the non-parametric model. The models that had highest chance of survival at 69 weeks were the exponential model and the loglogistic model. At median survival time, using the parameters of the Weibull model to predict overall survival, at 69 weeks, the probability of survival was closest to that of the K-M model compared to the other parametric models. The semi-parametric model, Cox PH, at 69 weeks it gave a chance of survival of 55%. The chance of survival was even lower compared to that of the parametric models but it was higher than that of the non-parametric model, the K-M model. The chance of survival from the Cox PH is closest to that given by the K-M model compared to the parametric models.

The log-likelihood of the log-logistic model was, however, found to be the highest when compared to the other three models fitted, followed by the lognormal model. In turn, according to the forecast measures table found in Table A.5 in Appendix A, the semi-parametric model had the lowest root mean square error and the lowest mean absolute error, which indicated that the Cox PH model fit had the lowest errors in respect to the predictions conducted using the test sample. The Weibull model, in comparison, was found to have the lowest mean absolute percentage error. On average, the forecast conducted by the Cox PH model was found to be better than the other model fits because it has the better results for the forecast measures. The mean absolute deviation further indicated that values for the lognormal model were closer to the true mean than the other models.

Using IBS and c-index to compare the predictions of the parametric models, the loglogistic had the lowest IBS, closest to 0 and highest c-index which was closest to 1 making it the model that made the best predictions compared to other parametric models. The loglogistic model when compared to the semi-parametric model, the Cox PH, it does not make better predictions than the semi-parametric model. The Cox PH had the lowest value of IBS. The exponential model had the least favorable c-index compared to all models which made it the model with the least favorable predictions.

The lognormal model presented better forecast results when compared to the other parametric models; however, the log-logistic model was found to have the lowest AIC value amongst the parametric models. According to the AIC and BIC values, for the parametric models, the log-logistic and lognormal models were the best-fitted parametric models. Furthermore, while all the models had fairly good predictive power in explaining the factors affecting AML, the Weibull model had the greatest predictive power. The AICc, which accounts for the sample size, also supported the Cox PH model as being the best fitted model. Specifically, the Cox PH model minimised all three information-based criteria.

4.5 Significant variables

This section presents the significant variables for the models fit: exponential, Weibull, log-normal, loglogistic and Cox PH. A 5% level of significance was used to choose the variables for each model. Table 4.4 shows all the significant variables used to fit the respective models. As seen in Table 4.4, the Weibull distribution had the most significant variables when compared to the other models in this study. From the group of clinically significant variables, 'Age' and 'cyto.cat' were the only two that were found to be significant for all models used. Variables such as 'SEX', 'Chemo.Simplest', 'HGB' and 'Infection' were, by comparison, not found to be significant by all models used; however, clinically, they were found to be significant and were therefore retained in the models.

Table 4.4: Significant variables at 5 percent level of significance.

Exponential	Weibull	Log-normal	Loglogistic	Cox PH
Age	Age	Age	Age	Age
PA2G4.pS65	cyto.cat	cyto.cat	ABS.BLST	cyto.cat
TGM2	ARC	ALBUMIN	cyto.cat	TGM2
ARC	TGM2	CAV1	ARC	CCNE2
cyto.cat	ERG	TGM2	CD19	PRIOR.CHEMO
ERG	EGFR.pY992	HDAC1	BILIRUBIN	PARP1.cl214
CAV1	CCND3	ARC	CCND3	PRIOR.CANCER
WBC	HDAC1	CDKN2A	KIT	GATA3
TRIM24	EIF2AK2.pT451	EGLN1	H3K27Me3	H3K27Me3
CTNNA1	SMAD3	SRC	GAB2.pY452	ITD
	CAV1	GAB2.pY452	HDAC1	EGLN1
	H3K27Me3	GSKA_B.pS21_9	AKT1	ALBUMIN
	SRC	CCND3	MYC	
	GAB2.pY452	CDKN1B.pS10	EGLN1	
	CDKN1B.pS10	AKT1	TGM2	
	PTPN11	PTPN11	PDK1.pS241	
	CTNNA1	H3K27Me3	EGFR.pY992	
	ODC1			
	FOXO3.S318_321			
	TSC2			

Chapter 5

Conclusion and Discussion

A study on survival analysis was conducted and presented in this research report using the data collected from AML cancer patients. The data for this analysis were retrieved from the M.D. Anderson Cancer Center, with focus on patients who were treated using ara-C-based chemotherapy. This research report specifically focused on comparing two different groups of survival models, namely one semi-parametric model (i.e., the Cox PH model), and various parametric models (i.e., the exponential, Weibull, lognormal, and log-logistic models). The comparison presented in this study was based on attempting to determine which model best explains the predictors of AML as well as which best fits the data.

The purpose of the study, as noted in Chapter 1, was to find the best diagnosis for AML and to determine the factors that lead to either death or survival at a later stage for patients afflicted by the disease. In pursuit of this aim, it was necessary to establish the variables that all the models identified as significant via the variable selection method. All models chose the following variables: the age at diagnosis and the cytogenic category in which a patient falls; and the specific total number of myeloid blast cells measured in blood samples, represented by the variable 'ABS.BLST'. All models also chose 'ARC' and 'TGM2' as being significant, which are RPPA variables.

Of particular note is that the 'age.at.dx' was found to be highly significant in and across all the included models. This finding concurs with the descriptive results presented

previously, as it was found that the chance of survival doubled in the younger age groups. Specifically, the youngest age group (4-25 years) was 5 times more likely to survive than die according to the descriptive statistics. However, the 65+ age group were reported as having a 42.9% chance of dying. The age at diagnosis may, thus, help to categorise patient as higher or lower risk. The cytogenic category and Anthra-based treatment option administered were also found to be important to consider, as these significantly affected patients' chances of survival. These findings indicate that the majority of patients studied were given the Anthra-HDAC chemotherapy treatment option and were in the intermediate-low group of the cytogenic category.

Only 25.65% of the 191 patients who initially enrolled in the study were alive at the end of the study. This finding supports the assertion that there was an undeniable need for a more efficient diagnosis method. Furthermore, the majority of the patients did not survive beyond 250 weeks. The K-M estimator gave a good illustration of the underlying survival curve of the AML data without making any prior assumptions about the data. In particular, the K-M estimator survival curve illustrated the drastic decline of the rate of survival for patients in the first 200 weeks. This finding translates to the importance of the first 200 weeks for patients after diagnosis in determining whether or not they are likely to survive the cancer. The K-M estimator also provided the median survival rate of 69 weeks which is critical in comparing the predictions made by the models. From this research it is noted that the Cox PH model gave a similar approximation of the chance of survival at the median survival time as that of the K-M model. The parametric models predicted high chances of survival above 60%.

The hazard rates from the models further indicated that the older group was more likely to die due to AML compared to the younger group. The cytogenic category of the patient was also found to be a variable that may help to determine the risk level of a patient. Specifically, the hazard rates indicated that having a high cytocat level meant that a patient could be considered low-risk. Of further note was that the log-logistic hazard rate for the number of myeloid blast cells in a patient's blood sample was less than 1, with a p-value close to 0. This finding suggests that there is a strong relationship between patients with above-average total number of myeloid blasts cells and decreased death risk. It would be useful to use the variables found to be significant

in this study as prior information that may assist in predicting the overall survival potential of patients.

The results obtained from fitting the parametric models, using the variables selected by stepwise regression, revealed that the log-logistic model best explained the effects of the predictors of AML patients. Specifically, the log-logistic model minimised the AIC, but the lognormal model minimised the BIC. The log-logistic model also had the highest log-likelihood value as well as the highest significant number of variables. The log-logistic model also minimised the IBS index and maximized the C-index which meant that the model provided better predictions of the overall survival rate than the other parametric models. These findings indicate that the log-logistic model is the best predictive model when compared to the other parametric models. The residual test for the log-logistic model's overall fit further showed that the model fitted the AML data well.

In fulfillment of the main aim of the research report, the best model, overall, was found to be the Cox PH model. Similar to the previously noted models, the Cox PH model also identified the age at diagnosis and cytogenic category as important prior-information variables. In addition to those two variables, however, the model also identified the following variables: whether or not the patient had prior chemotherapy, ITD, and whether or not the patient had been diagnosed with a prior cancer. It was determined that the verification of these variables would be useful when diagnosing a patient. It should be noted although the Cox PH model did not identify the sex of a patient as a variable to consider at diagnosis, according to the hazard ratio for the variable 'SEX', being a male patient did indicate a slightly lower likelihood of death (approximately 38%).

Furthermore, according to the Cox-Snell residuals, the Cox PH model did not fit the data well for large values of time. This finding also applied to the log-logistic model. For small values of time, however, both models follow the 45-degree line closely. As a result, the Cox PH model was compared to the log-logistic model, using information-based criteria, where it was found that both the AIC and the BIC for the Cox PH model were minimised. The Cox PH model gave the best results for the IBS

and C-index. Thus, while the cumulative hazard for the log-logistic model followed the 45-degree line more closely than that of the Cox model, the Cox PH cumulative hazard was generally found to be above the expected 45-degree line. The log-logistic model was, therefore, able to identify more prognostic factors when compared to the Cox PH model; hence it was determined that the log-logistic model explains the effects of AML best.

The findings of this research are similar to those obtained by one researcher who analysed at this dataset. Xihui Lin and Hunter (2014) obtained a similar conclusion that the Cox PH model best fits this data. It explained the data well and gave better predictions. However, the standard Cox PH model was found not to give as good results as the bagged Cox PH model which the researchers used.

Chapter 6

Limitations and Recommendations for future

To improve comparison, it would be best to work with datasets that were previously widely used in research purposes. There are many other parametric models that can be included in future work to increase the comparison spectrum and increase the probability of finding very important covariates that will assist in predicting overall survival. A bagged Cox PH model could be used to increase the quality of the predictions. Obtaining the best model that predicts AML outcome may assist medical researchers to find the better ways of treating the patients, hence increasing overall survival time for patients with AML.

To account for the heterogeniety of patients, it would be best to use heterogeniety survival models that include frailty modeling in future studies.

References

Alamartine, E., Sabatier, J.-C., Guerin, C., Berliet, J.-M., and Berthoux, F. (1991). Prognostic factors in Mesangial Iga glomerulonephritis: an extensive study with univariate and multivariate analyses. *American Journal of Kidney Diseases*, 18(1):12–19. 11, 12

Ali, J., Khan, R., Ahmad, N., and Maqsood, I. (2012). Random Forests and Decision Trees. *International Journal of Computer Science Issues (IJCSI)*, 9(5):272. 24

Allison, P. D. (2010). *Survival analysis using SAS: A practical guide*, volume 1. Sas Institute, Beverly Hills, 2nd edition. 8, 9

Brentnall, A. R. and Cuzick, J. (2018). Use of the Concordance index for predictors of Censored Survival data. *Statistical methods in Medical research*, 27(8):2359–2373. 27

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2):261–304. 26

Chase Jr, C. W. (1995). Measuring Forecast Accuracy. *The Journal of Business Forecasting*, 14(3):2. 28

Cox, D. R. (1972). Models and life-tables regression. *Journal of the Royal Statistical Society Series B (Methodological)*, 34:187–220. 17

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using Random forests. *Pattern Recognition Letters*, 31(14):2225–2236. 24, 25

Gill, H., Yim, R., Pang, H. H., Lee, P., Chan, T. S., Hwang, Y.-Y., Leung, G. M., Ip, H.-W., Leung, R. Y., Yip, S.-F., et al. (2020). Clofarabine, cytarabine, and mitoxantrone in refractory/relapsed Acute Myeloid Leukemia: High response rates and effective bridge to allogeneic hematopoietic stem cell transplantation. *Cancer Medicine*, 10:3371–3382. 2

Goel, M. K., Khanna, P., and Kishore, J. (2010). Understanding survival analysis: Kaplan-meier estimate. *International Journal of Ayurveda Research*, 1(4):274–278. 16

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and Comparison of Prognostic Classification Schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545. 27

Grimwade, D. (2012). The changing paradigm of prognostic factors in acute myeloid leukaemia. *Best Practice & Research Clinical Haematology*, 25(4):419–425. 4

Guo, S. (2010). *Survival Analysis*, volume 1. Oxford University Press, 1 edition. 5

Han, H., Guo, X., and Yu, H. (2016). Variable selection using Mean decrease accuracy and Mean decrease gini based on Random forest. In *2016 7th ieee International Conference on Software Engineering and Service Science (ICSESS)*, pages 219–224. IEEE. 24, 25

Hassan, S. and Smith, M. (2014). Acute Myeloid Leukaemia. *Hematology*, 19(8):493–494. 3, 4

Heuser, M., Ofran, Y., Boissel, N., Mauri, S. B., Craddock, C., Janssen, J., Wierzbowska, A., and Buske, C. (2020). Acute Myeloid Leukaemia in Adult patients: ESMO Clinical Practice Guidelines for Diagnosis, Treatment and Follow-up. *Annals of Oncology*, 31(6):697–712. 4

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

Kazemitabar, J., Amini, A., Bloniarz, A., and Talwalkar, A. S. (2017). Variable importance using decision trees. *Advances in Neural Information Processing Systems*, 30(5):426–435. 23, 24

Klein, J. P. and Moeschberger, M. L. (2006). *Survival analysis: Techniques for Censored and Truncated data*, volume 1. Springer Science & Business Media, New York, 2 edition. 5, 9, 10, 14

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research*, 33(2):188–229. 26

Levis, M. (2011). Flt3/itd Aml and The Law of Unintended Consequences. *Blood*, 117(26):6987–6990. 5

Lowenberg, B., Downing, J. R., and Burnett, A. (1999). Acute Myeloid Leukemia. *New England Journal of Medicine*, 341(14):1051–1062. 1, 2, 3

Mardis, E. R., Ding, L., Dooling, D. J., Larson, D. E., McLellan, M. D., Chen, K., Koboldt, D. C., Fulton, R. S., Delehaunty, K. D., McGrath, S. D., et al. (2009). Recurring mutations found by sequencing an acute myeloid leukemia genome. *New England Journal of Medicine*, 361(11):1058–1066. 2, 3, 4

Miller Jr, R. G. (2011). *Survival analysis*, volume 66. John Wiley & Sons, New York, 2 edition. 15, 18

Nardi, A. and Schemper, M. (2003). Comparing Cox and Parametric models in clinical studies. *Statistics in Medicine*, 22(23):3597–3610. 29

Olusegun, A. M., Dikko, H. G., and Gulumbe, S. U. (2015). Identifying the limitation of stepwise selection for variable selection in regression analysis. *American Journal of Theoretical and Applied Statistics*, 4(5):414–419. 25

Oran, B., Giralt, S., Couriel, D., Hosing, C., Shpall, E., De Meis, E., Khouri, I., Qazilbash, M., Anderlini, P., Kebriaei, P., et al. (2007). Treatment of AML and MDS relapsing after reduced-intensity conditioning and allogeneic hematopoietic stem cell transplantation. *Leukemia*, 21(12):2540–2544. 2, 3, 4

Pourhoseingholi, M. A., Hajizadeh, E., Moghimi Dehkordi, B., Safaee, A., Abadi, A., Zali, M. R., et al. (2007). Comparing Cox regression and Parametric models for survival of patients with Gastric Carcinoma. *Asian Pacific Journal of Cancer Prevention*, 8(3):412–416. 9, 10, 11

Ravangard, R., Arab, M., Rashidian, A., AKBARI, S. A., Zare, A., and Zeraati, H. (2011). Comparison of the results of cox proportional hazards model and parametric models in the study of length of stay in a tertiary teaching hospital in Tehran, Iran. *ACTA MEDICA IRANICA*, 49(10):650–658. 12

Rowe, J. M. and Tallman, M. S. (2010). How i Treat Acute Myeloid Leukemia. *Blood, The Journal of the American Society of Hematology*, 116(17):3147–3156. 4

Schlichting, P., Christensen, E., Andersen, P. K., Fauerholdt, L., Juhl, E., Poulsen, H., Tygstrup, N., and for Liver Diseases, C. S. G. (1983). Prognostic factors in cirrhosis identified by Cox's regression model. *Hepatology*, 3(6):889–895. 9, 11

Schober, P. and Vetter, T. R. (2018). Survival Analysis and Interpretation of Time-to-Event data: The Tortoise and The Hare. *Anesthesia and Analgesia*, 127(3):792. 9

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464. 26

Sengsayadeth, S., Labopin, M., Boumendil, A., Finke, J., Ganser, A., Stelljes, M., Ehninger, G., Beelen, D., Niederwieser, D., Blaise, D., et al. (2018). Transplant Outcomes for Secondary Acute Myeloid Leukemia: Acute Leukemia working party of the European Society for Blood and Bone marrow Transplantation study. *Biology of Blood and Marrow Transplantation*, 24(7):1406–1414. 4

Sephton, S. E., Sapolsky, R. M., Kraemer, H. C., and Spiegel, D. (2000). Diurnal cortisol rhythm as a predictor of breast cancer survival. *Journal of the National Cancer Institute*, 92(12):994–1000. 9, 11, 12

Steck, H., Krishnapuram, B., Dehing-Oberije, C., Lambin, P., and Raykar, V. C. (2008). On ranking in Survival Analysis: Bounds on the Concordance index. *Advances in Neural Information Processing Systems*, pages 1209–1216. 27

Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990). Martingale-based residuals for Survival models. *Biometrika*, 77(1):147–160. 29, 30

Thirman, M. J. and Larson, R. A. (1996). Therapy-related myeloid leukemia. *Hematology/Oncology Clinics*, 10(2):293–320. 4

Valentini, C. G., Fianchi, L., Voso, M. T., Caira, M., Leone, G., and Pagano, L. (2011). Incidence of Acute Myeloid Leukemia after breast cancer. *Mediterranean Journal of Hematology and Infectious diseases*, 3(1). 5

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, 17(2):228–243. 26

Wang, S. J., Kalpathy-Cramer, J., Kim, J. S., Fuller, C. D., and Thomas Jr, C. R. (2010). Parametric Survival Models for predicting the benefit of adjuvant chemoradiotherapy in gallbladder cancer. *AMIA Annual Symposium Proceedings*, 2010:847–851. 12, 13

Willmott, C. J. and Matsuura, K. (2005). Advantages of the Mean Absolute Error (mae) over the Root Mean Square Error (rmse) in assessing Average Model Performance. *Climate Research*, 30(1):79–82. 28

Xihui Lin, Honglei Xie, G. C. and Hunter, D. G. (2014). A bagged Semi-parametric model for predicting Overall Survival time in AML Subchallenge 3. *Acute Myeloid Leukemia Outcome Prediction Challenge*, 1(1):1–2. 13, 58

Xu, X.-Q., Wang, J.-M., Lü, S.-Q., Chen, L., Yang, J.-M., Zhang, W.-P., Song, X.-M., Hou, J., Ni, X., and Qiu, H.-Y. (2009). Clinical and Biological characteristics of Adult biphenotypic Acute Leukemia in comparison with that of Acute Myeloid Leukemia and Acute Lymphoblastic Leukemia: A case series of a Chinese population. *Haematologica*, 94(7):919. 2

Xue, X., Xie, X., Gunter, M., Rohan, T. E., Wassertheil-Smoller, S., Ho, G. Y., Cirillo, D., Yu, H., and Strickler, H. D. (2013). Testing the Proportional Hazards Assumption in Case-cohort analysis. *BMC Medical Research Methodology*, 13(1):1–10. 30

Appendix A

Extra results

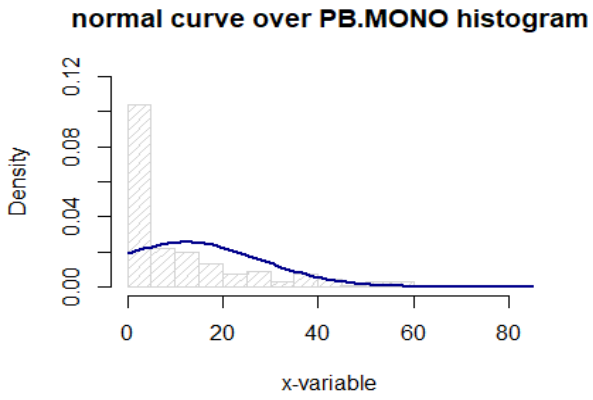


Figure A.1: Distribution for PB.MONO.

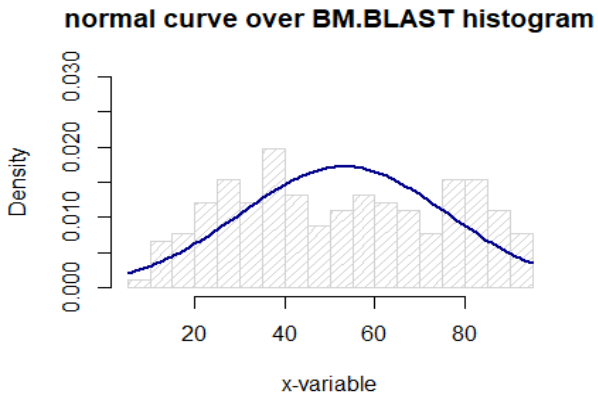


Figure A.2: Distribution for BM.BLAST.

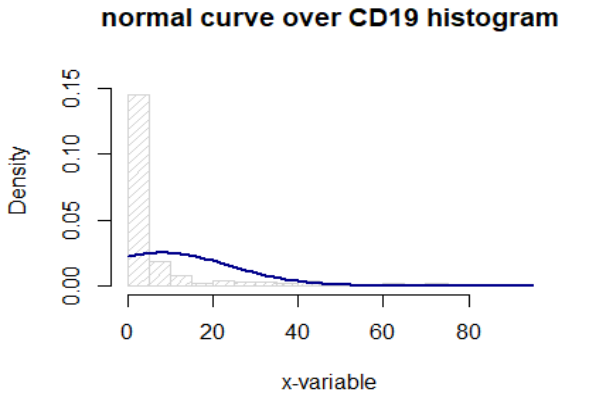


Figure A.3: Distribution for CD19.

KAPLAN-MEIER ESTIMATES

```
sts list

        failure _d:  vitalstatus == 2
  analysis time _t:  Overall_Survival

           At                  Survivor      Std.
```

Time	Risk	Fail	Lost	Function	Error	[95\% Conf.	Int.]
3.29	191	1	0	0.9948	0.0052	0.9634	0.9993
5.86	190	1	0	0.9895	0.0074	0.9588	0.9974
6.43	189	1	0	0.9843	0.0090	0.9521	0.9949
.							
.							
.							
43.43	116	1	0	0.6095	0.0355	0.5362	0.6749
43.86	115	1	0	0.6042	0.0355	0.5308	0.6698
50	110	1	0	0.5777	0.0359	0.5041	0.6444
.							
.							
.							
122	74	1	0	0.3911	0.0356	0.3214	0.4600
235.4	53	1	0	0.2891	0.0331	0.2260	0.3550
250.7	52	1	0	0.2835	0.0329	0.2209	0.3492
.							
.							
.							
501	14	0	1	0.2290	0.0336	0.1667	0.2973
703.9	2	0	1	0.2290	0.0336	0.1667	0.2973
734.9	1	0	1	0.2290	0.0336	0.1667	0.2973

Parametric Model fits

Table A.1: Exponential model fit for all clinical covariates.

Variable	Hazard rate	Value	Std. Error	Z-value	P-value
Age.at.Dx	1.0025	-0.0426	0.0089	-4.7600	0.0000
PA2G4.pS65	1.0814	-0.3690	0.1208	-3.0500	0.0023
TGM2	1.0894	-0.4366	0.1351	-3.2300	0.0012
ARC	1.1000	-0.6013	0.1327	-4.5300	0.0000
cyto.cat	0.9074	0.8062	0.1764	4.5700	0.0000
ERG	1.1784	-0.2858	0.1292	-2.2100	0.0269
CAV1	0.9315	0.3955	0.1477	2.6800	0.0074
WBC	0.9991	-0.0065	0.0024	-2.7200	0.0064
TRIM24	1.0709	-0.3702	0.1673	-2.2100	0.0269
CTNNA1	1.0685	-0.4186	0.1324	-3.1600	0.0016
replace_median_CD33	1.0003	-0.0022	0.0041	-0.5400	0.5925
SEX	0.9775	0.4144	0.2220	1.8700	0.0620
replace_median_HGB	0.9939	0.0286	0.0733	0.3900	0.6965
ALBUMIN	1.0782	0.1590	0.1678	0.9500	0.3434
Chemo.Simplest	1.0313	0.0788	0.0760	1.0400	0.2997
Infection	1.2220	-0.2555	0.2689	-0.9500	0.3421

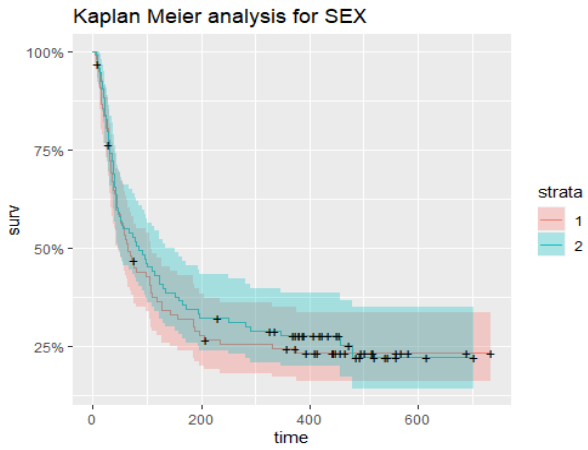


Figure A.4: SEX harzard plot.

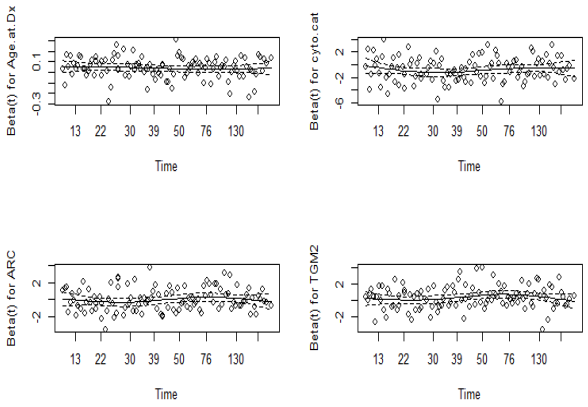


Figure A.5: Schoenfeld residuals (1).

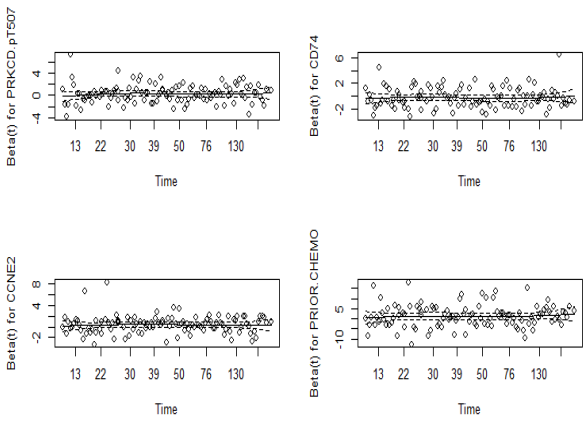


Figure A.6: Schoenfeld residuals (2).

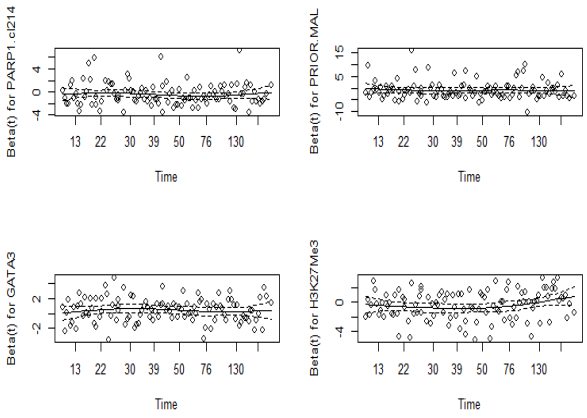


Figure A.7: Schoenfeld residuals (3).

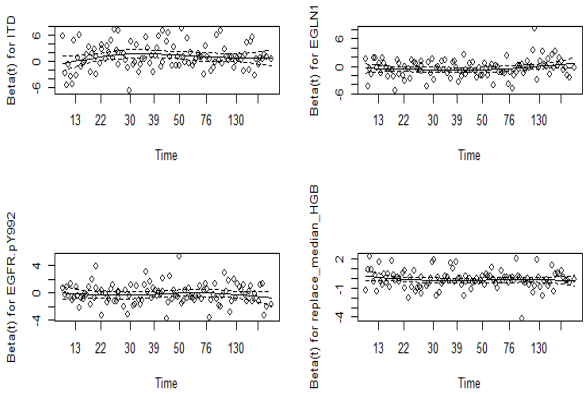


Figure A.8: Schoenfeld residuals (4).

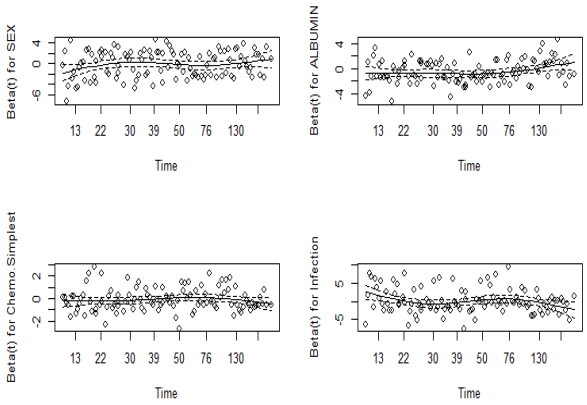


Figure A.9: Schoenfeld residuals (5).

Table A.2: Weibull model fit for all clinical covariates.

Variable	Hazard rate	Value	Std. Error	Z-value	P-value
Age.at.Dx	0.9988	-0.0289	0.0075	-3.8800	0.0001
cyto.cat	0.9266	0.9729	0.1440	6.7500	0.0000
ARC	1.3691	-0.6651	0.1280	-5.2000	0.0000
TGM2	1.1282	-0.3146	0.1251	-2.5100	0.0119
ERG	1.2014	-0.4401	0.1150	-3.8300	0.0001
EGFR.pY992	0.8982	0.4986	0.1759	2.8300	0.0046
CCND3	1.0838	-0.3663	0.1186	-3.0900	0.0020
HDAC1	0.9610	-0.6405	0.1273	-5.0300	0.0000
EIF2AK2.pT451	0.9050	0.5976	0.1205	4.9600	0.0000
SMAD3	0.9192	0.3174	0.1328	2.3900	0.0169
CAV1	0.8621	0.3323	0.1484	2.2400	0.0252
H3K27Me3	1.0453	0.3384	0.1366	2.4800	0.0132
SRC	0.6441	0.4877	0.1273	3.8300	0.0001
GAB2.pY452	1.0346	-0.3497	0.1323	-2.6400	0.0082
CDKN1B.pS10	0.7217	0.2702	0.1315	2.0600	0.0399
PTPN11	1.8410	-0.5983	0.2007	-2.9800	0.0029
CTNNA1	1.3113	-0.6257	0.1223	-5.1100	0.0000
ODC1	0.8091	0.2776	0.1351	2.0500	0.0399
FOXO3.S318_321	1.5431	-0.4650	0.1318	-3.5300	0.0004
TSC2	0.7213	0.4286	0.1569	2.7300	0.0063
SEX	1.0493	0.3274	0.1935	1.6900	0.0907
replace_median_HGB	1.0108	0.0468	0.0627	0.7500	0.4551
ALBUMIN	1.2081	0.1043	0.1551	0.6700	0.5013
Chemo.Simplest	1.1157	-0.0077	0.0664	-0.1200	0.9071
Infection	1.0302	-0.2032	0.2314	-0.8800	0.3798
Log(scale)	0.0724	-0.2470	0.0762	-3.2400	0.0012

Table A.3: Lognormal model fit for all clinical covariates.

Variable	Hazard rate	Value	Std. Error	Z-value	P-value
Age.at.Dx	1.0042	-0.0273	0.0068	-4.0200	0.0001
cyto.cat	0.9489	0.6645	0.1339	4.9600	0.0000
ALBUMIN	0.8941	0.3386	0.1394	2.4300	0.0151
CAV1	1.3202	0.4042	0.1410	2.8700	0.0042
TGM2	0.8117	-0.3362	0.1453	-2.3100	0.0206
HDAC1	0.9513	-0.4087	0.1335	-3.0600	0.0022
ARC	0.8901	-0.4665	0.1245	-3.7500	0.0002
CDKN2A	1.2191	0.2429	0.1126	2.1600	0.0310
EGLN1	1.1238	0.5607	0.1560	3.5900	0.0003
SRC	1.4458	0.4465	0.1286	3.4700	0.0005
GAB2.pY452	0.8858	-0.4408	0.1171	-3.7700	0.0002
GSKA_B.pS21_9	1.2053	0.4309	0.1391	3.1000	0.0019
CCND3	0.9433	-0.4564	0.1167	-3.9100	0.0001
CDKN1B.pS10	1.4046	0.3097	0.1396	2.2200	0.0265
AKT1	1.1848	0.3315	0.1334	2.4800	0.0130
PTPN11	0.6783	-0.4853	0.1902	-2.5500	0.0107
H3K27Me3	0.9124	0.3838	0.1351	2.8400	0.0045
SEX	1.0558	0.3175	0.1864	1.7000	0.0885
replace_median_HGB	0.9367	0.0511	0.0550	0.9300	0.3525
Chemo.Simplest	0.9049	-0.0101	0.0670	-0.1500	0.8806
Infection	0.9401	-0.0063	0.2499	-0.0300	0.9798
Log(scale)	1.2927	-0.0650	0.0699	-0.9300	0.3519

Table A.4: Loglogistic model fit for all clinical covariates.

Variable	Hazard rate	Value	Std. Error	Z-value	P-value
Intercept	0.6986	3.0500	1.0900	2.7900	0.0053
Age.at.Dx	0.9992	-0.0314	0.0064	-4.8900	0.0000
ABS.BLST	1.0000	0.0000	0.0000	-4.3200	0.0000
cyto.cat	0.9758	0.6680	0.1270	5.2500	0.0000
ARC	1.0062	-0.4580	0.1180	-3.9000	0.0001
replace_median_CD19	0.9996	0.0194	0.0073	2.6600	0.0079
replace_median_BILIRUBIN	0.9619	-0.4110	0.1660	-2.4800	0.0132
CCND3	0.9931	-0.3960	0.1070	-3.7100	0.0002
KIT	1.0104	-0.3210	0.1060	-3.0200	0.0025
H3K27Me3	1.0020	0.5520	0.1180	4.6700	0.0000
GAB2.pY452	0.9971	-0.4640	0.1230	-3.7800	0.0002
HDAC1	0.9930	-0.5290	0.1320	-4.0100	0.0001
AKT1	0.9981	0.2750	0.1240	2.2200	0.0266
replace_median_HGB	0.9975	0.0928	0.0493	1.8800	0.0597
MYC	0.9904	0.3890	0.1210	3.2200	0.0013
EGLN1	0.9992	0.5490	0.1320	4.1700	0.0000
TGM2	0.9976	-0.2800	0.1100	-2.5600	0.0106
PDK1.pS241	1.0053	-0.2250	0.1110	-2.0200	0.0433
EGFR.pY992	1.0086	0.3290	0.1270	2.5900	0.0097
replace_median_CD33	0.9999	-0.0031	0.0029	-1.0600	0.2914
SEX	1.0017	0.2510	0.1710	1.4700	0.1415
ALBUMIN	1.0119	0.1710	0.1470	1.1600	0.2443
Chemo.Simplest	1.0028	0.0726	0.0550	1.3200	0.1868
Infection	1.0212	0.0676	0.2170	0.3100	0.7557
Log(scale)	0.7202	-0.7340	0.0814	-9.0200	0.0000

Table A.5: A summary of the forecast results from training and testing the semi-parametric and parametric models.

Model	**RMSE**	**MAD**	**MAE**	**MAPE**
Exponential	3081.2160	126.0577	944.3498	0.8725
Weibull	2042.3720	108.2825	831.6822	0.7250
Lognormal	1761.5130	43.0330	654.5751	1.0611
Loglogistic	3763.0380	81.3889	815.5974	1.0458
Cox PH	236.8715	60.4927	148.9592	4.2525

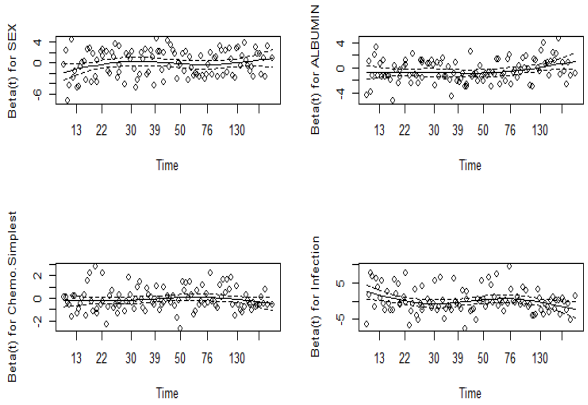


Figure A.10: Schoenfeld residuals (5).

Appendix B

R and STATA code

```
setwd("C:/Users/Maxeen/Desktop/Back up/aml challenge")
#Loading packages
library(ranger)       # a faster implementation of randomForest
library(randomForest) # basic implementation
library(survival)
library(randomForestSRC)
library(dplyr)
library(ggplot2)
library(flexsurv)
library(survminer)
library(ggfortify)



set.seed(123)
#Loading the data
RFDATA<- read.csv("DataSurv.csv")
```

```
sapply(RFDATA, class)

#Removing missing values
DT <- select(RFDATA, -X.Patient_id)
DATA<- na.omit(DT)

write.csv(DATA, "FitData.csv")
#Creating survival object
fit <- with(DATA, Surv(Overall_Survival, vital.status))

####Random forests for variable selection####
rd_fit <- ranger(Surv(Overall_Survival, vital.status)~ ., DATA,
 importance = "permutation",
                 splitrule = "extratrees",
                 verbose = TRUE)

var_impi <- data.frame(sort(round(rd_fit$variable.importance, 4),
decreasing = TRUE))
var_impi
head(var_impi)

write.csv(var_impi, "variable import.csv")


####Stepwise regression method####

stepwdata<-read.csv("Subsetdata.csv")
```

```
sw_data <- na.omit(stepwdata)
stw_data <- select(sw_data, -X.Patient_id)

#EXPONENTIAL DISTRIBUTION
Fitexp <- survreg(Surv(Overall_Survival, vital.status) ~ .
                  , data = stw_data, dist = "exponential")
summary(Fitexp)
Fitstartexp <- survreg(Surv(Overall_Survival, vital.status) ~ 1,
                       data = stw_data, dist = "exponential")
head(Fitstartexp)
stepModexp <- step(Fitstartexp, direction = "both", scope =
formula(Fitexp))

# Get the shortlisted variable.
shortlistedVars <- names(unlist(stepModexp[[1]]))
shortlistedVarsexp <- shortlistedVars[!shortlistedVars %in%
"(Intercept)"] # remove intercept

# Show in spreadsheet
print(shortlistedVarsexp)
write.csv(shortlistedVarsexp, "vars_exp.csv")

#WEIBULL DISTRIBUTION
Fitwei <- survreg(Surv(Overall_Survival, vital.status) ~  .
                  , data = stw_data, dist = "weibull")
Fitstartwei <- survreg(Surv(Overall_Survival, vital.status) ~ 1,
                       data = stw_data, dist = "weibull")
```

```
head(Fitstartwei)
stepModwei <- step(Fitstartwei, direction = "both", scope =
formula(Fitwei))

# Get the shortlisted variable.
shortlistedVars <- names(unlist(stepModwei[[1]]))
shortlistedVarswei <- shortlistedVars[!shortlistedVars %in%
 "(Intercept)"] # remove intercept

# Show in spreadsheet
print(shortlistedVarswei)
write.csv(shortlistedVarswei, "vars_wei.csv")

#LOGLOGISTIC DISTRIBUTION
Fitllg <- survreg(Surv(Overall_Survival, vital.status) ~ .
                  , data = stw_data, dist = "loglogistic")
Fitstartllg <- survreg(Surv(Overall_Survival, vital.status) ~ 1,
                       data = stw_data, dist = "loglogistic")
head(Fitstartllg)
stepModllg <- step(Fitstartllg, direction = "both", scope =
 formula(Fitllg))

#Get the shortlisted variable.
shortlistedVars <- names(unlist(stepModllg[[1]]))
shortlistedVarsllg <- shortlistedVars[!shortlistedVars %in%
"(Intercept)"] # remove intercept
```

```
# Show in spreadsheet
print(shortlistedVarsllg)
write.csv(shortlistedVarsllg, "vars_llg.csv")

#LOG-NORMAL DISTRIBUTION
Fitln <- survreg(Surv(Overall_Survival, vital.status) ~ .
                 , data = stw_data, dist = "lognormal")
Fitstartln <- survreg(Surv(Overall_Survival, vital.status) ~ 1,
                      data = stw_data, dist = "lognormal")
head(Fitstartln)
stepModln <- step(Fitstartln, direction = "both", scope =
 formula(Fitln))

#Get the shortlisted variable.
shortlistedVars <- names(unlist(stepModln[[1]]))
shortlistedVarsln <- shortlistedVars[!shortlistedVars %in%
 "(Intercept)"] # remove intercept

# Show in spreadsheet
print(shortlistedVarsln)
write.csv(shortlistedVarsln, "vars_ln.csv")


#COX PH MODEL

Fitcoxph <- coxph(Surv(Overall_Survival, vital.status) ~
    ABS.BLST+ACTB+Age.at.Dx+AHD+AKT1+AKT1_2_3.pS473
```

```
    +AKT1_2_3.pT308+ALBUMIN+ARC+ASH2L+BAD.pS136
    +BAD.pS155+BAX+BCL2+BCL2L11+BECN1+BID
      +BILIRUBIN+BIRC5+BM.BLAST+BMI1+BRAF+CASP3.cl175
      +CASP9+CASP9.cl330+CAV1+CCNB1+CCND3+CCNE1+CCNE2
     +CD10+CD13+CD19+CD33+CD74+CDK1+CDK2+CDK4
      +CDKN1B.pS10+CDKN2A+Chemo.Simplest+CLPP
     +COPS5+CREATININE+CREB1+CTNNA1+cyto.cat+DIABLO
      +EGFR.pY992+EGLN1+EIF2AK2+EIF2AK2.pT451+EIF2S1
      +EIF2S1.pS51.+EIF4E+ELK1.pS383+ERBB2.pY1248+ERG
      +FIBRINOGEN+Fli1+FN1+FOXO3+FOXO3.S318_321
        +GAB2.pY452+GAPDH+GATA3+GSKA_B+GSKA_B.pS21_9
       +H3histon+H3K27Me3+H3K4Me2+H3K4Me3+HDAC1+HDAC3
        +HGB+HNRNPK+HSP90AA1_B1+HSPB1+IGF1R+IRS1.pS1101
         +ITD+JMJD6+KDR+KIT+LGALS3+MAPK1+MAPK14.pT180Y182
         +MAPT+MET.pY1230_1234_1235+MSI2+MYC+NPM1+NPM1.3542
       +NR4A1+ODC1+PA2G4.pS65+PA2G4.pT37_46+PA2G4.pT70
        +PARK7+PARP1.cl214+PB.BLAST+PDK1+PDK1.pS241
         +PIK3CA+PIM2+PPARG+PRIOR.CHEMO+PRIOR.MAL
          +PRIOR.XRT+PRKCB.II+PRKCD.pT507+PTEN.pS380T382T383
        +PTPN11+RAC1_2_3+RPS6KB1.pT389+SIRT1+SMAD2+SMAD3
         +SMAD4+SMAD5+SMAD6+SOCS2+SPP1+SRC+SRC.pY527
         +STAT1.pY701+STAT3.pS727+STAT5A_B+STK11+STMN1
         +TGM2+TNK1+TP53+TP53.pS15+TRIM24+TRIM62
        +TSC2+VASP+WBC+YAP1+YAP1p+ZNF296
                  ,data = stw_data)
Fitstartcoxph <- coxph(Surv(Overall_Survival, vital.status) ~ 1
,data = stw_data)
```

```
head(Fitstartcoxph)
stepModcoxph <- step(Fitstartcoxph, direction = "both",
 scope = formula(Fitcoxph))

# Step 4: Get the shortlisted variable.
shortlistedVars <- names(unlist(stepModcoxph[[1]]))
shortlistedVarscoxph <- shortlistedVars[!shortlistedVars %in%
 "(Intercept)"] # remove intercept

# Show in spreadsheet
print(shortlistedVarscoxph)
write.csv(shortlistedVarscoxph, "vars_coxph.csv")



####Checking Percentage of missingness####
Data<- read.csv("Subsetdata.csv")
sapply(Data, class)

summary(Data)
#Removing missing values
is.na(Data)
complete.cases(Data)
mean(complete.cases(Data)) #0.895288 ->
16 cases with missing values

####Checking for skewness, kurtosis, summary
 of variables with missingness####
```

```
Data2<- na.omit(Data)

#1)Bilirubin
summary(Data2$BILIRUBIN)
sd(Data2$BILIRUBIN)
skewness(Data2$BILIRUBIN)
kurtosis(Data2$BILIRUBIN)

#2)Bm.Blast
summary(Data2$BM.BLAST)
sd(Data2$BM.BLAST)
skewness(Data2$BM.BLAST)
kurtosis(Data2$BM.BLAST)

#3)CD10
summary(Data2$CD10)
sd(Data2$CD10)
skewness(Data2$CD10)
kurtosis(Data2$CD10)

#4)CD13
summary(Data2$CD13)
sd(Data2$CD13)
skewness(Data2$CD13)
kurtosis(Data2$CD13)
```

```
#5)CD19
summary(Data2$CD19)
sd(Data2$CD19)
skewness(Data2$CD19)
kurtosis(Data2$CD19)

#6)CD33
summary(Data2$CD33)
sd(Data2$CD33)
skewness(Data2$CD33)
kurtosis(Data2$CD33)

#7)FIBROGEN
Data2<- transform( Data2, FIBRINOGEN =
as.numeric(Data2$FIBRINOGEN))
summary(Data2$FIBRINOGEN)
sd(Data2$FIBRINOGEN)
skewness(Data2$FIBRINOGEN)
kurtosis(Data2$FIBRINOGEN)

#8)HGB
summary(Data2$HGB)
sd(Data2$HGB)
skewness(Data2$HGB)
kurtosis(Data2$HGB)

####Imputing with the median of each variable as
```

```
they are numeric variables.####

#Loading the data
Imputedata<- read.csv("Subsetdata.csv")

#Function: Imputing Missing Values with median

# Return the column names containing missing observations
list_na <- colnames(Imputedata)[ apply(Imputedata, 2, anyNA) ]
list_na

Imputedata<- transform( Imputedata, FIBRINOGEN =
as.numeric(Imputedata$FIBRINOGEN))

median_missing <- apply(Imputedata[,colnames(Imputedata) %in% list_na],
                        2,
                        median,
                        na.rm =  TRUE)
Imputedata_replace <- Imputedata %>%
  mutate(replace_median_CD19  = ifelse(is.na(CD19),
median_missing[1], CD19),
         replace_median_BILIRUBIN  = ifelse(is.na(BILIRUBIN),
 median_missing[1], BILIRUBIN),
         replace_median_CD13  = ifelse(is.na(CD13),
median_missing[1], CD13),
         replace_median_CD33  = ifelse(is.na(CD33),
 median_missing[1], CD33),
```

```
        replace_median_CD10  = ifelse(is.na(CD10),
median_missing[1], CD10),
        replace_median_HGB  = ifelse(is.na(HGB),
median_missing[1], HGB),
        replace_median_FIBRINOGEN  = ifelse(is.na(FIBRINOGEN),
 median_missing[1], FIBRINOGEN),
        replace_median_BM.BLAST  = ifelse(is.na(BM.BLAST),
median_missing[1], BM.BLAST))
head(Imputedata_replace)


list_na2 <- colnames(Imputedata_replace)[ apply(Imputedata_replace,
 2, anyNA) ]
list_na2

write.csv(Imputedata_replace,"Imputeddata.csv")

cleanset5 <- select(Imputedata_replace, -c(BILIRUBIN, BM.BLAST,
CD13, CD19, CD33, CD10, HGB, FIBRINOGEN))

write.csv(cleanset5,"Imputeddatafinal.csv")


#### Model fits ####

#Model fits with variable selected from rf and stepwise, to choose
 significant variables.
```

```
#EXPONENTIAL(1)
Expmodel <- survreg(Surv(Overall_Survival, vital.status) ~
Age.at.Dx+PA2G4.pS65+ABS.BLST+TGM2
                    +ARC+cyto.cat+ERG+HDAC1+EIF2AK2.pT451
                    +SMAD3+CAV1+WBC+EIF2AK2+replace_median_HGB+TRIM24
                    +TNK1+CTNNA1+replace_median_CD33,
                    data = cleanset5, dist = "exponential")

summary(Expmodel)


#WEIBULL(1)
Weimodel <- survreg(Surv(Overall_Survival, vital.status) ~
Age.at.Dx+ABS.BLST+cyto.cat+ARC+TGM2
                    +ERG+AKT1+EGFR.pY992+CCND3+HDAC1
                    +EIF2AK2.pT451+SMAD3+EIF2AK2+CAV1+ALBUMIN
                    +BMI1+H3K27Me3+H3K4Me2+SRC+TNK1
                    +NPM1+RPS6KB1.pT389+AKT1_2_3.pT308+ERBB2.pY1248
+GAB2.pY452
                    +BAD.pS155+CDKN1B.pS10+PTPN11+CTNNA1+EGLN1
                    +CD74+GATA3+ODC1+TP53.pS15+FOXO3.S318_321
                    +MSI2+STAT1.pY701+TSC2+COPS5+PDK1.pS241
                    +EIF2S1.pS51.+BECN1+PIM2+PA2G4.pT70+BAX
                    +DIABLO+STMN1+replace_median_BILIRUBIN,
                    data = cleanset5, dist = "weibull")
```

```
summary(Weimodel)



#LOG-LOGISTIC(1)
Llgmodel <- survreg(Surv(Overall_Survival, vital.status) ~
 Age.at.Dx+ABS.BLST+cyto.cat+ARC
                    +replace_median_CD19+replace_median_BILIRUBIN
+ZNF296+CCND3
                    +KIT+H3K27Me3+GAB2.pY452+HDAC1
                    +AKT1+EIF2AK2+EIF2S1.pS51.+replace_median_HGB
                    +MYC+EGLN1+TGM2+PDK1.pS241
                    +EGFR.pY992+MAPK1+NPM1+CDKN1B.pS10
                    +GSKA_B.pS21_9+CD74+SMAD4+replace_median_CD33
                    +ERBB2.pY1248+NR4A1+COPS5+CASP3.cl175
                    +MAPK14.pT180Y182,
                    data = cleanset5, dist = "loglogistic")

summary(Llgmodel)



#LOGNORMAL(1)
Lnmodel <- survreg(Surv(Overall_Survival, vital.status) ~
Age.at.Dx+ABS.BLST+cyto.cat+ALBUMIN+CAV1
                   +TGM2+HDAC1+ARC+ERG+RPS6KB1.pT389
                   +CDKN2A+EGFR.pY992+SMAD6+EGLN1+BAD.pS155
                   +SRC+GAB2.pY452+ERBB2.pY1248+GSKA_B.pS21_9
+replace_median_CD10
```

```
                    +WBC+CCND3+CDKN1B.pS10+STK11+CDK2
                    +TNK1+AKT1+PIM2+replace_median_FIBRINOGEN+TP53.pS15
                    +PTPN11+H3K27Me3+AKT1_2_3.pT308+H3K4Me2+TRIM62
                    +COPS5,
                     data = cleanset5, dist = "lognormal")

summary(Lnmodel)


#COXPH(1)
Coxphmodel <- coxph(Surv(Overall_Survival, vital.status) ~
Age.at.Dx+ABS.BLST+cyto.cat+ARC+TGM2
                      +ZNF296+PRKCD.pT507+CD74+CCNE2+CDK1
                      +TP53+PRIOR.CHEMO+PARP1.cl214+PB.BLAST+TNK1
                      +PRIOR.MAL+GATA3+AKT1_2_3.pS473+AKT1
+replace_median_BILIRUBIN
                      +MAPT+H3K27Me3+Fli1+ITD+EGLN1
                      +HDAC3+Chemo.Simplest+SMAD3+PIM2+EGFR.pY992
                      +SMAD4+EIF4E ,
                     data = cleanset5)

summary(Coxphmodel)


####Model fits with testing####

##Model fits with variables selected from rf, stepwise and
```

```
those found as significant
#by the respective models.
#Model fits include clinically significant variables:
#Age.at.Dx, SEX, cyto.cat, HGB, ALBUMIN, Chemo.Simplest and Infection.

library(forecast)
library(tidyverse)
library(caret)
library(MLmetrics)
library(h2o)


#Loading Data
Modeldata <- read.csv("Imputeddatafinal.csv")
model_data <- select(Modeldata, -c(X.Patient_id,X))

#Checking for categorical variables

categorical_variables <- factor(model_data)

model_data<- transform( model_data,
                        AHD = as.numeric(model_data$AHD),
                        cyto.cat = as.numeric(model_data$cyto.cat),
                        PRIOR.CHEMO = as.numeric(model_data$PRIOR.CHEMO),
                        SEX = as.numeric(model_data$SEX),
                        Infection = as.numeric(model_data$Infection),
                        vital.status = as.numeric(model_data$vital.status),
```

```
                        ITD = as.numeric(model_data$ITD),
                        PRIOR.XRT = as.numeric(model_data$PRIOR.XRT),
                        PRIOR.MAL = as.numeric(model_data$PRIOR.MAL),
                        Chemo.Simplest = as.numeric(model_data$Chemo.Simplest)
)


set.seed(123) #Set seed for reproducibility


 #Split data into 80\% training and 20\% test
modeldata <- sort(sample(nrow(model_data), nrow(model_data)*0.8))

train<-model_data[modeldata,]
test<-model_data[-modeldata,]


#EXPONENTIAL(2)

Exp2model <- survreg(Surv(Overall_Survival, vital.status) ~
                       Age.at.Dx+PA2G4.pS65+TGM2+ARC
                     +cyto.cat+ERG+CAV1+WBC+TRIM24
                     +CTNNA1+replace_median_CD33+SEX
                     +replace_median_HGB+ALBUMIN
                     +Chemo.Simplest+Infection,
                     data = train, dist = "exponential")
```

```
Exp_con = concordance(Exp2model)
summary(Exp2model)

ExpPred <-  predict(Exp2model,test)
Expdata <- data.frame(

  Exp_RMSE = RMSE(test$Overall_Survival, ExpPred),

  Exp_R2 = R2(test$Overall_Survival, ExpPred),

  Exp_MAD = mad(test$Overall_Survival, ExpPred),

  Exp_MAE = MAE(test$Overall_Survival, ExpPred),

  Exp_MAPE = MAPE(test$Overall_Survival, ExpPred),

  AIC = extractAIC(Exp2model),

  BIC = BIC(Exp2model),

  AICc = AICc(Exp2model)

)

fitexpn<- flexsurvreg(Surv(Overall_Survival, vital.status) ~
                        Age.at.Dx+PA2G4.pS65+TGM2+ARC
                      +cyto.cat+ERG+EIF2AK2.pT451+SMAD3
```

```
                        +CAV1+WBC+EIF2AK2+TRIM24+CTNNA1
                        +replace_median_CD33+SEX+replace_median_HGB
                        +ALBUMIN+Chemo.Simplest+Infection,
data =train,dist = "exponential" )


plot(fitexpn, xlab="Time(days/weeks)", ylab= "Survival",
main="EXPONENTIAL MODEL")

summary(fitexpn, times=69)

#WEIBULL(2)

Wei2model <- survreg(Surv(Overall_Survival, vital.status) ~
                          Age.at.Dx+cyto.cat+ARC+TGM2
                        +ERG+EGFR.pY992+CCND3+HDAC1
                         +EIF2AK2.pT451+SMAD3+CAV1
                         +H3K27Me3+SRC+GAB2.pY452
                         +CDKN1B.pS10+PTPN11+CTNNA1
                         +ODC1+FOXO3.S318_321
                         +TSC2+SEX+replace_median_HGB
                        +ALBUMIN+Chemo.Simplest+Infection,
                        data = train, dist = "weibull")

summary(Wei2model)

WeiPred <- predict(Wei2model, test)
```

```
Weidata <- data.frame(

  Wei_RMSE = RMSE(test$Overall_Survival, WeiPred),

  Wei_R2 = R2(test$Overall_Survival, WeiPred),

  Wei_MAD = mad(test$Overall_Survival, WeiPred),

  Wei_MAE = MAE(test$Overall_Survival, WeiPred),

  Wei_MAPE = MAPE(test$Overall_Survival, WeiPred),

  AIC = extractAIC(Wei2model),

  BIC = BIC(Wei2model),

  AICc = AICc(Wei2model)

)


Wei_con = concordance(Wei2model)

fitwei<- flexsurvreg(Surv(Overall_Survival, vital.status) ~
                       Age.at.Dx+cyto.cat+ARC+TGM2
                     +ERG+EGFR.pY992+CCND3+HDAC1
                     +EIF2AK2.pT451+SMAD3+CAV1
```

```
                      +H3K27Me3+SRC+GAB2.pY452
                      +CDKN1B.pS10+PTPN11+CTNNA1
                      +ODC1+FOXO3.S318_321
                      +TSC2+SEX+replace_median_HGB
                      +ALBUMIN+Chemo.Simplest+Infection,
                      data = train, dist = "weibull")

plot(fitwei, xlab="Time(days/weeks)", ylab= "Survival",
main="WEIBULL MODEL")

summary(fitwei)


#LOGNORMAL(2)

Lln2model <- survreg(Surv(Overall_Survival, vital.status) ~
                       Age.at.Dx+cyto.cat+ALBUMIN+CAV1
                      +TGM2+HDAC1+ARC+CDKN2A+EGLN1
                      +SRC+GAB2.pY452+GSKA_B.pS21_9
                      +CCND3+CDKN1B.pS10
                      +AKT1+PTPN11+H3K27Me3+SEX
                      +replace_median_HGB+Chemo.Simplest
                      +Infection,
                      data = train, dist = "lognormal")

summary(Lln2model)
```

```
LlnPred <-  predict(Lln2model, test)
Llndata <- data.frame(

  Lln_RMSE = RMSE(test$Overall_Survival, LlnPred),

  Lln_R2 = R2(test$Overall_Survival, LlnPred),

  Lln_MAD = mad(test$Overall_Survival, LlnPred),

  Lln_MAE = MAE(test$Overall_Survival, LlnPred),

  Lln_MAPE = MAPE(test$Overall_Survival, LlnPred),

  AIC = extractAIC(Lln2model),

  BIC = BIC(Lln2model),

  AICc = AICc(Lln2model)
)



Lln_con = concordance(Lln2model)

fitlln<- flexsurvreg(Surv(Overall_Survival, vital.status) ~
                       Age.at.Dx+cyto.cat+ALBUMIN+CAV1
                     +TGM2+HDAC1+ARC+CDKN2A+EGLN1
                     +SRC+GAB2.pY452+GSKA_B.pS21_9
```

```
                      +CCND3+CDKN1B.pS10
                      +AKT1+PTPN11+H3K27Me3+SEX
                      +replace_median_HGB+Chemo.Simplest
                      +Infection,
                      data = model_data, dist = "lognormal")

plot(fitlln, xlab="Time(days/weeks)", ylab= "Survival",
main="LOGNORMAL MODEL")


#LOGLOGISTIC(2)

Llg2model <- survreg(Surv(Overall_Survival, vital.status) ~
                       Age.at.Dx+ABS.BLST+cyto.cat+ARC
                      +replace_median_CD19+replace_median_BILIRUBIN
                      +CCND3+KIT+H3K27Me3+GAB2.pY452+HDAC1
                      +AKT1+replace_median_HGB
                      +MYC+EGLN1+TGM2+PDK1.pS241+EGFR.pY992
                      +replace_median_CD33
                      +SEX+ALBUMIN+Chemo.Simplest
                      +Infection,
                      data = train, dist = "loglogistic")

summary(Llg2model)

LlgPred <-  predict(Llg2model, test)
Llgdata <- data.frame(
```

```
  Llg_RMSE = RMSE(test$Overall_Survival, LlgPred),

  Llg_R2 = R2(test$Overall_Survival, LlgPred),

  Llg_MAD = mad(test$Overall_Survival, LlgPred),

  Llg_MAE = MAE(test$Overall_Survival, LlgPred),

  Llg_MAPE = MAPE(test$Overall_Survival, LlgPred),

  AIC = extractAIC(Llg2model),

  BIC = BIC(Llg2model),

  AICc = AICc(Llg2model)
)


Llg_con = concordance(Llg2model)


#PLOT
fitllg<- flexsurvreg(Surv(Overall_Survival, vital.status) ~
                       Age.at.Dx+ABS.BLST+cyto.cat+ARC
                     +replace_median_CD19
+replace_median_BILIRUBIN
```

```
                    +ZNF296+CCND3+KIT+H3K27Me3
+GAB2.pY452+HDAC1
                    +AKT1+EIF2AK2+EIF2S1.pS51.
+replace_median_HGB
                    +MYC+EGLN1+TGM2+PDK1.pS241
+EGFR.pY992+MAPK1
                    +GSKA_B.pS21_9+CD74+SMAD4
+replace_median_CD33
                    +ERBB2.pY1248+COPS5+SEX
+ALBUMIN+Chemo.Simplest
                    +Infection,
                    data = model_data, dist = "llogis")

plot(fitllg, xlab="Time(days/weeks)", ylab= "Survival",
main="LOG-LOGISTIC MODEL")

#Cox sell for log-logistic model
streg AgeatDx ABSBLST cytocat ARC
replace_median_CD19
replace_median_BILIRUBIN ZNF296 CCND3 KIT
H3K27Me3 GAB2pY452 HDAC1 AKT1 EIF2AK2
EIF2S1pS51 replace_median_HGB
MYC EGLN1 TGM2 PDK1pS241 EGFRpY992
MAPK1 GSKA_BpS21_9 CD74 SMAD4
replace_median_CD33 ERBB2pY1248 COPS5 SEX
ALBUMIN ChemoSimplest Infection, dist(loglogistic)
predict double cs, csnell partial
```

```
predict double lgcs, csnell partial
stset lgcs, failure (vitalstatus==2)
sts generate km=s
generate double H=-ln(km)
line H lgcs lgcs, sort



#CoxPH(2)

Coxph2model <- coxph(Surv(Overall_Survival, vital.status) ~
                       Age.at.Dx+cyto.cat+ARC+TGM2+PRKCD.pT507
                     +CD74+CCNE2+PRIOR.CHEMO+PARP1.cl214
                     +PRIOR.MAL+GATA3+H3K27Me3+ITD+EGLN1
                     +EGFR.pY992+replace_median_HGB+SEX
                     +ALBUMIN+Chemo.Simplest+Infection ,
                    data = train)



summary(Coxph2model)
cox.zph(Coxph2model) #proportinal hazard test
plot(cox.zph(Coxph2model))
par(mfrow=c(2,2))

CoxphPred <-  predict(Coxph2model, test)
Coxphdata <- data.frame(
```

```
  Coxph_RMSE = RMSE(test$Overall_Survival, CoxphPred),

  Coxph_R2 = R2(test$Overall_Survival, CoxphPred),

  Coxph_MAD = mad(test$Overall_Survival, CoxphPred),

  Coxph_MAE = MAE(test$Overall_Survival, CoxphPred),

  Coxph_MAPE = MAPE(test$Overall_Survival, (CoxphPred*100)),

  AIC = extractAIC(Coxph2model),

  BIC = BIC(Coxph2model),

  AICc = AICc(Coxph2model)

)

Coxphcon = concordance(Coxph2model)

# Plot the baseline survival function
coxphfit <- survfit(Coxph2model)
plot(coxphfit, main = "Cox PH regression", xlab="Days", ylab = "Survival")
autoplot(coxphfit, main = "Cox PH regression", xlab="Days")

#Cox snell
```

```
quietly stcox SEX AgeatDx Infection cytocat ChemoSimplest
BMMONOCYTES ABSBLST HGB ALBUMIN CD13 CCND3
H3K27Me3 TRIM62 GSKA\_B, nohr mgale(mgl)
predict crs, csnell
stset crs, failure(vitalstatus)
sts generate hcs = na
line hcs crs crs, sort xlab(0 1 to 4) ylab(0 1 to 4)



#Martingale
quietly stcox AgeatDx ARC ALBUMIN cytocat TGM2 PRKCDpT507
CD74 CCNE2 PRIORCHEMO PARP1cl214 PRIORMAL GATA3
 H3K27Me3 ITD EGLN1 EGFRpY992 SEX replace_median_HGB
ChemoSimplest Infection, nohr mgale(mgl2)
predict crs, csnell
stset crs, failure(vitalstatus)
sts generate hcs = na
line hcs crs crs, sort xlab(0 1 to 4) ylab(0 1 to 4)
stset Overall_Survival, failure(vitalstatus==2) scale(1)
stcox AgeatDx ARC ALBUMIN cytocat TGM2 PRKCDpT507 CD74
CCNE2 PRIORCHEMO PARP1cl214 PRIORMAL GATA3 H3K27Me3
 ITD EGLN1 EGFRpY992 SEX replace_median_HGB ChemoSimplest
 Infection
predict mresid, mgale
lowess mresid AgeatDx, mean noweight title("") note("") m(o)
lowess mresid ALBUMIN , mean noweight title("") note("") m(o)
lowess mresid ChemoSimplest , mean noweight title("") note("") m(o)
```

```
lowess mresid cytocat , mean noweight title("") note("") m(o)
lowess mresid Infection , mean noweight title("") note("") m(o)
lowess mresid SEX , mean noweight title("") note("") m(o)
lowess mresid replace_median_HGB , mean noweight title("") note("") m(o)

# Comparison of models using IBS
library(survival)
library(rms)
library(pec)

m1 <- psm(Surv(Overall_Survival, vital.status!=2) ~
            Age.at.Dx+PA2G4.pS65+TGM2+ARC
          +cyto.cat+ERG+CAV1+WBC+TRIM24
          +CTNNA1+replace_median_CD33+SEX
          +replace_median_HGB+ALBUMIN
          +Chemo.Simplest+Infection,
          data = train, dist = "exponential")
m2 <- psm(Surv(Overall_Survival, vital.status!=2) ~
            Age.at.Dx+cyto.cat+ARC+TGM2
          +ERG+EGFR.pY992+CCND3+HDAC1
          +EIF2AK2.pT451+SMAD3+CAV1
          +H3K27Me3+SRC+GAB2.pY452
          +CDKN1B.pS10+PTPN11+CTNNA1
          +ODC1+FOXO3.S318_321
          +TSC2+SEX+replace_median_HGB
          +ALBUMIN+Chemo.Simplest+Infection,
          data = train, dist = "weibull")
```

```
m3 <- psm(Surv(Overall_Survival, vital.status!=2)~
            Age.at.Dx+cyto.cat+ALBUMIN+CAV1
          +TGM2+HDAC1+ARC+CDKN2A+EGLN1
          +SRC+GAB2.pY452+GSKA_B.pS21_9
          +CCND3+CDKN1B.pS10
          +AKT1+PTPN11+H3K27Me3+SEX
          +replace_median_HGB+Chemo.Simplest
          +Infection,
          data = train, dist = "lognormal")
m4 <- psm(Surv(Overall_Survival, vital.status!=2)~
            Age.at.Dx+ABS.BLST+cyto.cat+ARC
          +replace_median_CD19+replace_median_BILIRUBIN
          +CCND3+KIT+H3K27Me3+GAB2.pY452+HDAC1
          +AKT1+replace_median_HGB
          +MYC+EGLN1+TGM2+PDK1.pS241+EGFR.pY992
          +replace_median_CD33
          +SEX+ALBUMIN+Chemo.Simplest
          +Infection,
          data = train, dist = "loglogistic")
m5 <- coxph(Surv(Overall_Survival, vital.status!=2)~
              Age.at.Dx+cyto.cat+ARC+TGM2+PRKCD.pT507
            +CD74+CCNE2+PRIOR.CHEMO+PARP1.cl214
            +PRIOR.MAL+GATA3+H3K27Me3+ITD+EGLN1
            +EGFR.pY992+replace_median_HGB+SEX
            +ALBUMIN+Chemo.Simplest+Infection ,
            data = train,x=TRUE,y=TRUE)
```

```
mbrier <- pec(list("Exponential"=m1, "Weibull"=m2,
                   "Lognormal"=m3, "Loglogistic"=m4,
                   "CoxPH"=m5),data=train,formula
              =Surv(Overall_Survival, vital.status!=2)~1)
print(mbrier)


count
hist AgeatDx, frequency norm
su AgeatDx, d
tab AgeatDx
gen Agegroup= AgeatDx
recode Agrgroup min/24=0 25/49=1 50/65=2 65/max=3
recode Agegroup min/24=0 25/49=1 50/65=2 65/max=3
tab Agegroup
recode Agegroup min/24.9=0 25/49.9=1 50/65=2 65/max=3
tab Agegroup
drop Agegroup
recode Agegroup min/24.2=0 25/49.9=1 50/65=2 65/max=3
gen Agegroup= AgeatDx
gen Agegroup= AgeatDx
recode Agegroup min/24.2=0 25/49.9=1 50/65=2 65/max=3
tab Agegroup
tab SEX Age group
tab SEX Agegroup
tab Agegroup SEX
tab Agegroup SEX, col chi
tab Agegroup SEX
```

```
tab AHD
tab Overall_Survival
tab vitalstatus
tab Overall_Survival
hist Overall_Survival, frequency norm
su Overall_Survival,d
tab WBC
su WBC, d
su ABSBLST- HGB, d
gen WBC_1= WBC
gen ABSBLST_1= ABSBLST
gen BMBLAST_1= BMBLAST
gen BMMONOCYTES_1= BMMONOCYTES
gen BMPROM_1= BMPROM
gen PBBLAST_1= PBBLAST
gen PBMONO_1= PBMONO
gen PBPROM_1= PBPROM
gen HGB_1= HGB
gen PLT_1= PLT
gen LDH_1= LDH
gen ALBUMIN_1= ALBUMIN
gen BILIRUBIN_1= BILIRUBIN
gen CREATININE_1= CREATININE
gen FIBRINOGEN_1= FIBRINOGEN
su WBC
su WBC, d
recode WBC_1 min/31=0 31.1/max =1
```

```
tab WBC_1
su ABSBLST_1
recode ABSBLST_1 min/18203=0 18203.1/max=1
tab ABSBLST_1
su BMBLAST_1
su BMBLAST_1
recode BMBLAST_1 min/53=0 53.1/max=1
tab BMBLAST_1
su BMMONOCYTES_1
recode BMMONOCYTES_1 min/4.4=0 4.41/max=1
tab BMMONOCYTES_1
su BMPROM_1
recode BMPROM_1 min/1.4=0 1.41/max=1
tab BMPROM_1
su PBBLAST_1
recode PBBLAST_1 min/35=0 35.1/max=1
tab PBBLAST_1
su PBMONO_1
recode PBMONO_1 min/12.4=1 12.41/max=1
tab PBMONO_1
gen PBPROM_2= PBPROM
recode PBMONO_2 min/12.4=0 12.41/max=1
recode PBPROM_2 min/12.4=0 12.41/max=1
tab PBPROM_2
tab PBMONO
drop PBPROM_2 PBMONO_1
gen PBMONO_1=PBMONO
```

```
su PBMONO_1
recode PBMONO_1 min/12.4=-0 12.401/max=1
tab PBMONO_1
su BMPROM_1
tab BMPROM_1
su PBBLAST_1
tab PBBLAST_1
tab PBPROM_1
su PBPROM_1
recode PBPROM_1 min/0.6=0 0.61/max=1
tab PBPROM_1
su HGB_1
recode HGB_1 min/9.7=0 9.701/max=1
tab HGB_1
su PLT_1
recode PLT_1 min/78=0 78.01/max=1
tab PLT_1
su LDH_1
recode LDH_1 min/1577=0 1577.01/max=1
tab LDH_1
su ALBUMIN_1
recode ALBUMIN_1 min/3.4=0 3.401/max=1
tab ALBUMIN_1
su BILIRUBIN_1
recode BILIRUBIN_1 min/0.6=0 0.601/max=1
tab BILIRUBIN_1
su CREATININE_1
```

```
recode CREATININE_1 min/1=0 1.01/max=1
tab CREATININE_1
su FIBRINOGEN_1
recode FIBRINOGEN_1 min/440=0 440.01/max=1
tab FIBRINOGEN_1
su PBMONO_1
su WBC
tab vitalstatus SEX
tab vitalstatus SEX, col chi
tab vitalstatus Agegroup
tab  Agegroup vitalstatus
tab  Agegroup vitalstatus, col chi
tab SEX vitalstatus
tab Overall_Survival
gen Overall_Surv= Overall_Survival
recode Overall_Surv min/90=0 91/180=1 181/352=2 353/max=3
tab Overall_Surv
recode Overall_Surv 90.1=0
recode Overall_Surv 90.1=1
br Overall_Surv
sort Overall_Surv
replace Overall_Surv = 1 in 191

tab Overall_Surv
tab CD13
su CD13- CD19
gen CD13_1= CD13
```

```
gen CD33_1= CD33
gen CD34_1= CD34
gen CD7_1= CD7
gen CD10_1= CD10
gen CD20_1= CD20
gen HLADR_1= HLADR
gen CD19_1= CD19
drop ACTB ZNF346
drop ACTB-ZNF346
drop AIFM1 -ZNF346
drop AIFM1 - ZNF296

su CD13_1- CD19_1
tab CD13_1
su CD13_1- CD19_1
recode CD13_1 min/80=0 80.01/max=1
tab CD13_1
recode CD33_1 min/78=0 78.01/max=1
tab CD33_1
recode CD34_1 min/50=0 50.01/max=1
tab CD34_1
recode CD7_1 min/16=0 16.01/max=1
tab CD7_1
recode CD10_1 min/4=0 4.01/max=1
tab CD10_1
recode CD20_1 min/0.94=0 0.94.01/max=1
recode CD20_1 min/0.94=0 0.9401/max=1
```

```
tab CD10_1
tab CD20_1
su HLADR_1
recode HLADR_1 min/0.78=0 78.01/max=1
tab HLADR_1
recode HLADR_1 min/78=0 78.01/max=1
tab HLADR_1
drop HLADR_1
gen HLADR_1=HLADR
su HLADR_1
recode HLADR_1 min/78=0 78.01/max=1
tab HLADR_1
su CD19_1
recode CD19_1 min/8.2=0 8.201.01/max=1
recode CD19_1 min/8.2=0 8.201/max=1
tab CD19_1
```