Semi-supervised Learning for Pancreas Cancer Survival Analysis

By

Poonam Ahir

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science (MSc.) in Computational Sciences

The Faculty of Graduate Studies
Laurentian University
Sudbury, Ontario, Canada

THESIS DEFENCE COMMITTEE/COMITÉ DE SOUTENANCE DE THÈSE
Laurentian Université/Université Laurentienne
Faculty of Graduate Studies/Faculté des études supérieures

Title of Thesis Titre de la thèse	Semi-supervised Learning for Pancreas Cancer Survival Analysis	
Name of Candidate Nom du candidat	Ahir, Poonam	
Degree Diplôme	Master of Science	
Department/Program Département/Programme	Computational Sciences	Date of Defence Date de la soutenance July 14, 2021

APPROVED/APPROUVÉ

Thesis Examiners/Examinateurs de thèse:

Dr. Kalpdrum Passi
(Supervisor/Directeur(trice) de thèse)

Dr. Ratvinder Grewal
(Committee member/Membre du comité)

Dr. Peter Adamic
(Committee member/Membre du comité)

Dr. Bhabha Krishna Mohanty
(External Examiner/Examinateur externe)

Approved for the Office of Graduate Studies
Approuvé pour le Bureau des études supérieures
Tammy Eger, PhD
Vice-President Research (Office of Graduate Studies)
Vice-rectrice à la recherche (Bureau des études supérieures)
Laurentian University / Université Laurentienne

ACCESSIBILITY CLAUSE AND PERMISSION TO USE

Abstract

In this thesis, the survival analysis was of interest for the high dimensional data, in which the number of observations in the study is much less than the number of parameters, usually the clinical datasets are in this type because there are few experiments and each one includes many gene expressions. Treating with high dimensional data is necessary, because the redundant and non-prognostic genes can lead the researchers to incorrect results. In this study $L_{1/2}$ regularization has been used to shrink the coefficients of unimportant genes toward zero. Four methods of Single Cox, Single AFT, Semi-supervised Cox and Semi-supervised AFT have been used to implement survival analysis. The aim of the simulation study was to compare the four models in correctly detecting the prognostic genes. So, for two types of correlated and uncorrelated simulated data with 3 different sample sizes, the four models were compared. The single cox is sensitive to sample size but the semi-supervised cox model is less sensitive to sample size because we get a high value of average correctly selected parameters also in low sample size. The total number of selected parameters is lower in correlated data compared with uncorrelated data. Hence the precision is higher for correlated data using the semi-supervised cox model. The Semi-supervised Cox model implemented in this study was done by using Modified Newton Raphson method and coordinate descent to minimize the loss function. In the Semi-supervised method, the censored data were imputed by using the mean imputation method. The fraction of censoring right in our study is more. We get slightly more parameters in simulation compared with previous study. But we also found much correctly the prognostic genes in our study. The results of semi-supervised cox were seen to be better than previous study, in both simulation study and the real dataset.

Acknowledgements

I cannot express enough thanks to my thesis supervisor Dr. Kalpdrum Passi for his continued support and encouragement. His patience, motivation and enthusiasm helped me in all the time of research of the thesis. I must express my gratitude to my friends and family members for being with me at every moment and providing continues moral boosting and affection during thesis work. And thanks to my brother and sister for believing me that I can achieve this. A special thanks to my parents without them none of this would indeed possible. Your encouragement when the times got rough are much appreciated.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AFT	Accelerated Failure Time
IPW	Inverse Probability Weighting
OLS	Ordinary Least Squares
DLBCL	Diffuse Large B-Cell Lymphoma
AML	Acute Myeloid Leukemia
BMI	Body Mass Index
IBS	Integrated Brier Score
NSP	Number of Selected Parameters
CI	Concordance Index
CDF	Cumulative Distribution Function

Chapter 1

Introduction

In the field of clinical research, the main and foremost intention is the development of specific tools for the accurate and timely prediction of both risk profiling and survival time of the patients depending upon the microarray data of their DNA along with various other specific clinical parameters. There are many techniques that exists in the related literature that are timely used to carry out survival analysis and risk profiling mainly. Many of the models are used by these literatures. Among these models are the AFT (Accelerated Failure Time) model and Cox proportional hazards model which are most widely used [1]. The main reason why Cox model is widely used for the assessment of in the field of clinical research and survival analysis for the cancer patients is its ability to assess the different kind of genes with their particular significance [2]. As compared to Cox model, the significance of AFT (Accelerated Failure Time) model comes from the requirement of time data series analysis during the examination and analysis of relationship linking both profiles of microarray high dimensional genes and survival analysis outcomes. Due to this perspective, the AFT model is analyzed and implemented widely in recent case studies. Since, due to wide use of these aforementioned models [3], various other models for cancer survival analysis are not used as they have not exemplified and indicated their accuracy according to the expectations of researchers. However, in cancer survival analysis, the problems linked to accuracy are some of the preliminary difficulties [4]. For the improvement of accuracy and results in cancer survival analysis, one of the two main difficulties needs to be compromised either the preliminary dilemma of small sized sample data and censored survival analysis data in comparison with high dimensional co-heterogeneities in Cox model [5]. Due to the larger interest

in microarray gene studies, the high dimensional cancer survival analysis enticed more curiosity. This is due to the reason that it becomes very challenging in statistical terms. As compared to simple microarray data sample say n, the number of genes we say p is much greater which is indicated as p>>n. In cancer survival analysis, the availability of analyzed follow up data samples significantly reduce size of sample data [6]. Thus, only a small amount of microarray sample datasets of human tumor dispenses follow up data [2] for clinical research purposes even in the case of gene expression databases which are publicly available for clinical research. In case of Cox model, both the low-risk classification and high-risk classification mainly depends on conventional and established supervised learning techniques. For the process of learning in these traditional techniques, only completed data is useful. Complete data refers to specific data samples which proceeds with proper clinical follow up processes. Whereas, the censored data samples which without any follow up clinical processes are not used, hence ignored. However, in Cox model, both the censored data samples and small sized data samples are obstructions for the achievement of accurate and robust results for cancer survival analysis [7]. Although the Cox model and AFT model are of great importance in cancer survival analysis, another technique is coming in place for the achievement of better results from the combination of complete data(uncensored) and censored data. This technique is called semi-supervised machine learning technique. This technique indicates that both censored data and complete data can be used in limited conjunction to obtain more accurate and robust results with the substantial amount of improvement in accuracy and learning capabilities. The technique of semi supervised learning is also seeming to be efficient while solving certain biological problems mainly protein classification including prediction of certain drug interaction and predictions, prediction of human disease and proteins interaction. There are many approaches based on semi supervised

learning techniques which are widely used to analyze and examine gene expression data [5]. Some of the techniques include semi supervised predictions based on principal component regression and semi supervised learning and classification based on k-nearest neighbor clustering.

Censoring occurs once incomplete data is accessible regarding the survival time of some samples in our study. Censoring is classified into three types, namely, point censoring, interval censoring and left censoring. Point-censoring, the most common type of censoring, occurs when the survival time is “incomplete” at the right side of the follow-up period. This censoring is known as right censoring. By right censoring, it's meant that the survival time is just legendary to exceed an explicit time. We've used the right censoring in this study. Interval censoring happens once the place to begin is outlined by an occasion of treatment. Left censoring occurs when the person’s survival time is less than or equal to the observed survival time [8].

In case of AFT (Accelerated Failure Time) model, the main focal point is the analysis of cancer survival analysis through same phenotype disease in comparison with different genotype cancers. The gathering of more accurate and robust results for survival analysis in AFT model, the sample size is increased through the replacement of every censored observation with imputed values by using certain estimators with the use of specific methods such as mean imputed variation method, inverse probability weighting method referred as IPW method, rank based methodology and Buckley James methodology [9]. However, all of these estimation methodologies make assumptions about AFT model and its usage relevant to specific patients having exact same cancer phenotype. The common and unspecified distribution of probabilities should be estimated and specified by the similar survival times of patient [10]. Additionally, the inconsistencies linked to certain disease progressions and response of patient treatment can be ascribed to

phenotype cancer similarity may result in unconditionally different molecular genotype disease level that occurs in AFT model. This brings up the need of identification for cancer genotypes that are different from each other. As the main leading cause of death among humans is cancer. The timely and accurate prediction and survival analysis becomes much more complex problem in the clinical research over the recent times. Various models based on quantitative observations are developed for the modeling of various survival outcomes and also for multiple variables for explanatory purposes. These models include both fully and semi-parametric models [11]. The main example is Cox proportional hazards model discussed earlier. The Cox proportional hazards model generally makes assumptions based on parameters and certain predictors along with their effects on its hazard function. Also, this model does not make any direct assumptions about the hazard function itself. For most of the problems regarding real world synopsis for Cox proportional hazards model, the true hazard function and its form remains unknown or much complex which is utmost admired and accepted model for cancer survival analysis in clinical research. In clinical research and practice for cancer survival analysis, the survival analysis of patient always depends upon certain low dimensional parameters and characteristics like age, gender, and various other clinical parameters and other histo pathological assessments including stage and grade of disease. The advancements in technology of sequencing throughput provides huge amount of genomic high dimensional data due to which it becomes very easy to discover more molecular bio makers in order to determine and examine survival for improved treatment of patients [12]. Another aspect of research in patient survival analysis is RNA sequencing. This technique is also becoming very popular and useful. The huge degradation in cost of RNA sequencing makes it very feasible to prognosticate. As the cost RNA sequencing decreased greatly, from $ 1OOM per genotype from year 2001 to $1 K only per genotype in year

2015.However, the genomic data in case of RNA sequencing normally contains huge amount of variables. So, to analyze and examine those variables require much more enhanced algorithms for working with high dimensional data. More variables in data means more dimensions to address these challenges while handling high dimensionality of data, different Cox model implementations are also proposed [7]. In Cox model, the complexity penalty is added by regularized model for decreasing the model over fitting likelihood. In recent years, deep learning networks and their increased power for modeling provided much ease in the development of survival analysis models to tackle the issue of high dimensional features of sample data. The examples include auto encoder architecture which is developed for the extraction of features from genomic data for the prediction of liver cancer prognosis [13]. For increased flexibility and modeling strategy, the integration of Cox proportional hazards model and networks is also implemented. For clinical research of specific cancer survival results, one of the main problems is making the available data useful for cancers which are common and then use that particular information to produce more improved and accurate survival predictions [14]. However, this problem can be solved with another form of learning which is transfer learning. In transfer learning, a specific model is initially trained on one task and then it is trained further on a relevant target. The field of transfer learning has greatly improved the accuracy in survival analysis prediction. In combination with transfer learning, deep neural networks can also be used for analysis and examination of biomedical imaging data to grab the benefits from transfer of information from sample data in different environments [15].

1.1 Objectives

In clinical research and survival analysis prediction, one of the fundamental objectives of

research is diagnosing of patient cancer with greater accuracy through the gene expression of patient profiles. In this regard, all the three models Cox proportional hazards models, AFT (Accelerated Failure Time) model and semi supervised learning techniques are used greatly in survival analysis for the both high and low risk classifications along with predictions of patient survival time. However, the there are two main issues that reduces the accuracy of results of all these survival analysis models. One issue is linked with small sized sample data and other is censored data which obstructs the training of these models and affects the accuracy of results and predictions. Also, at the molecular and genotype levels, same phenotype tumors and prognoses are not same diseases. Hence, this limits the use of AFT model to its complete utility and power for survival analysis and prediction in clinical research. This study aims to utilize the semi supervised learning technique with Cox proportional hazard, and AFT model for the survival analysis and prediction by utilizing the Pancreas datasets. The main objectives of the study are also to check the validity and effectiveness of these four models and for that we compare our results with Liang et al [16]. on two datasets, DLBCL (2003) and AML.

Chapter 2

Literature Review

Liang at al. [16] study proposed a novel semi-supervised method based on the COX and AFT models to accurately predict the treatment risk and the survival time of the patients. In this the study shows the semi-supervised learning can significantly improve the predictive performance of COX and AFT models on survival analysis. In accelerated failure time (AFT) model, to increase the available sample size and get more accurate results, each censored observation time is replaced with the imputed value using some estimators. The estimation methods assume that the AFT model was used for the patients with similar phenotype cancer. Semi- supervised learning method has been proved to be effective in solving high-dimensional and small sample size biological data such as protein classification and prediction of interaction between disease and human proteins.

Early diagnosis of cancer could overcome many problems and complications in cancer patients. Systematic methods should be used as early diagnosis methods are of no use in population due to lack of knowledge and other factors involved. Early diagnosis improves quality of life in patients and is cost effective [17]. It prevents from over diagnosis and over prediction process. Technology advancement is introducing many new methods and techniques for the diagnosis purposes which includes many biomarkers, imaging devices, sensors, and other artificial intelligence algorithms. Many medical disciplines such as pathology, pharmacology, radiology and biochemistry frameworks have developed many diagnostic tests at national as well as

international level. Many clinicians require evidence regarding diagnostic purposes and evidence of how much useful and trustworthy is that test and how much accurate diagnostic results will be evaluated. It is also considerable that many results show false positive. Therefore, it is necessary to develop such a test which is very useful and time saving and predicts the disease prognosis along with survival rate in many cancer patients in order to develop a useful treatment planning necessary for saving from many additional adverse reactions in patients suffering from cancer. So, a useful technique is needed. There are many techniques that are based on COX and ATF analysis which can give survival rate prediction ratio in cancer patients [18].

The most leading cause of death throughout the world is cancer which is increasing day by day. Many methods have been developed to predict the survival rate of cancer patients through which better treatment options can be developed [19]. It is predicted that different subtypes are already existing by using prediction techniques. Many techniques are available but these techniques are not useful in case of subtypes of cancer identification. A technique that predicts wide variety of such subtypes is needed in case of no information is available for cancer. Survival time prediction is needed for better treatment planning. In this study [20]a technique was developed to predict the survival time of patients. Class labels were generated using clinical data in this study. Patients were divided into different groups based on their subtype of cancer. Supervised clustering method was used in this regard which is a semi-supervised method. It gave the predictable results in many cancer patients. A subset of genes was used as predictor of survival analysis. Stage of tumor can also be predicted using this method. It was concluded that this method is really useful in diagnosis of many subtypes of cancer. But attention is needed as many tests when used for diagnostic purposes are also false alarms as they may waste time and are also costly. As a result of which, false diagnosis leads to false treatment plans which increases the

complications and adverse reaction already suffering from such disease [20].

In previous studies [21], sample labelling is needed in many supervised models in order to achieve better performance and efficient results specially in case of cancer studies. As cancer is diagnosis- based disease, better diagnosis leads to better treatment and treatment planning which is helpful for both patients and physicians. In many biological data, only a small number of samples are needed for sampling of data. Some samples remain unlabeled, and to label these samples is costly task and is time consuming. Semi-supervised method has significant importance in modelling the data and in case of cancer diagnosis as it is reliable and is cost effective. Many methods have been used for labelling such as semi-supervised or self- learning methods. In this way, performance of the model is improved. There are some problems regarding self-learning model as some manual work may be needed and that it is biased and short sighted. In comparison, semi-supervised method is easy to use and is cost effective. It is also important in disease classification and is accurate method. In this study, in addition to the semi-supervised method another method is also used known as pseudo-labelled sample method. It reduces the false samples. When performed experimentally, it was concluded that in comparison to self-learning and semi-supervised, pseudo-learning method showed better performance. It also was very efficient for gene selection and classification of disease specially in case of cancer. Diagnosis and treatment strategies lie on these techniques. Early diagnosis can help to improve the quality of life of patients and treatment. So, it is necessary to use those treatment and diagnostic plans that are important for both patients and physicians and helpful in clinical research [21].

Diagnosis of cancer is most important and first step in cancer research. It is based on profile of gene expression of the patient. Both ATF model and COX model have been widely used for the

prediction of survival time of low risk or high-risk classification of the patient in this regard. But the diagnosis is not limited only to these steps. But the use of ATF model is limited to unidentified biological differences of this disease in case of phenotype and genotype of the disease. To overcome this, a semi-supervised model was proposed in this study. $L_{1/2}$ regularization genes were adopted in this semi-supervised model. It was predicted that this semi-supervised model can overcome the problems in COX and ATF analysis. This method has successfully been adopted and is being applied in evaluating clinical performance of gene array datasets. There are many advantages of this semi-supervised model such as training samples can be increased from censored data. Survival rate and identification rate from COX analysis is increased. Patient survival time prediction from ATF model is also increased. Bio-markers selection is also strong. In short, this semi supervised model is the best and more important and appropriate tool in order to check the survival rate and ratio in many cancer patients [22].

Selection of genes is a very important task in case of survival of cancer analysis. Supervised learning method can give solution for making future predictions based on labeled data. Weakly labelled data which is also known as censored data is ignored in many cases for model building in cancer patients. In this study [23] , a combined method of COX and ATF methods were used with the censored or weak data. This framework when compared with single COX or single AFT showed better results when used in combination form and also performance was better. But the problem in this method is noise disturbance. This problem was overcome by combining this COX-AFT combination with another technique which is self-placed learning method. This method employs the data more effectively to the censored data as it is self-learning method. It is the most stable and most reliable method so far. It is recently introduced which helps to simulate

the human learning process as a result of which, AFT automatically identifies the samples. As a result of which, noise disturbance is controlled. It has many advantages such as, it promotes the utilization of censored data and improves it in different ways. It also helps to reduce the noise that is produced and that disturbs the data process. When performed experimentally, the advantages were seen over- using single COX and single AFT and improved the survival rate and combination with self-learning process gave best results [23].

A significant challenge in cancer patient survival analysis is to estimate the accurate survival time with low sample size and high dimension genes data set. The most efficient method for treating the cancer is to identify the relevant genes that are associated with the tumor. It also helps in cancer research and its diagnosis which leads to treatment plan. Only cox method was being used for this purpose in the last decade. With some modifications, this method helps in biomarker identification and also risk classification. If the data is not like hazards assumptions, this COX method may not be helpful. For clinical treatment plan, patients' survival time estimation is much important task. So, accelerated failure time model is used if COX model fails. As small sample size remains the problem, so some modifications are done in AFT model along with combining with other models. Beckley-James model has importance in this regard [24]. By using Kaplan-Meier, it estimates the small sized data or also known as censored data. In this paper, Kaplan Meier approach was used in RS-AFT model and small sized data was collected. Prediction model was designed by using ordinary least squares. But the problem with OLS is it is really sensitive to noise produced in the system. Different genes were collected by using this method. By using this approach, a unique pair of genes was selected. It was concluded that RS-AFT model not only collects the genes, it also performs some other approaches such as survival

approach. Accuracy was affected by large number of data set. Some future work is needed as to combine this RS-AFT model with some machine learning approaches. In this way, data set will improve and it will also improve and effect the RS-AFT model in efficacy point of view. So, improvement is needed in this regard [23].

For the classification of cancer, genome profiles can be used. This helps to study the response of many drugs and also patient outcomes and their response. Gene expression study has been very useful in cancer research and is very promising in this regard. But in case of phenotypes, it is not that much helpful. For analysis of regression, COX model is very helpful specially in case of censored data. But this method cannot be applied directly in case of high dimensional predictors. Due to high dimension, some genes expression is also high which creates the problem and collinearity occurs. To deal with this, L2 and L1 are used. It minimizes the negative log and is known as lasso procedure. This lasso helps in selection n of variables. But the problem with this L2 procedure is that it selects all the genes available and the required genes are not expressed. For the prediction of phenotype, only a small number of genes are required. In this study LARS-COX model was compared with other L2 models and it was predicted that other procedure's performance was better than this L2 performance. One advantage of this LARS-COX model is that it selects only required genes and automatically performs the data selection procedure. It is also worth noticing that this method has no limitations over patient's time to clinical data performance and also number of genes that are basic requirements in building the models for clinical evaluation of patient's treatment planning and survival rate. This LARS-COX model can be helpful in building the high risk and low risk patient's groups and also in gene expression. This method can also be used for the selection of many important gene expressions in order to

identify the survival rate in the cancer patients and further studies are needed in this regard [24].

Dey and Mukherjee[25] research shows that lung cancer is a very critical form of the cancer. Patients are at high risk of mortality. Survival rate estimation is an important task in this regard. It is necessary to plan the best treatment options available and to identify the genes that are involved with tumors. Many methods are used to identify the genotype and phenotype of different types of tumors and different models are applied to check the survival rate of the patients in almost all types of cancer patients. Risk assessment and survival rate identification is a necessary task in this regard as it helps in better planning and treatment choices available in this field. The most commonly used methods are COX and AFT methods. Statistical data is analyzed by using these methods and then it is concluded which option is the best. As cox has some limitations like it does not monitor the censored data so it is mostly used with ATF or some other models in order to overcome its limitations. Many models are used such as cubic regression and hazard model. But for censored data analysis, only a valid model is applied. So, in this paper [26], cox regression model along with 5 M-spline functions was introduced. It is more convenient and flexible in a way that it can detect the hazards functions almost all directions such as convex, concave, increasing and decreasing direction. A gene expression data set was used on lungs cancer in order to illustrate the functionality and usefulness of this model. A comparison study was performed on lungs cancer patients between the new technique of this gene expression model and other models. It was concluded and proved statistically that this method is more useful than the existing models. Further studies are needed to compare the effects of this model in other aspects [26].

Urbanska and Sokolowska [27] in their study show that the glioblastoma is a fatal form of brain cancer as its progression is rapid, reoccurrence probability is also high and has resistance to common therapeutics agents. The survival rate of patients is low only 12-15 months and may be five years in rare cases. So, a very powerful diagnostic method is needed to check the prognosis of this disease. Mostly biomarkers are used to identify the prognosis of this brain disease. In this study, a model was proposed that was mostly based on pathways prediction. This model was constructed using the cox model and it was mostly based on L1 model [28]. The succession and risk management of this model was mostly based on the three sets. For the prediction assurance and assessment this pathway model was compared with the gene-based model. In order to improve the prognosis ability of this pathway model, this model was integrated with many different types of clinical features as clinical assessment is needed. For therapeutic interventions, these prognostic improvements are necessary [29]. It was concluded later on that this method is helpful in order to identify the prognosis of glioblastoma in patients. It was also applied clinically. This comparison study was based on this model that is a combination of PDS-based Pathifier (which represent the extent of pathway deregulation based on expression data) and LASSO based cox model. Improvement in prognostic studies was observed when pharmaceutical information was introduced in this study. Application of this prediction model may be helpful in therapeutic management of this brain cancer in near future better treatment options.

Chapter 3

Materials and Methodology

3.1. The Data Source

Three datasets have been used in this study to implement the semi-supervised learning. The three datasets used in this study are DLBCL2003 dataset [30], AML dataset [31] and Pancreas dataset [32]. In this section the three datasets will be introduced and summary statistics for them will be presented.

3.1.1. DLBCL (2003)

This dataset includes a sample of 92 patients with 8810 gene expressions from each patient, 28 out of 92 patients are censored observations which is 30.43% of the total observations. The survival time of the 69.56% is completely recorded. For the 30.43% of the censored observations the survival time is not recorded completely, since they were still alive at the time of releasing the data. The term DLBCL refers to “Diffuse Large B-Cell Lymphoma”. The patients in this dataset are those which have diffuse large B-cell lymphoma and passed the chemotherapy and

monitoring schedules.

In Table 3.1, five patients of the DLBCL (2003) data with the gene expressions ID and 26 gene expressions are presented. Some gene expression values are missing and the missing gene expressions are shown by NA in this dataset.

Table 3.1. Gene expressions of DLBCL2003 data

UNIQID	MCL_CyclinD1pos_Lym610	MCL_CyclinD1pos_Lym613	MCL_CyclinD1pos_Lym623	MCL_CyclinD1pos_Lym626	MCL_CyclinD1pos_Lym632
15841	0.2041	-0.0516	0.6930	-0.3406	0.6152
15842	NA	NA	0.0768	0.2962	0.2650
15843	-0.4479	-0.3005	-0.2826	-0.2326	0.4723
15844	NA	0.1456	0.0074	0.0615	0.3314
15845	-0.7302	NA	-0.1943	-0.2962	NA
15846	NA	0.2951	-0.2771	-0.5173	-0.0513
15847	0.3546	0.0710	0.5275	-0.4036	-0.4946
15848	-0.7628	0.3365	-0.0274	-0.0298	-0.1122
15849	-0.1092	0.5600	-0.1611	-0.3234	0.1137
15850	1.0211	-0.3402	0.1230	-1.1193	-0.7158
15851	-0.1817	0.8957	-0.7491	-1.1271	-0.5438
15852	-1.1804	0.5062	0.9862	-0.3318	-0.3010
15853	0.3606	0.2161	0.4494	0.0705	-0.3226

15855	-0.1210	0.0071	0.1403	-0.0051	-0.2874
15856	0.2673	NA	1.2121	0.4433	NA
15857	-0.3175	-0.3203	0.1079	NA	0.0625
15858	0.9995	0.9252	0.0160	0.4551	0.3352
15859	-0.7528	0.2199	0.2414	-0.3157	-0.2318
15861	-0.2363	0.2148	0.2146	-0.4433	-0.2390
15862	-0.4582	0.0443	0.1936	-0.3567	-0.3582
15863	0.0329	-0.2515	0.0785	0.2121	-0.4311
15864	-5.0000	NA	1.2097	0.2470	-0.1838
15866	-0.4858	0.6537	0.1689	-0.4052	-0.0934
15867	-0.1944	-1.0435	0.5175	NA	-0.1286
15868	-0.9015	-0.1035	1.0374	0.1113	0.5035
15869	0.0638	-0.0431	0.4046	-0.4233	-0.1077

The descriptive statistics for the gene expression are presented in Table 3.2. The five-number summary, minimum, 1st quartile, median, 3rd quartile, maximum and the mean of gene expression values for 7 patients are reported.

Table 3.2. Descriptive statistics of 7 patients in DLBCL2003 data

statistics	MCL_CyclinD1pos_Lym610	MCL_CyclinD1pos_Lym613	MCL_CyclinD1pos_Lym623	MCL_CyclinD1pos_Lym626	MCL_CyclinD1pos_Lym632	MCL_CyclinD1pos_Lym634	MCL_CyclinD1pos_Lym644

Min	-5.0	-5.0	-5.0	-2.6895	-5.0	-2.46311	-5.0
1st Qu.	-0.4485	-0.2826	-0.2115	-0.3747	-0.3466	-0.26481	-0.4035
Median	-0.1366	0.0308	0.0735	-0.1146	-0.0870	-0.05868	-0.1120
Mean	-0.1625	-0.0060	0.0796	-0.0839	-0.0777	-0.05641	-0.1369
3rd Qu.	0.1540	0.3074	0.3652	0.1845	0.1805	0.15884	0.1568
Max	5.0	3.1699	5.0	2.8251	5.0	2.86387	3.3094
NA's	1424	527	356	667	1111	289	930

As we can see in the descriptive statistics, for each patient several of gene expression values are missing. For example, 1111 gene expressions of "MCL_CyclinD1pos_Lym632" are missing. The gene expressions are between -5 to 5 for all 8810 gene expressions.

DLBCL 2003 dataset also includes survival times of the patients and status of the patients, namely censored (status = 0) or uncensored (status = 1). In Table 3.3, the survival time, status, BMI (Body Mass Index), proliferation average and 4 other parameters in the DLBCL 2003 data can be seen. The dataset is presented for the 26 patients in the dataset. Time of Follow-up column represents the survival time, status at follow up represents the status (censoring = 0 and dead = 1).

Table 3.3. Survival time, status and other parameters for 26 patients of DLBCL2003

Array ID	status at follow up	Time of Follow-up	INK/ARF deletion	ATM deletion	P-53 deletion	CyclinD-1 taqman	BMI expression	Proliferation.average

						results		
MCL_CyclinD1pos_Lym610	1	0.7529	0	0	1	0.381	-0.4049	-0.1230
MCL_CyclinD1pos_Lym613	1	3.2772	0	0	0	0.656	0.2125	-0.3192
MCL_CyclinD1pos_Lym623	1	2.1218	1	1	1	0.532	-0.3697	0.4747
MCL_CyclinD1pos_Lym626	1	14.0534	0	0	0	0.67	-0.0886	-1.1303
MCL_CyclinD1pos_Lym632	1	3.2361	0	0	0	0.505	-0.2735	-0.4045
MCL_CyclinD1pos_Lym634	1	4.4873	0	0	0	0.522	0.0168	-0.1024
MCL_CyclinD1pos_Lym644	0	0.7778	0	0	0	0.618	0.1034	-1.1241
MCL_CyclinD1pos_Lym620	1	0.4298	0	0	0	0.233	-0.3466	0.0061
MCL_CyclinD1pos_Lym625	1	1.0568	0	1	0	0.737	-0.1871	0.1799
MCL_CyclinD1pos_Lym631	1	3.2882	NA	NA	NA	0.632	-0.3488	-0.0260
MCL_CyclinD1pos_Lym635	0	6.8966	NA	NA	NA	0.477	-0.6770	-0.3544
MCL_CyclinD1pos_Lym653	1	0.2656	0	0	0	0.87	0.0120	0.4414
MCL_CyclinD1pos_Lym666	0	0.5503	0	1	0	0.833	0.1422	-0.4077
MCL_CyclinD1pos_Lym600	0	0.7529	0	0	0	0.462	-0.3483	0.2824
MCL_CyclinD1pos_Lym601	1	9.2320	0	0	0	0.672	0.8919	-0.3334
MCL_CyclinD1pos_Lym503	0	1.3963	0	0	0	0.551	0.0414	-0.4091
MCL_CyclinD1pos_Lym504	0	1.4045	1	1	0	0.552	0.0517	-0.3129
MCL_CyclinD1pos_Lym501	0	2.3135	0	0	0	0.305	-0.2512	-0.3818
MCL_CyclinD1pos_Lym508	1	0.8049	1	1	0	1.806	-0.3007	0.5533
MCL_CyclinD1pos_Lym514	0	7.2334	NA	NA	NA	0.95	-0.2029	-0.5593
MCL_CyclinD1pos_Lym516	1	8.4654	0	0	0	0.251	-0.1695	-0.0395
MCL_CyclinD1pos_Lym519	1	1.6235	0	1	1	0.604	-0.3567	0.4697

MCL_CyclinD1pos_Lym520	0	2.7981	0	0	0	0.185	-0.2280	-0.1833
MCL_CyclinD1pos_Lym521	1	0.8898	0	0	0	0.57	0.2297	-0.1786
MCL_CyclinD1pos_Lym522	1	1.0705	0	0	0	0.251	-0.3022	0.4519

The descriptive statistics for the second dataset of DLBCL 2003 is presented in Table 3.4. For the discrete variables in Table 3.4. Frequency and fraction of each category of "0" and "1" are reported. For the continuous parameters the minimum, maximum, 1st quartile, 3rd quartile, median and mean of the variable for 92 patients are reported. For 3 discrete variables (INK/ARF deletion, ATM deletion, P-53 deletion), there were 3 missing values. For the continuous parameters there were no missing value in the dataset.

Table 3.4. Descriptive statistics of survival data of DLBCL2003

Statistics	status at follow up	Time of Follow-up	INK/ARF deletion	ATM deletion	P-53 deletion	CyclinD-1 taqman results	BMI expression	Proliferation. average
Min	0: 28	0.0191	0: 67	0: 55	0: 76	0.1850	-0.9288	-2.0384
1st Qu.	1: 64	0.8268	1: 18	1: 30	1: 9	0.4725	-0.3483	-0.3820
Median		1.9644	NA: 7	NA: 7	NA: 7	0.6050	-0.0857	-0.05904
Mean		2.7624				0.6635	0.0	0.0023
3rd Qu.		3.2799				0.7550	0.1767	0.43137

Max		14.0533				2.1670	2.7180	1.7565
Percentage 0	30.43%		72.82%	59.78%	82.61 %			
Percentage 1	69.57%		19.57%	32.61%	9.78%			

The first dataset has size 8810x101. The first two columns are UNIQID and name, the rest of columns are the patients IDs. There are 99 patients in the first dataset. The second data set has the size 92x9. After checking both datasets, it was seen that in the first dataset, columns 95 to 101 are patients IDs which are not included in the second dataset. So, the patients from columns 95 to 101 are removed from the first dataset. For the first 92 patients in both datasets the order of the patients is same. After getting columns 3 to 94 in the first dataset and transposing it will give the size 92x8810. The second dataset has size 92x9. But we need only two columns of second dataset (status and time of follow-up). Two columns of second dataset were merged with the first dataset. The size of the merged data is 92 x 8812, where 8810 columns are for gene expressions, one column for survival time and one column for status. The merged data is ready for survival analysis.

3.1.2. AML Dataset

The AML dataset [31], includes the gene expression and survival times for 119 patients. This data includes 2828 gene expressions for 119 patients. The survival dataset of AML includes survival time and status for 116 patients. The survival time of 68 out of 116 patients in AML data are uncensored which is 58.62% and 48 of them are censored (41.38%). In Table 3.5 the heading of the gene expression dataset of AML is presented for 6 patients. The first column refers to the ID of gene expression. The other column names are the ID of the patients.

Table 3.5. Gene expression of AML dataset

CLID	AML 106	AML 52	AML 66	AML 49	AML 80	AML 107
143654	-0.9343	-0.07201	1.317	NA	-0.4999	0.07302
433257	-0.2151	0.8321	1.329	NA	-0.06238	0.1786
429093	-1.721	NA	1.153	0.4126	-0.3687	0.2062
878212	0.5484	0.1317	-0.6551	NA	-1.064	-0.9097
143759	1.309	-0.6348	0.6642	NA	0.1566	1.003
144762	0.8102	NA	0.1133	NA	NA	-0.1583
177737	-3.412	0.119	1.11	NA	-0.09959	0.4633
815279	1.785	0.6667	-0.3679	0.6748	0.3885	-0.2955
625693	1.281	0.4298	0.5214	NA	1.105	0.8707
897742	0.6619	NA	0.02898	NA	1.235	0.2144
40946	0.3963	1.179	0.01223	NA	0.9547	-0.008389
289936	-0.4229	2.776	0.7156	NA	-0.007939	NA
810603	0.6831	NA	-0.1836	NA	0.1879	0.3718

208969	NA	NA	0.3	NA	0.07949	-0.08857
381107	NA	NA	-0.4002	NA	-0.1977	0.7912
451546	1.392	NA	0.08459	1.672	2.251	1.756
448676	0.3637	0.9499	-0.7208	NA	-1.033	0.1796
796468	-2.858	NA	0.574	NA	-0.0716	0.7713
1521297	NA	NA	0.4187	NA	-0.8678	1.959
565826	NA	NA	0.5916	NA	0.314	NA
190325	-2.468	1.819	-0.6255	NA	-1.608	1.228
324815	-0.00281	0.009428	0.1116	NA	0.2351	-0.234
429642	NA	NA	-0.8599	NA	0.5175	0.8445
839101	-0.322	2.505	-0.7447	NA	0.2477	0.1996
725364	0.1163	-0.9114	0.8573	0.3619	-0.01525	-0.5343

The size of AML gene expression data is 2828x121. The first two columns are the IDs and name of the gene expressions (name of gene expression column is not shown in the above table). So, 119 patients are in the gene expression data. The survival data includes 116 patients as mentioned earlier. After checking both datasets, it was seen that 5 patients which are in columns {3, 19, 82, 120, 121} in the first dataset are not included in the second dataset which has survival times. So, these 5 patients in the first dataset were discarded. After that both datasets were ordered by patient's IDs. The gene expression data were transposed and then concatenated with the survival data. So, the merged data has the size 116x2830 (2828 gene expressions, 1 column for status and one column for survival time). Table 3.6 shows the five-number summary, minimum, 1st quartile, median, 3rd quartile, maximum and mean of the gene expressions. The

number of missing gene expressions for 7 patients is also shown in the last row.

Table 3.6. Gene expressions descriptive statistics of AML dataset for 7 patients

statistics	AML 106	AML 52	AML 66	AML 49	AML 80	AML 107	AML 87
Min	-6.2680	-7.6010	-4.7330	-5.2060	-5.3290	-3.4830	-5.7480
1st Qu.	-1.0290	-1.5095	-0.4991	-1.1325	-0.6046	-0.6357	-0.3916
Median	-0.1479	-0.3253	0.1401	-0.2818	-0.0222	-0.0512	0.2984
Mean	-0.2030	-0.4509	0.0665	-0.3093	-0.0421	-0.0830	0.2515
3rd Qu.	0.7236	0.6778	0.7178	0.4690	0.5730	0.5102	0.9354
Max	7.3560	4.9540	4.0380	4.4260	3.8050	4.1820	4.4310
NA’s	423	557	28	1794	27	317	870

In Table 3.6, “AML 49” shows 1794 gene expressions having missing values out of 2828 gene expressions. This dataset includes many missing values for some patients. The patients with many missing values need to be removed before getting the non-missing data for the survival analysis, otherwise all the gene expressions will be removed when getting the non-missing data.

The two columns of survival time and status of AML data can be seen in Table 3.7.

Table 3.7. Survival data of AML

Sample	Sex	Age (years)	WBC (x1000/ ul)	PB-Blasts (%)	BM-Blasts (%)	LDH (U/l)	Status	Overall Survival (days)	Training/Test Set?
AML 1	F	39.7	14.1	76	88	599	Dead	213	test set

AML 2	M	34.5	36.1	86	95	365	Alive	801	training set
AML 4	M	40.9	130.3	NA	79	3890	Dead	210	training set
AML 5	M	38.7	37.1	93	95	524	Dead	243	training set
AML 6	M	63.3	29.9	73	81	343	Dead	336	training set
AML 7	F	37.6	7.3	80	90	386	Dead	134	training set
AML 8	M	68.9	69.7	36	70	648	Alive	206	test set
AML 9	M	33.7	57.1	60	95	1011	Alive	438	test set
AML 10	M	34.1	177.5	90	NA	1621	Dead	233	test set
AML 11	F	46.7	65	NA	84	1109	Dead	204	test set
AML 12	M	31.2	112	NA	95	731	Alive	610	test set
AML 13	F	73.4	81	87	80	494	Dead	85	training set
AML 14	M	35.3	24.5	77	90	576	Dead	711	test set
AML 15	M	62.6	112	66	90	724	Dead	483	test set
AML 16	F	49.1	13	30	51	249	Alive	610	training set
AML 17	M	46.3	17.2	79	90	559	dead	570	test set
AML 18	M	28.1	26.2	76	90	409	dead	323	training set
AML 20	M	65	31.9	78	100	1336	alive	884	training set
AML 21	M	23.2	81.3	84	85	410	dead	21	test set
AML 22	M	52.4	60.4	56	56	735	dead	154	training set
AML 23	M	73	57.5	57	90	276	dead	28	training set
AML 24	F	41.9	91.2	96	80	883	dead	35	test set
AML 25	M	67.1	21.9	58	50	420	alive	400	test set
AML 26	M	71.8	11.6	59	78	574	alive	138	training set

AML 27	M	47.4	54.6	64	98	817	dead	326	training set

The summary statistics for AML data can be seen in Table 3.8, which includes five number summary, minimum, 1st quartile, median, 3rd quartile, maximum and mean. 42.24% of the patients are female and 57.76% of them are male. The patient's Age range is from 19.1 to 74.6 years. The survival time is reported in days. Half of the patients have survival time more than 333.5 days (approx. 1 year) and half of them have survival less than 1 year.

Table 3.8. Summary statistics for Survival data of AML

Statistics	Sex	Age (years)	WBC (x1000/ ul)	PB-Blasts (%)	BM-Blasts (%)	LDH (U/l)	Status	Overall Survival (days)	Training/Test Set?
Min	F:49 M:67	19.1	1.10	1	30	93	Dead: 68 Alive: 48	0.0	test set: 57 trainingset: 59
1st Qu.		38.58	21.20	52	71.50	397.5		145.0	
Median		49.85	37.90	72	83	599		333.5	
Mean		50.83	56.85	66.77	78.51	803.9		422.7	
3rd Qu.		63.42	74.70	84	90	917.5		610.2	
Max		74.60	427	100	100	3953		1625.0	

NA’s		4	3	9	16	5		206	

3.1.3. Pancreas Dataset

The Pancreas dataset [32], includes the gene expression and survival times for 90 patients. This data includes 28869 gene expressions for 90 patients. The survival dataset of Pancreas includes survival time and status for 90 patients. The survival time of 58 out of 90 patients (64.4%) is uncensored and 26 (28.8%) of them are censored and status for 6 patients are missing (NA). In Table 3.9 the heading of the gene expression dataset of Pancreas data is presented for 7 patients. The first column is for the ID of gene expressions. The other column names are the ID of the patients.

Table 3.9. Gene expressions for Pancreas dataset

ID_REF	GSM711904	GSM711905	GSM711906	GSM711907	GSM711908	GSM711909
7896736	3.30105	4.31783	2.98749	3.5481	2.97383	3.51183
7896738	1.85066	1.8463	2.14159	1.83839	2.37636	1.80252
7896740	2.02519	2.11336	2.15384	1.88753	2.22671	1.88364
7896742	4.47212	5.26114	4.45806	5.72147	5.18768	4.73796
7896744	4.33654	4.48297	4.44503	4.6213	4.09481	4.18046
7896746	7.11794	8.42338	6.24828	6.79179	6.96976	6.20861
7896748	8.14456	8.73852	7.04447	7.03643	7.28209	7.32149

7896750	2.45761	3.54018	2.02138	2.38217	1.83525	1.79839
7896752	6.78099	7.2457	5.94896	5.88869	6.02336	5.33305
7896754	5.69052	6.11363	6.49958	5.04972	4.9541	5.83686
7896756	2.57463	3.25761	2.85148	2.68036	2.67713	2.9894
7896759	4.6751	4.58873	4.41072	4.62062	4.08564	5.23173
7896761	4.67853	4.93868	4.14657	4.86673	4.63776	4.63271
7896779	4.69027	5.65066	4.50531	4.78355	4.749	4.61338
7896798	4.80957	5.42863	4.6985	4.94571	4.83227	4.68613
7896817	4.26581	4.21733	5.16008	4.16156	3.98531	3.96813
7896822	5.13505	4.83462	7.15667	5.54606	6.21325	4.73643
7896859	4.19797	4.08344	3.94949	4.33324	4.19142	4.0743
7896861	2.24616	2.21034	2.34461	2.46391	2.46553	2.26239
7896863	3.99632	4.92952	4.24649	4.76474	4.22626	4.05603
7896865	3.94536	4.92105	4.01757	4.39553	4.30383	4.2102
7896878	4.91523	6.01246	4.93032	5.29012	5.1706	4.97211
7896882	3.81208	4.20901	4.05487	4.24003	3.90322	3.79108
7896908	4.09896	4.26998	4.29329	4.11288	3.89184	3.9135
7896917	3.98895	4.35853	4.37263	4.07403	4.32297	4.06566

The descriptive statistics of 7 patients of Pancreas dataset for the gene expressions can be seen in Table 3.10.

Table 3.10. Summary statistics for gene expressions of Pancreas dataset

Statistics	GSM711904	GSM711905	GSM711906	GSM711907	GSM711908	GSM711909
Min	0.9905	0.9415	0.987	1.043	0.9896	0.9108
1st Qu.	2.8596	3.0745	2.80	2.996	2.8882	2.8843
Median	4.1728	4.2428	4.088	4.245	4.149	4.1707
Mean	4.2906	4.3130	4.282	4.301	4.30	4.288
3rd Qu.	5.4947	5.3353	5.530	5.414	5.4895	5.4778
Max	12.7306	12.8659	12.875	12.813	12.8603	12.7028

As we can see in the above table, the gene expression values for the 7 patients have almost the same range. The gene expression ranges for the rest of the patients are similar to the values in the table for 7 patients. The survival time and status of the patients with Pancreas cancer are presented in Table 3.12. The survival time and status of some patients are missing (NA). The status (1) is for uncensored or dead observations and status (0) is for observations which are censored or alive. The survival time unit for pancreas data is in months.

Table 3.11. Survival time and status of patients for pancreas dataset

Patient ID	Survival time (Months)	Status (cancer_death)
GSM711904	51	1
GSM711905	51	1
GSM711906	7	1
GSM711907	7	1
GSM711908	3	1

GSM711909	3	1
GSM711910	42	1
GSM711911	42	1
GSM711912	NA	NA
GSM711913	NA	NA
GSM711914	36	1
GSM711915	36	1
GSM711916	2	1
GSM711917	2	1
GSM711918	NA	NA
GSM711919	NA	NA
GSM711920	NA	NA
GSM711921	NA	NA
GSM711922	19	1
GSM711923	19	1
GSM711924	13	1
GSM711925	13	1
GSM711926	16	1
GSM711927	16	1
GSM711928	41	1
GSM711929	41	1

The summary statistics for Pancreas data is presented in Table 3.12.

Table 3.12. Summary statistics for Survival data of pancreas dataset

Patient ID	Survival time (Months)	Status (cancer_death)
Min	1.0	0: 26 (28.89%) 1: 58 (64.44%)
1st Qu.	8.0	
Median	14.5	
Mean	17.4	
3rd Qu.	24.0	
Max	51.0	
NA's	6	6 (6.67%)

As we can see in the above table, survival time of 6 patients is missing. So, after discarding the gene expression of these 6 patients and transposing gene expression data, the gene expression data and survival data are merged together to be used for survival analysis. After removing the data for 6 patients, there are 84 patients. From 84 patients 26 (30.95%) are censored and 58 (69.05%) are uncensored.

3.1.4. Simulating survival data

For simulating correlated and uncorrelated survival data and gene expressions, method proposed

by Bender et al. [33] has been used. The gene expressions were simulated by using randomly generated values from normal distribution, which are then related by correlation zero (uncorrelated parameters) or 0.3 (correlated parameters). Simulating process [16] is explained below:

Step 1. Randomly generating 1000 parameters with different sample sizes n = 100, 200, 300 using standard normal distribution. The generated parameters are $\gamma_{i0}, \gamma_{i1}, \ldots, \gamma_{ip}$, for i = (1,2,3,…,n) and p = 1000.

Step 2. Setting 10 coefficients equal to 1 and the rest 990 coefficients from 1000 coefficients equal zero. The 10 parameters which have coefficient equal 1 are prognostic genes.

Step 3. Using correlation coefficients $\rho = 0$ and $\rho = 0.3$, the standard normal parameters are correlated by using: $X_{ij} = \gamma_{ij}\sqrt{1-\rho} + \gamma_{i0}\sqrt{\rho}$. These generated parameters are the simulated gene expressions.

Step 4. For prognostic genes, we need to relate the survival outcome with the generated gene expressions. Using simulated coefficients and generated gene expression, the survival time is simulated by using $Y_i = 1 - \frac{\alpha \log(U)}{\omega e^{\beta X}}$. In this formula α and ω are shape and scale parameters of Gompertz distribution and U is a random uniform distribution with values between 0 and 1.

Step 5. 40% of observations of survival times were replaced by random numbers and considered as censored.

Data was simulated with sample size n = 100, 200 and 300, number of parameters p = 1000 using correlated ($\rho = 0.3$) and uncorrelated ($\rho = 0$) parameters.

3.2. Methodology

The methodology used for survival analysis includes single cox proportional hazard model with $L_{1/2}$ regularization, single accelerated failure time (AFT) model with $L_{1/2}$ regularization, semi-supervised cox proportional hazard model and semi-supervised accelerated failure time (AFT) model [16]. These four models have been employed for survival analysis of correlated and uncorrelated simulated datasets and three real datasets which were introduced in Section 3.1.

3.2.1. Cox proportional hazard with $L_{1/2}$ regularization

The Cox-proportional Hazard model [34] is a regression model which relates the hazard ratio with the independent variables by a regression model. In Cox-model the assumption is that the Hazard depends on the time but the ratio of the Hazard on the Baseline hazard is not time dependent. The formula for Cox-proportional hazard model is presented in equation (1).

$$\lambda(t\,|X_i) = \lambda_0(t)exp(X_i\beta) \quad (1)$$

Where $\lambda(t\,|X_i)$ is the conditional hazard at time t gives the covariates X_i , $\lambda_0(t)$ is the baseline hazard at time t and $exp(X_i\beta)$ is the exponential of the multiplication of covariates with the coefficients. The Hazard ratio will be independent from the time and will be just related with the exponential of multiplication of covariates and coefficients. The log of hazard ratio can be written as a linear equation presented in equation (2).

$$\log\left(\frac{\lambda(t\,|X_i)}{\lambda_0(t)}\right) = X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{ip}\beta_p \quad (2)$$

In equation {2}, index i shows the hazard ratio for the ith observation. Index of β from (1 to p)

shows the coefficients for p covariates.

The objective in the cox proportional hazard model is to calculate the coefficients. So that the hazard ratio can be calculated for each observation by using the calculated coefficients. The coefficients can be calculated by maximizing the partial likelihood of the cox model.

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{exp(X_i \beta)}{\sum_{r \in R(t_i)} exp(X_r \beta)} \right\}^{s_i} \quad (3)$$

Where $R(t_i)$ is the risk set at each time t_i. t_i is the survival time of the ith observation. In each time t_i all the observations which are still alive and the one which is dead or is censored at t_i are the observations which are at risk. So the risk set $R(t_i)$ contains all the observation t_r which have $t_r \geq t_i$. n is total observations in the sample. s_i shows the status of the ith observation whether it is uncensored (1) or censored (0).

The coefficients can be estimated by calculating the maximum of the partial log-likelihood or minimizing the negative partial log-likelihood. The partial log-likelihood of the cox model can be written as equation (4).

$$l(\beta) = \sum_{i=1}^{n} S_i \{ X_i \beta - \log \left\{ \sum_{r \in R(t_i)} exp(X_r \beta) \right\} \} \quad (4)$$

When there are many covariates in the model (number of parameters is much more than number of observations: p >> n) like the survival study using real data. In clinical study, normally there are many gene expressions and low number of observations due to cost of clinical experiments. So, the survival study deals with the high dimensional data. In such cases regularization parameter is used to reduce the dimension of the data and keeping only the gene expressions

which have notable effect on the hazard rate. The regularization parameter shrinks the coefficients somehow to set those which are not notable to zero. Cox proportional hazard model with $L_{1/2}$ regularization can be solved by minimizing the equation (5).

$$\beta = argmin\{-l(\beta) + \lambda \sum_{j=1}^{p} |\beta_j|^{\frac{1}{2}}\} \quad (5)$$

Where $-l(\beta)$ is the negative partial log-likelihood of the cox model from equation (4), λ is the tuning parameter. Minimizing the equation (5), the coefficients can be estimated and it will be clear which observations were set to zero by using $L_{1/2}$ Regularization.

To solve the cox proportional hazard model with $L_{1/2}$ regularization, a modified Newton Raphson method with coordinate descent algorithm has been proposed by Cheng et al [8].

Flow chart of the algorithm for estimating the coefficients in cox proportional hazard with $L_{1/2}$ penalty is presented in Figure 3.1.

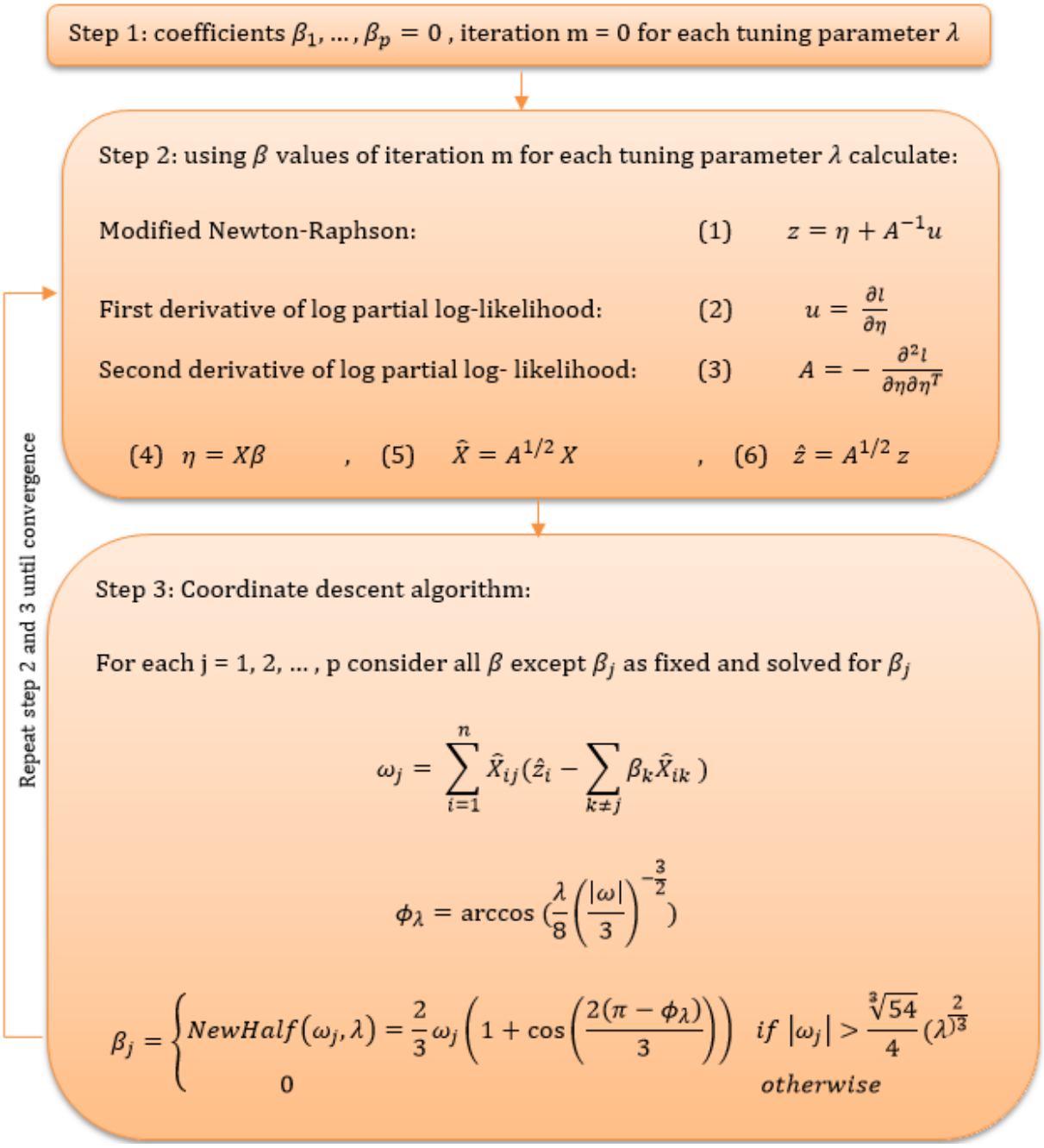

Figure 3.1. Flowchart of parameter estimation of cox-proportional hazard model with $L_{1/2}$

3.2.2. Accelerated failure time (AFT) with $L_{1/2}$ regularization

The Accelerated failure time (AFT) model [4], represents the relationship between the log of survival time with the covariates linearly. The equation (6) shows the linear relationship between survival time and covariates in log-linear AFT model.

$$\log(t_i) = X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{ip}\beta_p + \epsilon_i \quad (6)$$

Where t_i is the survival time of the ith observation, ϵ_i is the residual of the model. In log-linear model the residual is assumed to be normally distributed. The AFT model can simply be solved by using ordinary least square method in absence of the censoring. But most of the survival data and also the data which we have used in this study have censoring events. To make the analysis easier and be able to use least square method, in previous study [16] have proposed to use mean imputation to impute the censoring observations (mean imputation method is explained later in this section). Xu et al. [34] iteratively reweighted least square method with $L_{1/2}$ regularization was used for the solving the loss function of the AFT model. The AFT model with $L_{1/2}$ regularization can be solved by minimizing the loss function presented in equation (7).

$$\beta = argmin\,\{\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 \;\; + \lambda \;\; \sum_{j=1}^{p}\left|\beta_j\right|^{\frac{1}{2}}\} \quad (7)$$

Where $\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2$ is the sum of squared errors, λ is the tuning parameter and $\sum_{j=1}^{p}\left|\beta_j\right|^{\frac{1}{2}}$ represents the penalty term for $L_{1/2}$ regularization. The coefficients can be estimated by minimizing the equation (7).

To minimize the loss function using iterative reweighted least squares method, coordinate

descent algorithm can be used with New Half thresholding. Previously they have used Y_i which is the log of survival time in AFT model instead of $\hat{z}$ (instead of $\hat{z}$ in the equations of modified newton raphson for cox model, in this algorithm Y_i has been used). Hence, in AFT model the solution can be done by using log of survival time and updating the coefficients.

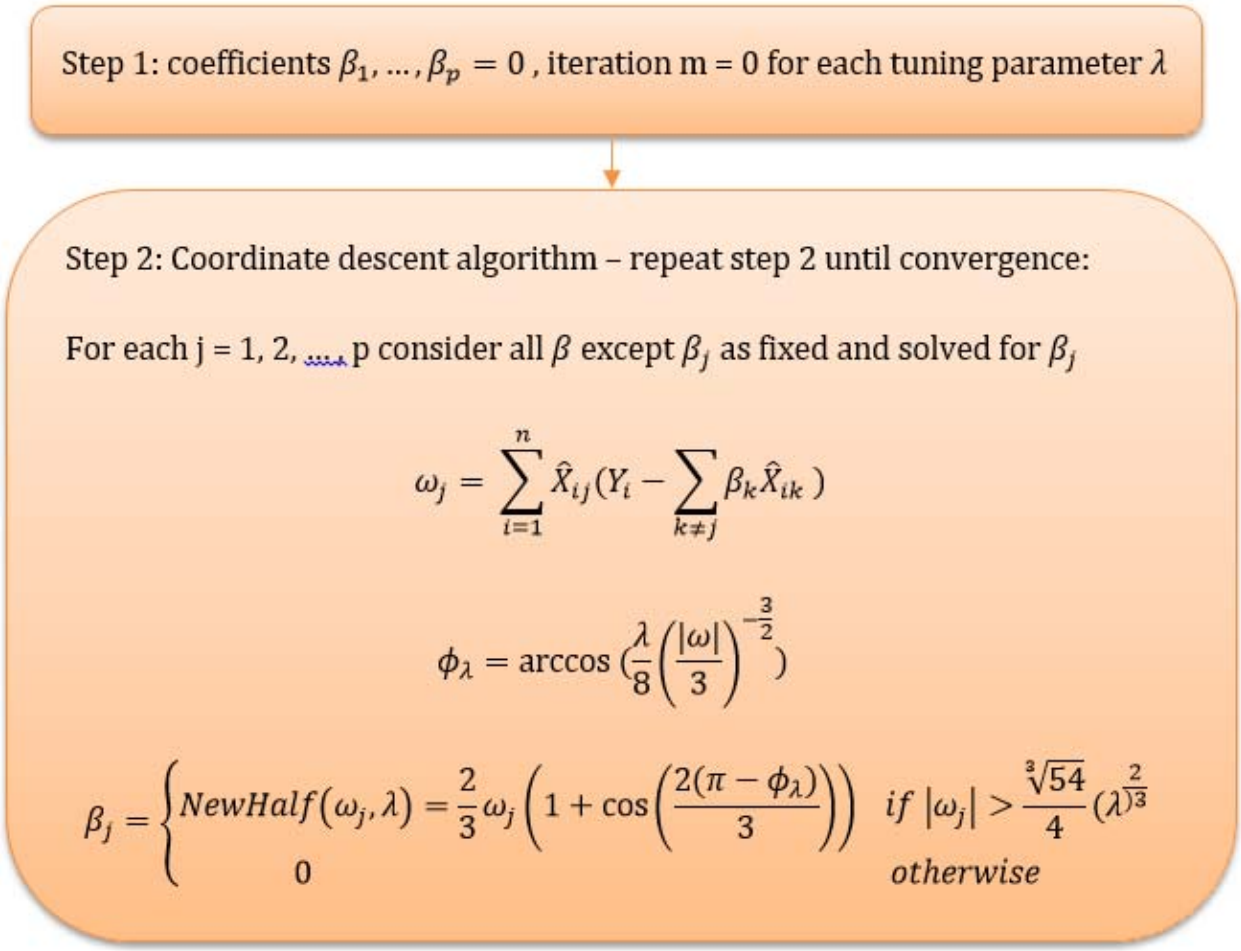

Figure 3.2. Flowchart of parameter estimation of Accelerated failure time model with $L_{1/2}$ regularization

AFT model can be solved by using the modified Newton Raphson algorithm shown in Figure 3.2. Instead of loss function of partial log-likelihood in that algorithm, we need to use likelihood function of normally distributed error term.

3.2.2.1. Mean Imputation method

In this method the censoring observation was imputed by using the survival function which was derived with the non-parametric Kaplan-Meier estimator [36] for the empirical distribution of the survival. The mean imputation method imputes the values of censored observations with the equation (8).

$$h(t_i^*) = \delta_i h(t_i) + (1 - \delta_i)\{\hat{S}\}^{-1} \sum_{t_{(r)} > t_i} h(t_{(r)}) \Delta\hat{S}(t_{(r)}) \qquad (8)$$

Where for each observation, δ_i is the status (1) uncensored and (0) censored, $h(t_i)$ is the log of survival time for the ith observation. If the survival time is uncensored, it will remain same using equation (8) and it will be imputed when the survival time is censored by the sum of multiplication of all log survival time $h(t_{(r)})$ in which $t_{(r)}$ are more than t_i with the difference in survival function at time step $t_{(r)}$. So for all $t_{(r)} > t_i$ the summation of $h(t_{(r)})\Delta\hat{S}(t_{(r)})$ multiplied by inverse of survival function of Kaplan-Meier estimator $\{\hat{S}\}^{-1}$ will be calculated and imputed instead of censored log survival time.

3.2.3. Semi-Supervised Cox-proportional hazard model

Liang et al. [16] have proposed the semi-supervised learning algorithm which can be performed as semi-supervised Cox and semi-supervised AFT. The semi-supervised Cox model process is described in this section. After preprocessing the gene expression data by making the parameter X to have zero mean and standard deviation equal to 1 and getting rid of the missing values, the data set is ready for the survival analysis. The data splits into completed data or non-censored data and censored data. For the completed data, Cox proportional hazard model with $L_{1/2}$ penalty was fitted with 3 folds cross validation. Then the best tuning parameter λ is selected as the one which has the highest concordance index in the testing data. The selected parameters and the coefficients of the cox model is determined. Then based on the prediction of the cox model the risk ratio which is more than 1 is considered as high risk and those risk ratios less than 1 is considered as low risk. Hence the dataset is classified into low risk and high risk. Then mean imputation method will be used to impute the censored observations. Those imputed values are accepted which has $\mathrm{h}(\mathrm{t}_\mathrm{i}^*) \geq h(t_i)$, otherwise the imputed value is not accepted. For mean imputation the Kaplan-Meier estimator for low risk and high risk data is derived separately. So, the high risk observations will be imputed based on survival function of high risk observations and the low risk observations will be imputed based on the survival function of low risk observations. Then the final cox proportional hazard model is fitted with $L_{1/2}$ penalty using both censoring and uncensored data. Then replacing $\mathrm{h}(\mathrm{t}_\mathrm{i})$ and δ_i with the imputed value the next iteration will be done. This operation is done for M iterations. The flowchart of the semi-supervised Cox model is shown in Figure 3.3

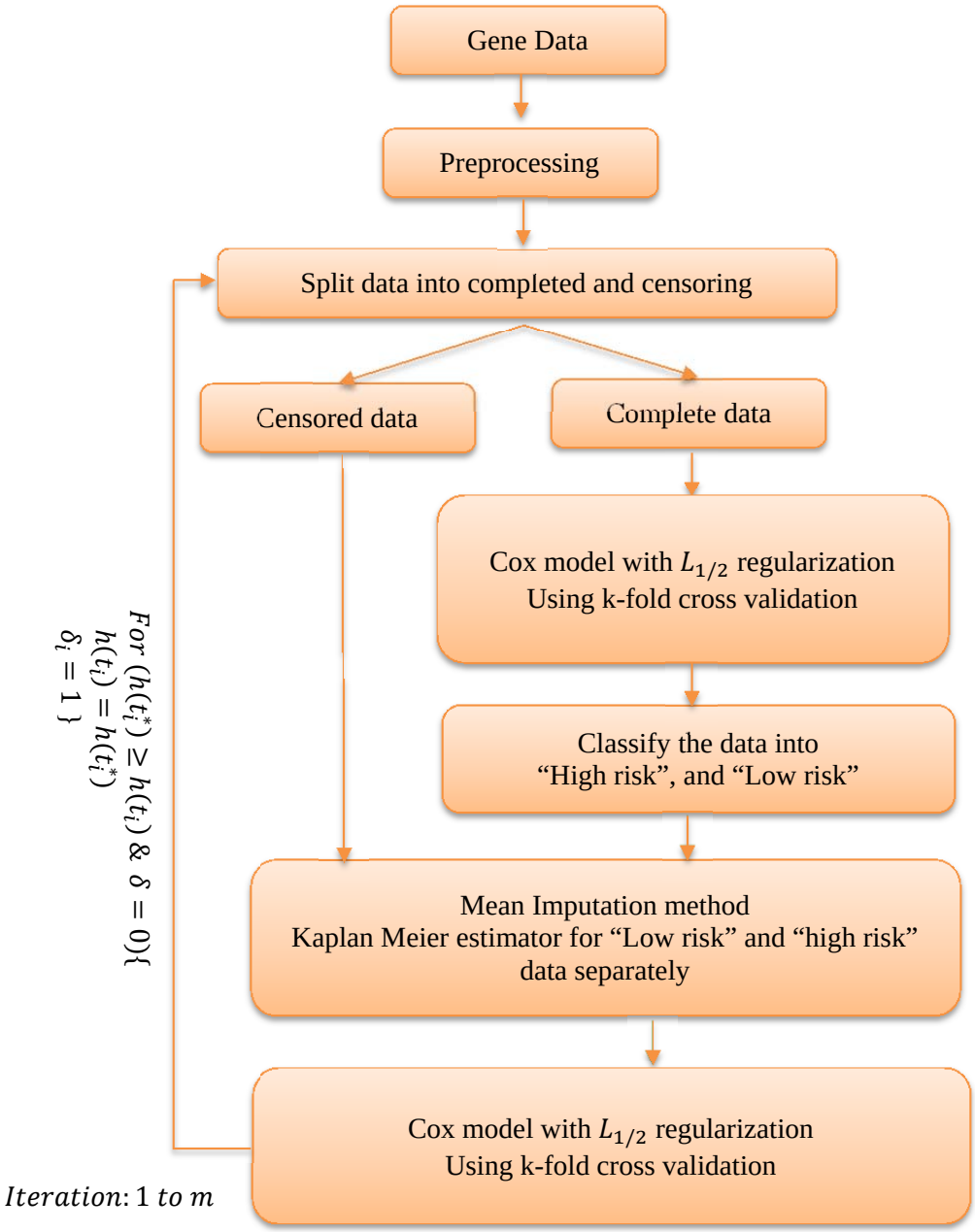

Figure 3.3. Flowchart of semi-supervised Cox model with $L_{1/2}$ penalty

3.2.4. Semi-Supervised Accelerated failure time (AFT) model

The semi-supervised AFT model is another model which is presented in this study. The model is like the aforementioned model in the Semi-supervised Cox model. By this difference that using Semi-AFT model, the prediction of survival time can be calculated by using the AFT model. So, in this model after preprocessing by scaling the gene expressions to have zero mean and 1 standard deviation and removing the missing values, the log of survival times is used as the output of the model. The completed data then will be selected (no censoring). The Cox-proportional hazard model with $L_{1/2}$ penalty is trained using completed data. 3 fold cross validation is used to select the parameters. Then the model is used to classify the data into high risk and low risk observations. The mean imputation method is used to impute the censoring observations. The high risk and low risk observations are considered separately in calculation of Kaplan-Meier estimator. Then the full data (censoring and non-censoring) is used by an AFT model with $L_{1/2}$ penalty to fit the final model. The prediction of survival time from final model was used to get the new survival time. Those predictions which have prediction survival time > censored survival time are replaced by the prediction and δ_i is set equal to 1. This process is iterated for M times. The flowchart of Semi-supervised AFT with $L_{1/2}$ penalty is presented in Figure 3.4.

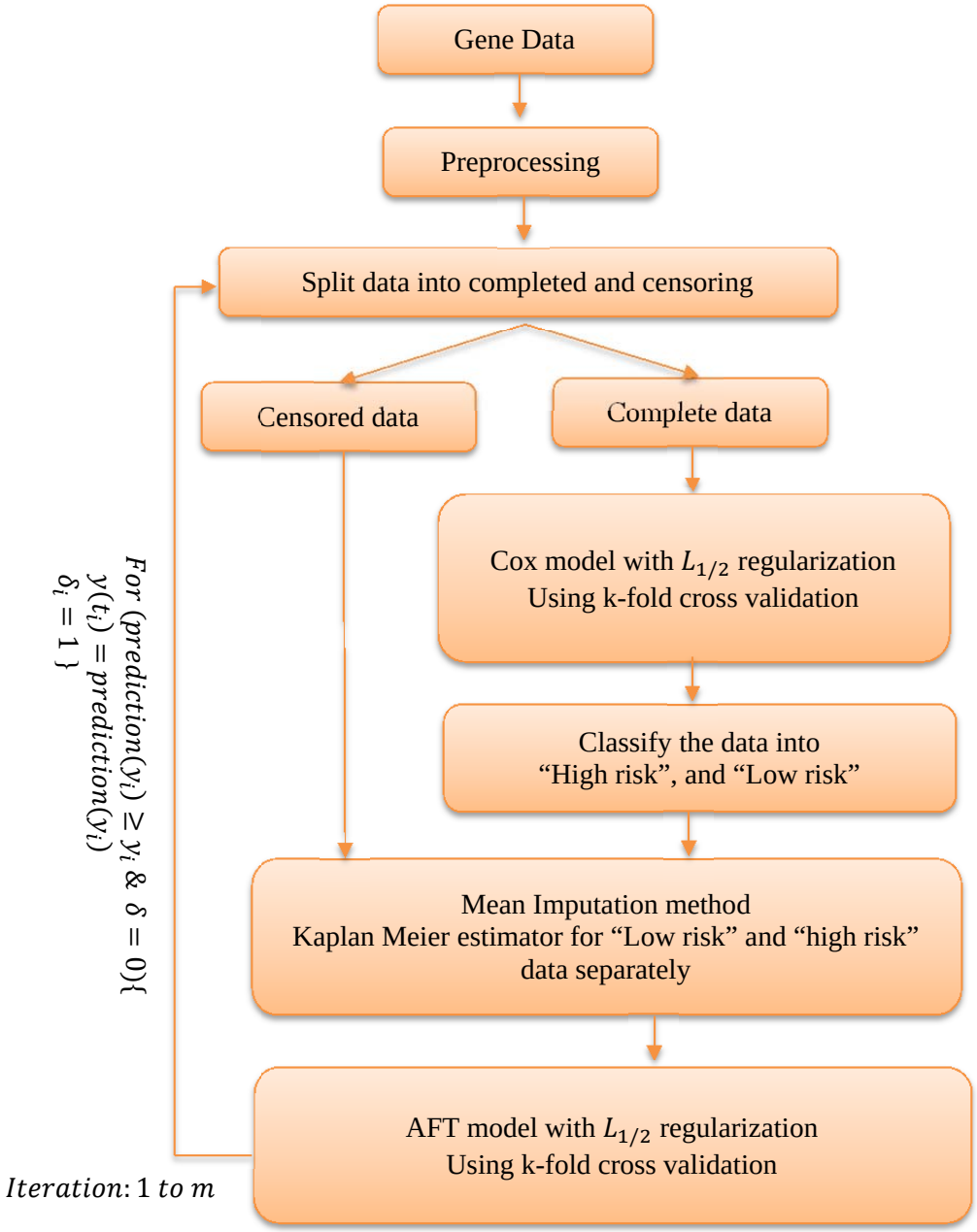

Figure 3.4. Flowchart of semi-supervised AFT model with $L_{1/2}$ penalty

For each of single Cox, single AFT, Semi-supervised Cox and Semi-supervised AFT the tuning

parameter λ is calculated by using 3-fold cross validation and accepting the tuning parameter λ which has the highest concordance index on the testing portion of the cross-validation splits. Each of these four survival analysis methods were evaluated by metrics of Concordance Index and Integrated Brier Score (IBS).

3.2.5. Concordance Index

The Concordance index explains that for all comparable pairs of observations i and j, if the predicted survival is less in comparison observations i and j, the recorded survival time is also less and if the predicted survival is more in comparison of i and j , then the survival time of the i is also more from survival time of j. Then the pair is considered as concordant. Otherwise, it will be considered as discordant. The concordance index calculates the proportion of concordant pairs in total comparable pairs. The concordance index can be calculated using equation (9).

$$CI = \frac{\sum_i \sum_j 1(f_i < f_j \wedge \delta_i = 1)}{\sum_i \sum_j 1(t_i < t_j \wedge \delta_i = 1)} \quad (9)$$

In equation (9) t is survival time, f is the predicted survival and δ is the status (censored = 0, uncensored = 1). Concordance index close to 0.5 shows equal concordant and discordant pairs so the pairs are simply random. So higher concordance index is better which shows the proportion of concordant pairs are more than discordant pairs. For a perfect model concordance index is close to 1.

3.2.6. Integrated Brier score

Brier score is time dependent and is calculated using equation (10).

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\hat{S}(t\,|X_i)^2\,1\,(t_i \le t \wedge \delta_i = 1)}{\hat{G}(t_i)} + \frac{(1-\hat{S}(t\,|X_i))^2\,1\,(t_i > t)}{\hat{G}(\mathrm{t})}\right] \qquad (10)$$

In equation (10), $\hat{G}(.)$ is the Kaplan-Meier estimator for censored data. So, using the censored data the Kaplan-Meier estimator or empirical Cumulative Distribution Function (CDF) of the censored observations will be calculated. $\hat{G}(t_i)$ is the Kaplan-Meier estimator at time t_i and $\hat{G}(\mathrm{t})$ is the Kaplan-Meier estimator at time t. $\hat{S}(.\,|X_i)$ is the estimated survival for the ith patient. For $t_i \le t$ it is expected that the censored Kaplan-Meier survival of t_i be more than estimated survival of patient i because at time t which is more than t_i the patient i has already passed away. So the estimated survival of patient i at time t is expected to be less than censored Kaplan-Meier estimator of observation i. The reason is that for observation i, the survival of observation at time t_i is more than the survival of that observation at time t. The second part of the equation (10) which is for $t_i > t$ the survival of the observation i at time t is expected to be high because at time t the observation i is still alive. So the probability of his survival up to time t should be high. So again 1 minus the estimated survival of patient i at time t should be small value. As we can see the good brier score is expected to be small. Integrating Brier score (IBR) from all interval times that we have in our study from 0 to max(t), we can calculate the integrated Brier score (IBR). The integrated Brier score (IBR) for a good model should be close to 0. For a random model the integrated brier score is approximately close to 0.25. The formula of integrated brier score (IBR) is shown in equation (11).

$$IBS = \frac{1}{\max(t_i)}\int_0^{\max(t_i)} BS(t)dt \qquad (11)$$

Chapter 4

Results and Discussion

The survival analysis was performed on three high dimensional real data using semi-supervised learning. Three real datasets DLBCL (2003) [30], AML [31] and Pancreas dataset [32] which was introduced in the previous chapter was used for the survival analysis. Four survival analysis methods which are single Cox, single AFT, semi-supervised Cox and semi-supervised AFT were used for the analysis. The proposed methods also were checked by using simulated survival datasets to assess the precision of the selected variables in high dimensional simulated data.

4.1. Survival analysis on Real datasets

Three real datasets of DLBCL (2003) [30], AML [31] and Pancreas [32] were used for the survival analysis. These datasets are high dimensional, because they have many gene expressions. For each dataset the number of patients is much less than the number of parameters ($n << p$). In such datasets, dimension reduction is required to reduce the number of parameters and keep only the gene expressions which affect the survival of the patients. In this study the dimension reduction was done by employing $L_{1/2}$ regularization in four methods of single Cox, single AFT, Semi-supervised Cox and semi-supervised AFT.

4.1.1. Results of four models for DLBCL (2003)

In this study the DLBCL (2003) [30] dataset is exactly the same as the one used previous study[16], where in their analysis they get (19, 13, 26, 22) as the number of selected parameters for single Cox, semi-cox, single AFT and semi-AFT respectively. In our analysis we get (47, 13, 18, 21) for four aforementioned methods respectively. The number of selected parameters for semi-supervised learning are almost the same for both studies. But for single cox and single AFT they are a bit different. The number of selected parameters for each method was derived by using cross validation. In Table 4.1, the results of DLBCL2003 data are presented for this study and compared with the previous study.

Table 4.1. Results for DLBCL (2003) data using four survival models

Metrics	Single Cox		Semi-Cox		Single AFT		Semi-AFT	
	Current study	Previous study	Current study	Previous study	Current study	Previous study	Current study	Previous study
NSP	47	19	13	13	18	26	21	22
CI	**0.67**	0.63	**0.82**	0.7	0.57	0.66	0.72	0.73
IBS	0.21	0.13	**0.06**	0.09	0.14	0.14	0.11	0.11

In Table 4.1, NSP is the number of selected parameters, CI is concordance index and IBS is the integrated Brier score. Previous study in Table 4.1 refers to the results obtained and current study

is the results obtained in this thesis. As we see in Table 4.1, the results found in this study using semi-cox models are better than the previous results. The Concordance index of the Semi-cox model is 0.82 > 0.7 and the integrated Brier score is 0.06 < 0.09. For concordance index metrics as explained in Chapter 3, higher value shows better results or more correctly predicted risk (concordant risk) and for IBS less value is better. For the semi-AFT the results are almost the same for both studies. Comparing the fraction of censoring in previous study the semi-supervised learning, only 3.25% of the observations remained as censoring after implementing a semi-supervised learning method. In our study, 7.61% of the observations remained censored using semi-supervised learning.

4.1.2. Results of four models for AML dataset

For AML dataset [31], after merging the dataset for gene expression and survival, the final data include 2828 gene expressions for 116 patients. The data also includes missing values for some gene expressions. The missing values were removed from the dataset. The AML dataset is different from the one used in which they had 6283 gene expression for the 116 patients. In our dataset 49 out of 116 patients are censored while in their dataset 48 of them are censored observations. Anyway, since the patients are same in both dataset by having different number of censoring observations, we will compare the results of the AML data on four models of single cox, single aft, semi cox and semi aft. The patients of AML data were analyzed by survival analysis models to obtain the number of selected parameters out of 2828 genes, Concordance index for the predicted risk and the integrated brier score. The results can be seen in Table 4.2. The results of the current study show less value for a selected number of parameters. This is

expected since the number of parameters in the AML data of this study is much lower.

Table 4.2. Results for AML data using four survival models

Metrics	Single Cox		Semi-Cox		Single AFT		Semi-AFT	
	Current study	Previous study	Current study	Previous study	Current study	Previous study	Current study	Previous study
NSP	22	26	9	20	18	39	22	32
CI	0.61	0.64	**0.81**	0.68	0.55	0.66	**0.73**	0.72
IBS	0.23	0.17	0.20	0.13	0.18	0.13	**0.11**	0.14

Comparing the results, we can see that for the semi-AFT model both concordance index 0.73 and integrated brier score 0.11 are a bit better for the current study. The numbers are shown in bold in table 4.2. For the semi-cox model, the concordance index is better than previous study but the integrated brier score is worse.

4.1.3. Results of four models for Pancreas dataset

The Pancreas dataset [32], includes the gene expression and survival times for 90 patients. This data includes 28869 gene expressions for these 90 patients. However, the survival time of 84 patients is available. From these 84 patients, survival time of 26 patients is censored and 58 of them are uncensored. Four survival analysis models were trained on the pancreas data as well by implementing 5-fold cross validation. The results for number of selected parameters, concordance index and integrated brier score for Pancreas data is presented in Table 4.3. As we

see in the table, Semi-AFT has the best performance compared with other three models, with CI = 0.85 and IBS = 0.14. After that semi-cox has better performance compared with single AFT and single cox model. The single AFT also is better compared with single Cox for Pancreas dataset.

Table 4.3. Results for Pancreas data using four survival models

Metrics	Single Cox	Semi-Cox	Single AFT	Semi-AFT
NSP	80	7	112	21
CI	0.66	0.83	0.78	**0.85**
IBS	0.25	0.18	0.21	**0.14**

The lowest number of selected parameters is in semi-cox which has only selected 7 parameters and the highest is for single AFT which has selected 112 parameters.

4.1.4. Comparing four survival models

For the three real datasets, the four models of single cox, single aft, semi supervised cox and semi supervised aft have been employed for survival analysis. For each model 5-fold cross validation was done by splitting the model into 5 folds of training and testing data. The survival model was trained on the training data and tested by testing data. The number of selected parameters for four models in three datasets are shown in Figure 4.1. As we can see the number of selected parameters is highest in single cox compared with other models, except for Pancreas

data in which we can see the number of selected parameters is higher in the single AFT (112) compared with single cox (80). The lowest number of selected parameters can be seen in semi-supervised models specially in semi-cox. Semi-cox in three datasets include the lowest number of selected parameters. Single AFT and semi-AFT are not much different in number of selected parameters. Only in Pancreas data which is a big data the number of selected parameters is relatively higher in the single AFT model.

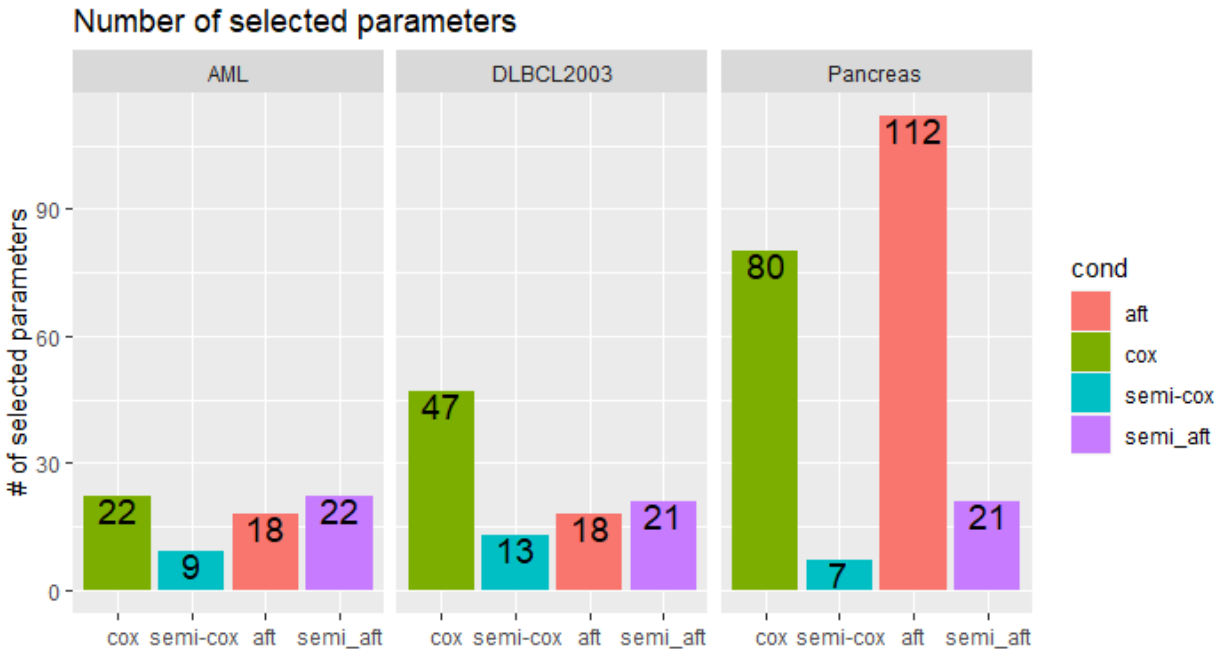

Figure 4.1. Bar plot for number of selected parameters in Real datasets

The Concordance index calculated on the testing data is showing the proportion of the events which are predicted in correct order of their survival times. The higher value for CI shows a better model which predicts the risks appropriately with the survival times. The bar plot for

concordance index of the three real datasets are shown in Figure 4.2. Comparing the four models we can see that the semi-cox model has a high concordance index in all three datasets. The CI for semi-supervised cox model is 0.83 for pancreas, 0.82 for DLBCL2003 and 0.81 for AML dataset. The semi-supervised AFT model has also high values in CI after the semi-cox model. The CI of semi-supervised AFT is 0.85 for the pancreas model which is better than the semi-cox model. But for two other real datasets (DLBCL 2003 and AML) the CI is lower compared with semi-cox. The CI for DLBCL2003 and AML are 0.72 and 0.73 respectively. We can see that the CI of single cox is more in AML and Pancreas data compared with single AFT. But for Pancreas data the CI of single AFT is higher compared with single cox. We can see a clear improvement in the CI by implementing semi-supervised learning models.

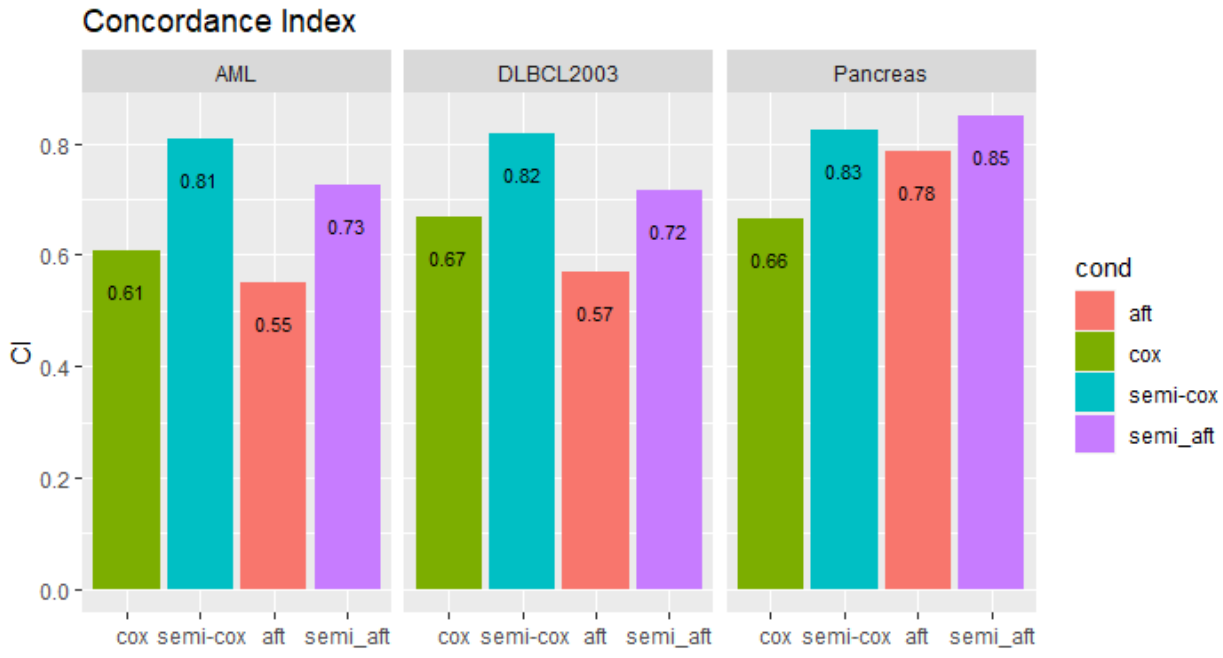

Figure 4.2. Bar plot for Concordance index in Real datasets

The integrated brier score is another measure for evaluating the survival model. This measure is time dependent. The Brier score at each time t shows the accuracy of the predicted survival for that time. The numerator of the Brier score is the predicted survival function and the denominator is the Kaplan-Meier estimator for the censored data. If the ith patient in the data with feature information X_i has survived until T_i, then the brier score for $t > T_i$ for this patient will be the predicted survival function until time t given information of X_i, given by the cumulative hazard (or cumulative risk) equation $S(t|X_i)$. As we know the patient t has not survived until time t, so we expect that probability of surviving of this patient until time t to be a low value, much less than the empirical cdf (Cumulative Distribution Function) of the survival time until T_i for the censoring data (Kaplan-Meier estimator for censored data at T_i). Hence, lower value for brier score is better, for a patient with T_i which is more than t ($T_i > t$). Then we expect to have a high probability of predicted survival at time t. The survival function closer to 1 is expected for such patients. For these patients the value of $(1 - S(t|X_i))$ will be close to zero. The brier score again is expected to be a low value. Because for patients with $T_i > t$. the brier score is calculated as $(1 - S(t|X_i))^2$ divided by the Kaplan-Meier estimator at time t. The denominator in brier score is used when censoring exists in the data to adjust the value of brier score. The integrated brier score, calculate the summation of brier score for all time ranges seen in the data. The low values of IBS show better fit because it shows lower distance between the survival function and the observed survival of the patients. For a random model the integrated brier score is close to 0.25. The values less than 0.25 indicate a fit better than a random model. The IBS for the three datasets and each four survival models is shown in Figure 4.3.

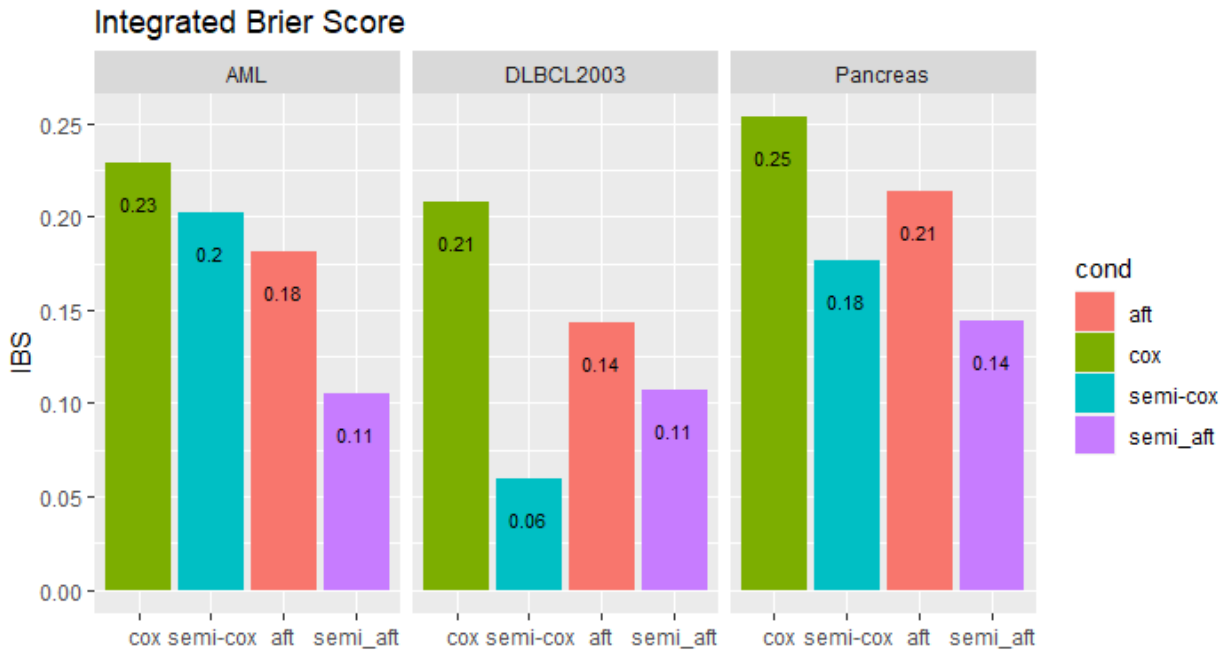

Figure 4.3. Bar plot for Integrated brier score for Real datasets

As we can see the IBS for Semi-AFT is the lowest for Pancreas and AML dataset. So according to the IBS, the Semi-AFT has the best performance in Pancreas and AML dataset. For DLBCL (2003), the semi-Cox model has the lowest IBS (0.06) which shows the best performance is for Semi-cox in DLBCL (2003) data. Comparing single AFT with single cox model, we can see the integrated brier score is lower for single AFT in all three datasets compared with single cox. So, the single AFT has better performance compared with single cox according to the IBS measure for evaluation of the survival model.

The semi supervised learning model has the ability of imputing the censored observations. The censored observations are imputed in a semi-supervised learning model by implementing cox model on the complete data and classifying the data into low risk and high risk. Then Kaplan-

Meier estimator will be used for low risk and high-risk data separately to impute the censored observations by mean imputation method. Then the imputed data will be used by the AFT model to predict the survival times. If the predicted survival times are equal or more than censoring time, it is correct otherwise it will be considered as an error. This process is iterated several times in a semi-supervised learning model. So that the censored observations are imputed as much as possible. The process of classifying the data into high risk and low risk is done by a single cox model and a semi-supervised learning model for three real datasets. The fraction of censoring for AML data was 41.54% in single cox, while using semi-supervised learning the fraction of censoring reduced to 29.23%. It means that 12.31% more of the observations which were censored, could be classified by using a semi-supervised learning model. For the DLBCL2003 data the fraction of censoring in single cox is 30.43% while using a semi-supervised learning model this fraction reduced to 7.61%. It means 22.82% more of the observations could be classified by using a semi-supervised learning model. For Pancreas data, the fraction of censoring is 30.95% in the single cox model while the value is reduced to 20.24% in the semi-supervised learning model. It means 10.71% more of the observations were classified using a semi-supervised learning model.

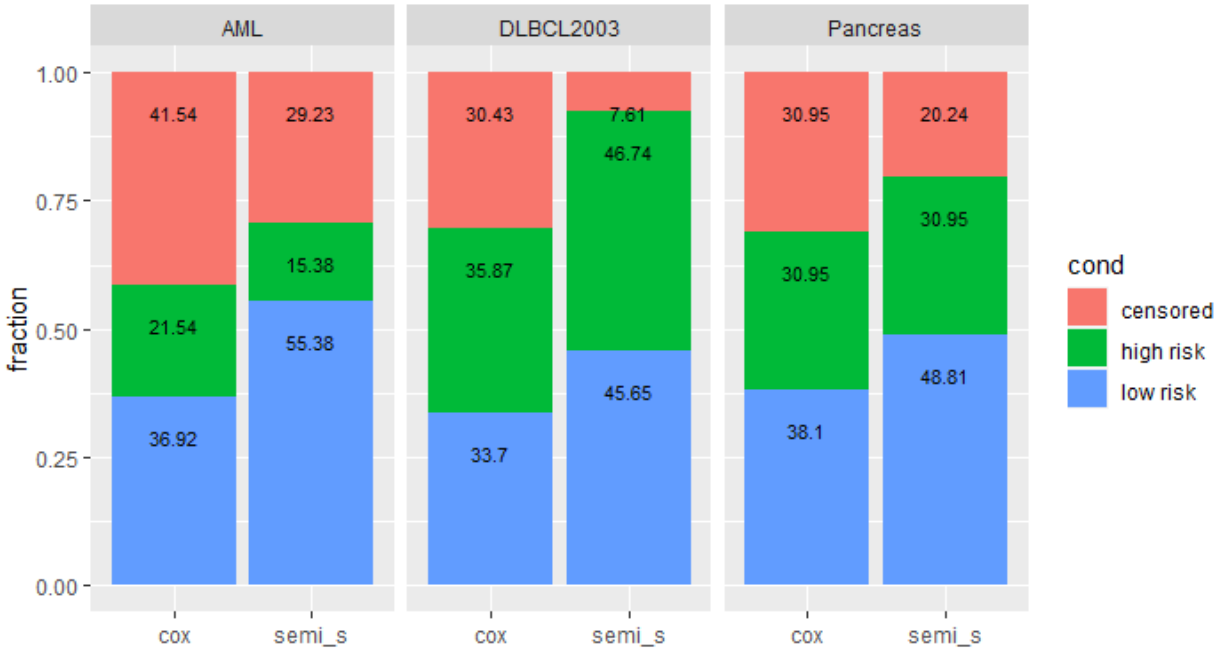

Figure 4.4. classifying the data using single cox and semi-cox model.

As we see in Figure 4.4 the proportion of high risk and low risk observations are different in single cox and semi-cox model. In the AML data 36.92% of the observations are at low risk in single cox and 21.54% as high risk, while in semi-cox the proportion of low risk is 55.38% and high risk is 15.38%. In two out of three real datasets, we can see that in semi-cox the proportion of low risk and high risk increases or remains the same compared with single cox model. Only in the AML dataset we can see that the proportion of high risk is reduced in semi-cox compared with single cox. In the Pancreas dataset the low risk was 38.1% in single cox and it was increased to 48.81% in the semi-cox model. The high risk in single cox is 30.95% and it remains as 30.95% in a semi-cox model. The DLBCL2003 data include 33.7% as low risk and 35.87% as high risk in single cox. The proportion of low risk and high risk are 45.65% and 46.74% respectively in the semi-cox model. The fraction of low risk, high risk and censoring in this study

was compared for DLBCL (2003) in Table 4.4.

Table 4.4. Classified data in single cox & semi-cox for DLBCL.

status	DLBCL (2003)			
	Single cox		Semi Cox	
	Current study	Previous study	Current study	Previous study
Low risk	33.7	31.52	45.65	43.47
High risk	35.87	38.04	46.74	53.26
censored	30.43	30.44	7.61	3.26

As we see in the previous section the CI and IBS of the semi-cox was better in this study. Looking at Table 4.4 we can see that after using semi-cox 7.61% of the data remained censored while it was less (3.26%) in previous study. The proportion of high risk is less in our study. Although in the previous study more censoring observations were imputed, but the proportion of high risk is 6.52% more compared with our study.

For the AML dataset, as we see in the previous section the semi-cox model had a high concordance index but the IBS was low for this data. In the proportion of Figure 4.4 we can see that the proportion of low risk increased to 55.38% using semi-cox and the proportion of high risk reduced to 15.38%. The performance of the Semi-AFT model was better for AML dataset in which we see both Concordance index and Integrated brier score were good and a bit better. The proportion of Low risk, high risk and censoring using semi-AFT for three real datasets can be seen in Figure 4.5.

Figure 4.5. classifying the data using single cox and semi-aft model.

As we can see (Table 4.5), the proportion of low risk is 33.85%, high risk is 38.46% and censoring is 27.69% in AML data for semi aft.

Table 4.5. Classified data in single cox & semi-AFT for AML.

status	AML			
	Single cox		Semi AFT	
	Current study	Previous study	Current study	Previous study
Low risk	36.92	31.03	33.85	47.11
High risk	21.54	26.72	38.46	47.72
censored	41.54	42.25	27.69	5.17

As we mentioned in the previous section, the AML data is not exactly the same in our study compared with previous study. The data has lesser number of genes, 2828 in this study, while 6283 genes were in the dataset of the previous study. The number of censors is also not the same, 49 patients in this study while 48 patients in the previous study. However, the performance of our semi-AFT model is almost the same and a bit better in our study. It can be seen that although the fraction of imputation is less in our study but this does not affect the model performance and we can see a good performance in the survival model.

4.2. Survival analysis on simulated survival datasets

The survival data was simulated using Gompertz distribution [37]. The methodology for simulating survival data and gene expressions is the method of Bender et al [33]. The data was drawn from independently and normally distributed features with zero correlation and also with correlation equal to 0.3. For each data 1000 variables were simulated. The coefficients were set

for 990 of them as 0 and for the first 10 features the coefficients were set equal 1. Then the survival time is computed using the coefficients and the randomly generated features from normal distribution with 0 and 0.3 correlation between the features. The survival time was calculated using the Gompertz distribution. For each simulated data, 40% of the data were considered as censoring. For these censored observations minimum of the simulated y and the randomly assigned survival was set to be used as survival time. For each sample size (N = 100, 200 and 300) 50 datasets were simulated with 1000 features. 10 out of these 1000 features are prognostic genes and the rest are zero. Four methods of single cox, single aft, semi-cox and semi-aft were used on each of these simulated data to find the prognostic genes. Three measures which were used for evaluation of the survival models are:

(1) number of correctly selected parameters: This measure is calculated as the average of the number of correctly selected parameters in 50 simulated datasets. Each simulated dataset includes 10 prognostic genes. So, this measure shows that on average how many of these 10 prognostic genes were detected.

(2) total number of selected parameters: This measure is calculated as the average of the total number of selected parameters in 50 simulated datasets. Each dataset includes at most 1000 parameters. So, this measure shows how many of these parameters are selected on average in 50 simulations.

(3) Precision: This measure is calculated as the division of the first item with the second item, which shows how many correct parameters were selected from total selected parameters.

A good model will be the one which has a higher number of correctly selected parameters and also higher precisions. This means that we want to have a higher value of the number of correctly

selected parameters while keeping the total number of selected parameters as low as possible, which causes high precision in finding the prognostic genes.

4.2.1. Single Cox Model

A single Cox proportional hazard model was fitted to each simulated dataset. From 1000 parameters which are much more than sample size, we want to keep only the prognostic genes and use them for predicting the hazard ratio. $L_{1/2}$ regularization is employed for shrinking the coefficients toward zero and finally keeping the prognostic genes which only have true influence on the survival in the model. The tuning parameter was calculated using 5-fold cross validation. The model chooses the tuning parameter as the value which has the highest concordance index in the validation splits. For the single cox model, 50 datasets were simulated and a single cox model with $L_{1/2}$ regularization was fitted on each of these simulated data. Then for each of the simulated dataset, the number of correctly selected parameters, total number of selected parameters and the precision were calculated. Then the average of these measures was calculated for 50 datasets. Three sample sizes of N = 100, 200 and 300 were used with correlation (0 and 0.3). For each of these sample sizes and correlation, 50 datasets were simulated and the average of the correctly selected parameters and precision were reported. So, for 3 sample sizes and 2 correlation coefficients, 300 datasets were simulated only for the single cox model. Table 4.6 shows the results for the single cox method. The highest number of correctly selected parameters (10) can be seen in N = 300 and correlation of (0 and 0.3). The lowest total number of selected parameters is for N = 300 and correlation equal zero (40.46). The highest precision (0.287) can be seen also in N = 300 and no correlation (cor = 0). In the Table, (S.e) means standard error of the measure.

For example, “correct (s.e)” means standard error of the number of correctly selected parameters.

Table 4.6. Results for simulating 50 datasets and fitting using the single cox model.

Cor.	Sample Size	Single Cox					
		Correct	Correct(s.e)	Total Selected	Selected(s.e)	Precision	Precision(s.e)
0	100	6.48	0.43	77.88	9.90	0.231	0.041
0	200	9.98	0.02	65.28	5.28	0.193	0.012
0	300	**10**	0.00	**40.46**	2.51	**0.287**	0.016
0.3	100	7.96	0.28	93.86	5.44	0.093	0.005
0.3	200	9.98	0.02	86.38	3.65	0.123	0.004
0.3	300	**10**	0.00	90.94	3.13	0.116	0.003

We can see that for correlated data the total number of selected parameters are more than uncorrelated datasets. We can see that the single cox model was able to find all the prognostic genes when sample size is N = 300 for both correlated and uncorrelated datasets. For N = 200 also the number of correctly selected parameters are high. For N = 100, the number of correctly selected parameters for correlated data is more than uncorrelated data. However, the correlated data tends to keep a greater number of parameters in the model. We can see clearly that the total number of parameters are more in correlated datasets. This is the reason that precision is less for correlated data compared with uncorrelated dataset. If we compare the results with the one obtained in previous studies the precision which they report is a bit different from the precision which was reported in Table 4.6. To explain the difference, let's give an example. Assume we have 5 values of {10, 8, 6, 4, 8} as the number of correctly selected parameters and {70, 60, 80, 20, 50} as the total number of selected parameters. We have calculated the precision as average

of {10 / 70, 8/60, 6/80, 4/20, 8/50} = 0.142. While they calculated the average of correct = 7.2 and average of total = 56. Then for precision they divided the average of correct with average of total = 7.2/56 = 0.128. So, the value is a bit different. For comparison we calculate the precision as they have calculated. In Table 4.6 the results of the single cox model are compared. As can be observed, the values of precision are different from Table 4.6. In Table 4.7, the results obtained for single cox in simulation study is presented and compared.

Table 4.7. Comparing results of single cox model.

Cor.	Sample Size	Single Cox					
		Correct	Correct (Liang et. al)	Total Selected	Total Selected (Liang et al.)	Precision	Precision (Liang et al.)
0	100	**6.48**	4.06	77.88	24.44	0.083	0.166
0	200	**9.98**	5.62	65.28	28.22	0.153	0.199
0	300	**10**	8.02	40.46	35.18	**0.247**	0.228
0.3	100	**7.96**	3.90	93.86	24.38	0.085	0.159
0.3	200	**9.98**	5.68	86.38	29.64	0.116	0.192
0.3	300	**10**	7.84	90.94	35.86	0.110	0.219

As we can see in Table 4.7 in terms of the number of correctly selected parameters, our results show higher value in all sample sizes and correlations. However, in terms of total number of selected parameters our model selects a greater number of parameters and hence it has lower precision compared to Liang et al. [16]. Only for N = 300 and correlation zero the precision of our results (0.247) is more than (0.228) in Liang et al [16].

4.2.2. Semi-Cox Model

Semi-supervised learning method was used to do survival analysis using simulation data. 50 datasets were simulated for each of N = 100, 200 and 300 and correlation of (0 and 0.3). So, like the single cox method, in total 300 data were simulated and trained using the semi-cox model. Semi-cox model also has regularization of $L_{1/2}$ to shrink the coefficients toward zero and keep only the prognostic genes. This method first fits a cox model on complete data (data which do not include censoring). Then it will classify the data into high risk and low risk. For the classified data the Kaplan-Meier estimator is used to get the empirical distribution of the dataset for low risk and high risk separately. Then the censored data were imputed using the mean imputation method. Finally, a cox model was fitted on the imputed dataset. The coefficients are calculated. From the calculated coefficients we determine which one is correctly detected and are actually prognostic genes and which one are not. The three measures explained in the single cox method is used to evaluate the performance of the semi-cox model. The semi-cox model is an iterative model. However, in this study we have used one iteration for the semi-cox model. The results obtained for the semi-cox model for 50 simulation data using each sample size and correlation are presented in Table 4.8.

Table 4.8. Results for simulating 50 datasets and fitting using the semi-cox model.

Cor.	Sample Size	Semi-supervised Cox model					
		Correct	Correct(s.e)	Selected	Selected(s.e)	Precision	Precision(s.e)
0	100	9.28	0.21	32.26	3.40	0.453	0.039
0	200	**10.00**	0.00	33.80	3.23	0.492	0.049
0	300	9.98	0.02	48.56	3.63	0.315	0.037
0.3	100	9.12	0.20	26.66	3.24	0.547	0.041
0.3	200	9.84	0.10	**23.56**	2.83	**0.655**	0.047
0.3	300	**10.00**	0.00	28.64	3.47	0.629	0.052

As we can see in Table 4.8, The number of correctly selected parameters are high in all of the sample sizes and correlations. For N = 200, cor. = 0 and for N = 300, cor. = 0.3 the value is 10. It means all the prognostic genes were successfully detected. The Total number of selected parameters are lower in correlated data. In single cox we saw that for correlated data the total number of selected parameters were higher, while here we can see that the total number of selected parameters are lower in correlated data compared with uncorrelated datasets. The lowest value of the total number of selected parameters can be seen for N = 200 and cor. = 0.3 which is 23.56. The precision of N = 200 and Cor = 0.3 is also the highest precision compared with other sample sizes and correlation. The precision of the correlated data is more than uncorrelated dataset.

For comparison, the results of the semi-cox method of this study on 50 simulated data are presented in Table 4.9. Again, to have same calculation for precision, the same method for calculating the precision as was used in Table 4.9.

Table 4.9. Comparing results of semi-cox model.

Cor.	Sample Size	Semi-supervised Cox model					
		Current study Correct	Previous study Correct	Current Study Total Selected	Previous Study Total Selected)	Current Study Precision	Previous Study Precision
0	100	**9.28**	6.58	32.26	16.96	0.288	0.388
0	200	**10.00**	8.68	33.80	17.84	0.296	0.487
0	300	**9.98**	9.76	48.56	19.02	0.206	0.513
0.3	100	**9.12**	6.46	26.66	17.08	0.342	0.378
0.3	200	**9.84**	8.62	23.56	17.86	0.418	0.483
0.3	300	**10.00**	9.42	28.64	18.54	0.349	0.508

We can see that although in all the sample sizes and correlations the number of correctly selected parameters are more in our results. But in terms of total number of selected parameters we are also higher compared. The closet value of precision is N = 200 and Cor = 0.3 which is 0.418 in our study compared with 0.483 in their study. However, although they get higher precision, but that led to reduce or miss some of the prognostic genes. So, our model tends to select as much prognostic genes as possible compared with their model.

4.2.3. Single AFT Model

A single accelerated failure time model was used to do simulation study. The AFT model with $L_{1/2}$ regularization was used to reduce the dimensionality of the data and shrink the coefficients

toward zero to keep only the effective parameters in the model. The AFT model was used on simulated data with $L_{1/2}$ regularization to see how accurate the model keeps the coefficients in the data. To evaluate the single AFT model, 50 random datasets were simulated. Each simulated data includes 1000 parameters where 10 of them are prognostic genes. The number of correctly selected parameters from these 10 prognostic genes among 1000 parameters were calculated. The average of the number of correctly selected parameters is used as the average value of these 50 simulations. The precision is the number of correctly selected parameters divided by the total number of selected parameters for each model. The average of precision is also calculated to evaluate the single AFT model in parameter selection. The regularization was done by $L_{1/2}$ penalty. The value of the tuning parameter was calculated using 5-fold cross validation. The tuning parameter was selected as the value which has the least mean squared error of the survival time in the AFT model in validation part of the 5-fold cross validation. Table 4.10, shows the results of simulation study for 50 simulated datasets.

Table 4.10. Results for simulating 50 datasets and fitting using the single AFT model.

Cor.	Size	Single AFT model					
		Correct	Correct(s.e)	Selected	Selected(s.e)	Precision	Precision(s.e)
0	100	9.32	0.17	336.02	58.25	**0.107**	0.015
0	200	**10**	0.00	483.18	60.29	0.050	0.005
0	300	**10**	0.00	136.86	18.44	0.093	0.005
0.3	100	8.8	0.16	166.68	34.96	0.087	0.004
0.3	200	9.88	0.05	**97.52**	1.73	0.103	0.002
0.3	300	**10**	0.00	97.96	1.83	0.104	0.002

As we see in Table 4.10, the single AFT model with correlation 0 and sample size 200 and 300 has found the 10 prognostic genes in all 50 simulations. The simulation data with sample size N = 300 and correlation equal 0.3 also has found the 10 prognostic parameters in all 50 simulations. The correlated datasets in the single AFT model have considerably lower total number of selected parameters compared with uncorrelated data. The minimum total number of parameters is for N = 200 and correlation 0.3 which is 97.52. The highest average precision (0.107) can be seen for N = 100 and correlation 0 but the standard error for average precision on this simulated data is relatively much higher compared with other datasets (0.015). The second highest precision (0.104) can be seen in N = 300 and correlation 0.3 with relatively lower standard error compared with the highest average precision.

The single AFT model results on simulated data were compared with the previous study. The precision is calculated as the average number of correctly selected parameters divided by the average number of total selected parameters to be the same as the precision obtained by Liang et al. [16]. For comparison. The results of two studies can be seen in Table 4.11. In this table the precisions are different from Table 4.10, because in table 4.10 the reported precision is the average precision on 50 simulated datasets while in Table 4.11, the precision is the average number of correctly selected parameters divided by average number of total selected parameters, which means division of column three by column five in the Table 4.11

Table 4.11. Comparing results of single AFT model.

Cor.	Sample Size	Single AFT					
		Current study Correct	Previous study Correct	Current Study Total Selected	Previous Study Total Selected)	Current Study Precision	Previous Study Precision
0	100	**9.32**	5.02	336.02	38.74	0.028	0.130
0	200	**10**	7.12	483.18	46.68	0.021	0.152
0	300	**10**	8.90	136.86	56.54	0.073	0.157
0.3	100	**8.8**	4.74	166.68	39.54	0.059	0.120
0.3	200	**9.88**	6.98	97.52	47.02	0.101	0.148
0.3	300	**10**	8.80	97.96	56.82	0.102	0.155

Comparing the results of our study with previous study we can see that in our study the number of correctly selected parameters are considerably higher compared with previous study. However, the total number of selected parameters are also higher in our study compared with previous study. So, the precision of the single AFT model is less in our study. The closet precision can be seen for N = 200 and Correlation 0.3 which is 0.101 compared with 0.148. Also, N = 300 with correlation 0.3 are close to the previous. Hence, we can see that for correlated data and various sample sizes, our model could select much more correct parameters. Also, the precision is close to the one obtained in the previous study for the correlated data.

4.2.4. Semi-AFT Model

Semi-supervised accelerated failure time model is used for simulation study. 50 random datasets were simulated and analyzed by the semi-supervised AFT model. Each dataset again includes 1000 parameters and 10 of them are prognostic genes. The semi-supervised AFT model with $L_{1/2}$ regularization is used to find the prognostic genes between these 1000 parameters. The $L_{1/2}$ penalty helps in shrinking the coefficients toward zero. The tuning parameter is calculated by using 5-fold cross validation and selected as the value which minimizes the mean squared error of the survival time. Simulation study is done for sample sizes, N = 100, 200 and 300. The data is simulated as correlated data (correlation = 0.3) and uncorrelated data (correlation = 0). The results of the semi-supervised AFT model for 50 simulation study are presented in Table 4.12.

Table 4.12. Results for simulating 50 datasets and fitting using the Semi AFT model.

Cor.	Size	Semi AFT model					
		Correct	Correct(s.e)	Selected	Selected(s.e)	Precision	Precision(s.e)
0	100	6.52	0.30	**25.52**	3.12	0.416	0.039
0	200	9.18	0.15	45.70	3.88	0.338	0.040
0	300	**9.68**	0.07	49.50	3.79	0.356	0.046
0.3	100	8.48	0.18	60.82	5.00	0.342	0.052
0.3	200	9.44	0.07	47.66	5.68	0.522	0.056
0.3	300	9.40	0.07	42.08	5.64	**0.585**	0.055

As we see in the Table 4.12, the maximum number of correctly selected parameters is for N = 300 and correlation = 0. In semi-AFT we can see that the total number of selected parameters are less for uncorrelated data except for N = 300. The minimum value of the total number of selected parameters on average for 50 simulations can be seen for N = 100 and correlation = 0. The

highest average precision also is seen for N = 300 and correlated data (0.585). We can see that in semi-supervised AFT, for low sample size (N = 100) the number of correctly selected parameters is higher for correlated data but with lower average precision. For other sample sizes (N = 200 and 300), the number of correctly selected parameters are almost equivalent in both correlated and uncorrelated data. While the average precision is higher for correlated data.

The results of the simulation study of semi-AFT is compared with previous study. Table 4.13 represents the results of two studies. The precision presented in Table 4.13 has the same calculation method as the previous study.

Table 4.13. Comparing results of Semi AFT model.

Cor.	Sample Size	Semi AFT model					
		Current study Correct	Previous study Correct	Current Study Total Selected	Previous Study Total Selected)	Current Study Precision	Previous Study Precision
0	100	6.52	6.84	**25.52**	35.52	**0.255**	0.192
0	200	**9.18**	8.84	45.70	42.16	0.201	0.210
0	300	9.68	9.86	**49.50**	50.84	**0.196**	0.194
0.3	100	**8.48**	6.72	60.82	35.84	0.139	0.188
0.3	200	**9.44**	8.78	47.66	44.96	**0.198**	0.195
0.3	300	9.40	9.78	**42.08**	49.31	**0.223**	0.198

As we see in Table 4.13, the results of our study for semi-supervised learning are close to the previous study. Our results are a bit better compared with their results. We can see in terms of the number of correctly selected parameters our results are better in 3 out of 6 simulation studies. In the other 3 also the results of the number of correctly selected parameters are close together.

The total number of selected parameters also is better in 3 out of 6 simulations. The precision calculated in our study is better in 4 out of 6 simulations. Among all the models which we have simulated the closest is the semi-supervised AFT. It is the most competitive one, in which we can see in both terms of correctly selected parameters and precision the results are close together. In previous sections we have models which show higher numbers of correctly selected parameters but lower precision. But in semi-AFT both measures are close together. However, the results obtained in this study are slightly better compared with previous study.

4.2.5. Classifying the observation using semi-supervised learning

The survival analysis on the simulated datasets were done using semi-supervised learning approaches (semi-Cox and semi-AFT). Each simulated sample has 40% of the data as censoring. Implementing single Cox or single AFT method only the 60% of uncensored data can be classified as “low risk” or “high risk” observations. But in the semi-supervised learning method we are imputing the censored observations in the survival analysis process by using mean imputation method. So, more proportion of the data can be classified as “high risk” or “low risk” observations in the semi supervised learning approach. For various sample sizes N = 100, 200 and 300 and for uncorrelated and correlated datasets the classified data can be compared in the semi-supervised learning method with the single survival model. Considering the classified data as: (test: single cox model, <a: N = 100, cor. = 0>, <b: N = 200, cor. = 0>, <c: N = 300, cor. = 0>, <d: N = 100, cor. = 0.3>, <e: N = 200, cor. = 0.3>, <f: N = 300, cor. = 0.3>). The data were simulated using the aforementioned sample size and correlations. The results of classifying the data using the single cox model and semi-supervised learning is presented in Figure 4.6. Figure

4.6 shows that for single Cox, 40% of the data is censored. While in the semi-supervised learning lower proportion remains censored. For example, for dataset “a” which is uncorrelated and with sample size 100, only 7% of the data remained as censored using the semi-supervised learning method. The rest of the data were classified into “low risk” and “high risk” observations. In the figure, the results of “a”, “b” and “c” are for uncorrelated datasets and “d”, “e” and “f” are for correlated data. The proportion remained as censoring is a bit less in the correlated dataset compared with uncorrelated data.

Comparing the results of classifying the dataset using semi-supervised learning in this study with Liang et al. [16] we can see that the proportion of censoring in semi-supervised learning in this study for “a”, “b”, “c”, “d”, “e”, “f” was 7%, 11%, 7%, 8%, 6% and 6% respectively, while it was 5.52%, 5.44%, 3.47%, 3.41%, 2.39% and 2.41% in Liang et al. [16], which is lower than in this study. The common thing is that the proportion of censoring is a bit less for correlated data compared with uncorrelated data. Except for N = 100, for uncorrelated data we get a lower fraction of observations remaining as censoring after implementing the semi-supervised learning method.

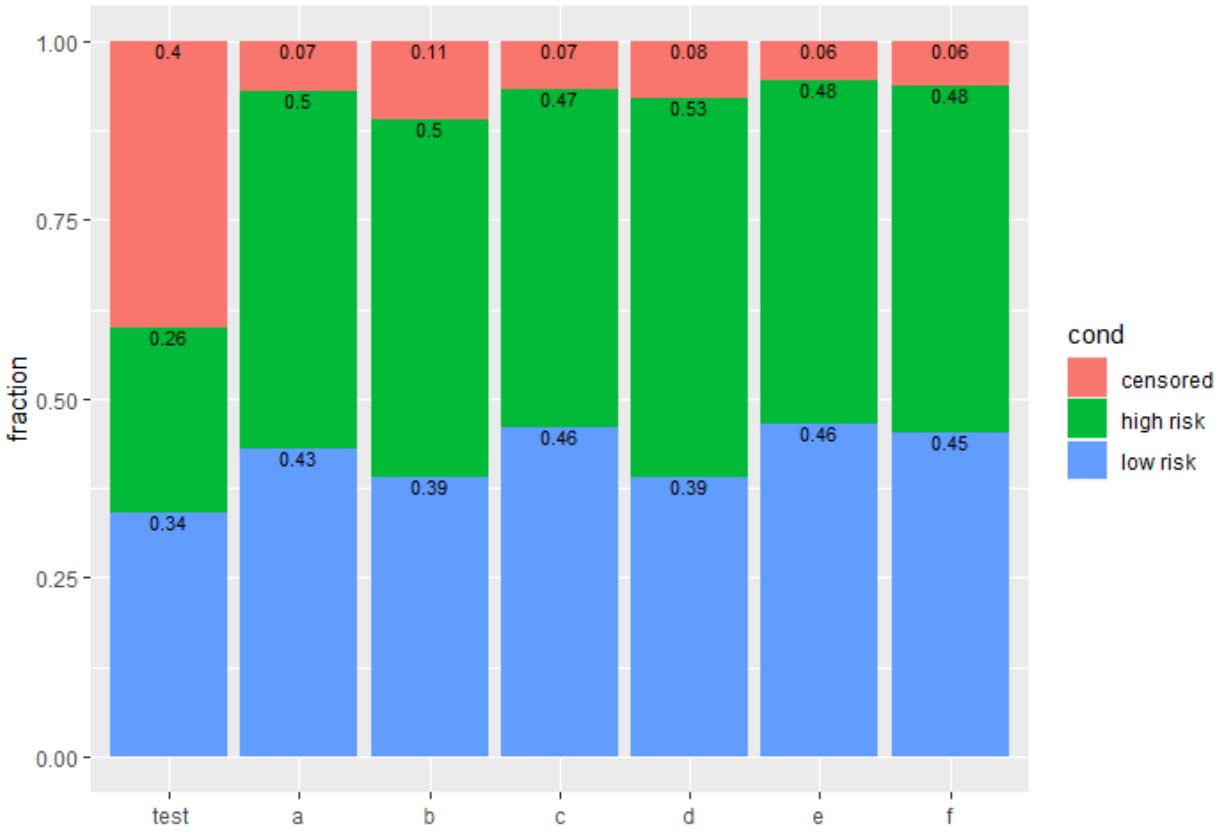

Figure 4.6. classifying the data using single cox and semi-cox models in simulation.

4.2.6. Testing data using single survival and semi-supervised learning

A testing data with sample size N = 200, were simulated using the same coefficients and correlations as in the simulation study (correlation = 0.3). Then using each model of single cox and semi-supervised cox models obtained using N = 100, 200 and 300 with correlation = 0.3 the testing data which has sample size N = 200 and correlation = 0.3 were tested. To evaluate the single cox and semi-cox models, these models were used for classifying the data of testing. The testing data also were classified by fitting single cox and semi-cox on it. Then the error rate was calculated as the classified data which are not equivalent with the classified data found by the

model fitted on the testing data. The bar plot presented in Figure 4.7 shows the error rate obtained for the testing data (N = 200) by using various models fitted on N = 100, 200 and 300.

We can see that the error rate is low for both single cox and semi-cox. However, the error is lower for higher sample size compared with lower sample size. The model of single cox fitted on a data with N = 100 has an error rate of 7%. The model of semi-cox with N = 100 has an error rate of 5%. The model with single cox with N = 200 has an error rate of 5%. The model of semi-cox with N = 200 has an error rate of 3%. The model with single cox with N = 300 has an error rate of 4% and finally for the semi-cox and N = 300 the error rate is only 2%.

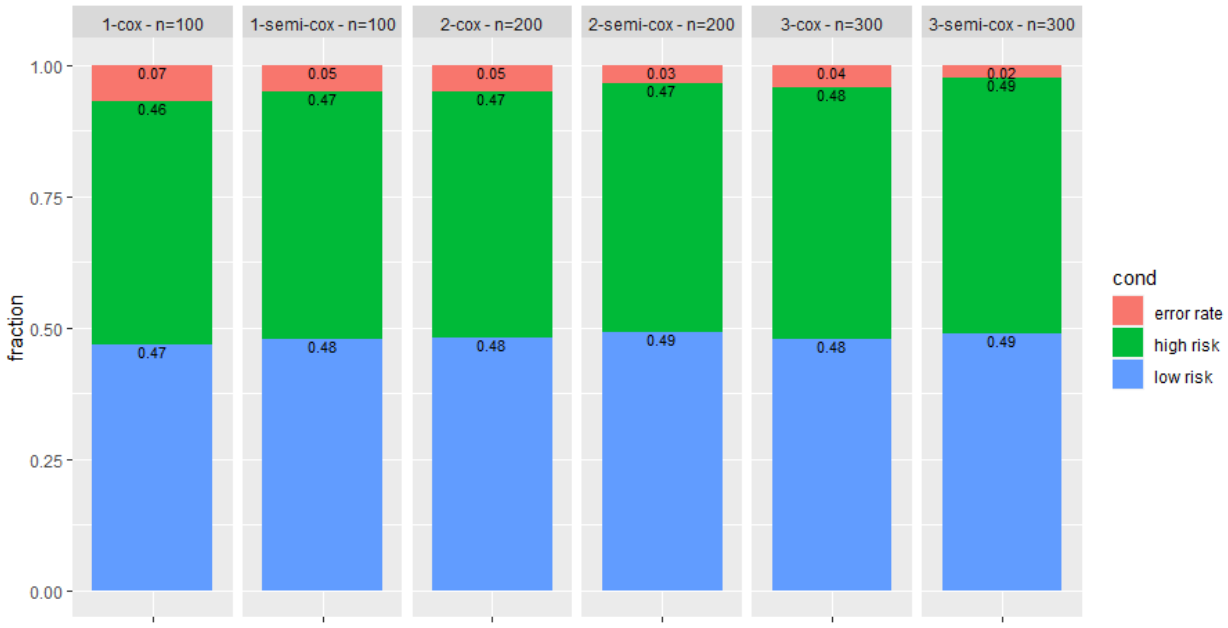

Figure 4.7: testing the single cox and semi-cox models in simulation.

Chapter 5

Conclusions and Future Work

5.1. Simulation study

The survival analysis was performed using four models of single cox, semi-supervised cox, single aft and semi-supervised aft models. For each of these four models 3 sample sizes of N = 100, 200 and 300 were used. The simulation of the data was done as two types of correlated (correlation = 0.3) and uncorrelated data (correlation = 0). For each sample size and correlation, 50 datasets were simulated. Each dataset includes 1000 parameters with 10 of them as prognostic genes. The aim of the simulation study was to find these 10 prognostic genes out of 1000 parameters. The parameter selection in each of these four methods is done by using $L_{1/2}$ regularization. The tuning parameter is calculated as the value which maximizes the concordance index in the cox models and the value which minimizes the mean squared error in the aft models. The measure of concordance index and mean squared error are calculated on the validation set using 5-fold cross validation.

The survival analysis using the mentioned four models were employed and the results were evaluated by three measures of average number of correctly selected parameters among 50 simulations. Average number of total selected parameters and the average precision of the 50 simulations is calculated by number of correctly selected parameters divided by total number of selected parameters. The results of the current simulation study also were compared with the previous study.

5.1.1. Single Cox

In the single cox model, it was seen that the values of the average number of correctly selected parameters is pretty high in sample sizes of N = 200, 300 and for both correlated and uncorrelated data (average of number of correctly selected parameters > 9). Only for lower sample size (N = 100) the average of correctly selected parameters is less than 9. For correlated data the value of average correctly selected parameters is a bit higher than uncorrelated data (7.96 in correlated and 6.48 in uncorrelated data). The total number of selected parameters are less in uncorrelated data compared with correlated data. For N = 200 and N = 300 we can see higher precision in uncorrelated data.

Comparing the single cox model, we can see that in all 6 simulations our average number of correctly selected parameters are more however, since the total number of selected parameters are also higher in our study so in terms of precision except for N = 300 and correlation = 0 (0.247 > 0.228), our precision is less than those calculated in the previous study. Tt means that our implemented method tends to get more correctly the prognostic genes however it also causes to include more redundant parameters in the model.

5.1.2. Semi-supervised Cox

In the semi-supervised Cox model for all the sample sizes the average number of correctly selected parameters are pretty high (> 9). While we saw that the single cox is sensitive to sample size, we can see that the semi-supervised cox model is less sensitive to sample size because we

get a high value (<9) of average correctly selected parameters also in low sample size (N = 100). The total number of selected parameters is lower in correlated data compared with uncorrelated data. Hence the precision is higher for correlated data using the semi-supervised cox model.

Comparing the results of simulation, we saw that for all 6 simulations the average number of correctly selected parameters are higher in our study. But the total number of selected parameters is also higher. So, the precision of our study is less than.

5.1.3. Single AFT

The single AFT model gets better results in terms of average number of correctly selected parameters compared with single cox. The single AFT gets the value of the average number of correctly selected parameters > 9 for 5 out of 6 simulations. Only for N = 100 and Correlation = 0.3 the value is 8.8 which is slightly less than 9. But in terms of total number of selected parameters and precision the single AFT is worse than single cox. Because the total kept parameters in the model are relatively high compared with single cox. Comparing correlated and uncorrelated data we can see that the total number of parameters are lower in correlated data, so the precision is higher in correlated data compared with uncorrelated data in the single AFT model.

The results of the single AFT model of our study were compared, while we saw that in all 6 simulations the average number of correctly selected parameters is higher in our study compared. But since the total number of selected parameters is also higher, hence the precision is less.

5.1.4. Semi-Supervised AFT

The results of the semi-supervised AFT model also have high value in the number of correctly selected parameters. However, this model is a bit sensitive to the sample size. In 4 out of 6 simulations the average number of correctly selected parameters is more than 9. Only for low sample size, sample size N = 100, the average number of correctly selected parameters is less than 9. The total number of selected parameters is lower than single cox and single aft models. Comparing the correlated and uncorrelated data in the semi-supervised AFT model we saw that the correlated data had a lower total number of selected parameters and relatively higher precision compared with uncorrelated data.

Comparing results of Semi-supervised AFT in our study with previous study, we can say that the most similar results which we found in our study with previous is the results of semi-supervised AFT. In terms of the three measures of average number of correctly selected parameters, total number of correctly selected parameters and precision. However, in our study we get a higher precision in 4 out of 6 simulations compared with their study. We also get a higher number of correctly selected parameters in 3 out of 6 simulations.

5.2. Conclusions on Real datasets

5.2.1. DLBCL (2003)

For DLBCL 2003, the best performance is for the Semi-supervised Cox model among the four survival models with concordance index of 0.82 and Integrated brier score of 0.06. The number

of selected parameters is the least in the semi-cox model. After that the Semi-supervised AFT, the single cox and single AFT. Comparing the results of our study with Liang et al. [16], we observe that the most similar results are obtained for Semi-Supervised AFT. This is also seen in the simulation study that the most similar results were for Semi-AFT in the simulation study also. The results of our Semi-Supervised Cox model are better. The CI = 0.82 was seen in our study compared with 0.7 in previous study. The integrated brier score was IBS = 0.06 in our study which is better than previous study (0.09).

5.2.2. AML dataset

For the AML dataset, looking at the obtained results, considering only concordance index as the measure for evaluation, the best model is semi-supervised Cox which has CI = 0.81. But considering both measures of IBS we can say that Semi-supervised AFT has the lowest IBS = 0.11. The least number of selected parameters was seen in the semi-supervised cox model (9 parameters).

Although the data of AML used in this study was not the same with previous one but the results were compared and it was seen that again the results of Semi-supervised AFT are close together in both studies. Just the results are a bit better in our study. Comparing the Semi-cox of our study with previous, we can see that CI is higher for our study (0.81 > 0.68) but the IBS is better in the previous study (0.2 > 0.13).

5.2.3. Pancreas dataset

For the Pancreas data, among four survival models, the Semi-supervised AFT had the best results with CI = 0.85 and IBS = 0.14. After that the Semi-Supervised Cox with CI = 0.83 and IBS = 0.18. Then Single AFT and single cox. The least selected parameters again are for semi-cox with 7 selected parameters. In semi-AFT 21 parameters were selected. The best performance was seen for Semi-Supervised AFT in the Pancreas datasets.

5.3. Overall discussion and conclusions

The Semi-supervised Cox model implemented in this study was done by using Modified Newton Raphson method and coordinate descent to minimize the loss function. The results we get by this method were best in simulation study and also in 2 out of 3 Real datasets. We have done 1 iteration for semi-supervised learning methods. The fraction of censoring left in our study is more. It means that we did not do many iterations to impute the censoring observations as much as possible. In our results although we get slightly more parameters in simulation compared with previous study, we found higher number of correctly the prognostic genes in our study. The results of semi-supervised cox were seen to be better, in both simulation study and the real dataset. The only real datasets which were exactly the same as in Liang et al. [16] are DLBCL2003 and the results of semi-cox are better in both the measures of CI and IBS in our study.

For the Semi-Supervised AFT we have used the same loss function and minimizing algorithm as the previous study and as we see the results of both simulation and real dataset is close to the previous study.

For single Cox and Single AFT we have found the number of correctly selected parameters

better compared with previous study but with having less precision which means the total number of selected parameters were also more in our study. In terms of average correctly selected parameters single AFT is better than single cox, but in terms of precision single cox is better than single AFT.

5.4. Suggestions for future study

In this study we have done Semi-supervised learning using modified Newton-Raphson with coordinate descent algorithm. It was seen that the model performed better than previous study in terms of correctly selecting parameters. In our results we get more correctly the prognostic genes with using 1 iteration in semi-supervised learning. However, the precision in simulation study was lower compared with previous study. Also, the fraction of observations which remained as censored in our method is more compared with previous study. As a suggestion for future work, more iterations of semi-supervised learning can be done to see whether it is possible to keep the number of correctly selected genes high and also leading to increase in precision or not. Probably increasing the iterations will reduce the fraction of remaining as censored too. But this should not impact and cause to miss the prognostic genes. Further study can be done for improving the precision while avoiding to miss the prognostic genes.

Another type of distribution like Weibull, Gamma, Inverse Gaussian and log-logistic distribution, for the AFT model also can be tried with maximizing the likelihood. In our study we have used log-normal distribution.

References

[1] Cox, D. R. (1975) ‘Partial likelihood’, Biometrika. doi: 10.1093/biomet/62.2.269.

[2] Bair, E. and Tibshirani, R. (2004) ‘Semi-supervised methods to predict patient survival from gene expression data’, PLoS Biology. doi: 10.1371/journal.pbio.0020108.

[3] Malik, S. S. et al. (2019) ‘Survival analysis of breast cancer patients with different treatments: A multicentric clinicopathological study’, Journal of the Pakistan Medical Association, 69(7).

[4] Wei, L. J. (1992) ‘The accelerated failure time model: A useful alternative to the cox regression model in survival analysis’, Statistics in Medicine. doi: 10.1002/sim.4780111409.

[5] Wasito, I. and Veritawati, I. (2012) ‘Subtype of cancer identification for patient survival prediction using semi supervised method’, Journal of Convergence Information Technology. doi: 10.4156/jcit.vol7.issue14.25.

[6] Chapelle, O., Sindhwani, V. and Keerthi, S. S. (2008) ‘Optimization techniques for semi-supervised support vector machines’, Journal of Machine Learning Research.

[7] Tibshirani, R. et al. (2002) ‘Diagnosis of multiple cancer types by shrunken centroids of gene expression’, Proceedings of the National Academy of Sciences of the United States of America. doi: 10.1073/pnas.082099299.

[8] Shankar Prinja , Nidhi Gupta and Ramesh Verma Censoring in clinical trials: review of survival analysis techniques.

[9] Wang, Z. and Wang, C. Y. (2010) ‘Buckley-James boosting for survival analysis with high-dimensional biomarker data’, Statistical Applications in Genetics and Molecular Biology. doi: 10.2202/1544-6115.1550.

[10] Xia, Z. et al. (2010) ‘Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces’, BMC Systems Biology. doi: 10.1186/1752-0509-4-6.

[11] Patients., [1] Jorunal. erceived exertion/pain and attention allocation in healthy subjects and chronic pain et al. (2009) ‘Effects of single-task versus dual-task training on bsalance performance in older adults: a double-blind, randomized controlled trial’, Archives of Physical Medicine and Rehabilitation.

[12] Koestler, D. C. et al. (2010) ‘Semi-supervised recursively partitioned mixture models for identifying cancer subtypes’, Bioinformatics. doi: 10.1093/bioinformatics/btq470.

[13] Huang, J., Ma, S. and Xie, H. (2006) ‘Regularized estimation in the accelerated failure time model with high-dimensional covariates’, Biometrics. doi: 10.1111/j.1541-0420.2006.00562.x.

[14] Golub, T. R. et al. (1999) ‘Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring’, Science. doi: 10.1126/science.286.5439.531.

[15] Korenberg, M. J. (2002) ‘Prediction of treatment response using gene expression profiles’, Journal of Proteome Research. doi: 10.1021/pr015510m.

[16] Liang, Y., Chai, H., Liu, XY. et al. (2016). Cancer survival analysis using semi-supervised learning method based on Cox and AFT models with $L_{1/2}$ regularization. BMC Med Genomics 9, 11.

[17] Walter, Fiona M., et al. “Evaluating Diagnostic Strategies for Early Detection of Cancer: The CanTest Framework.” BMC Cancer, vol. 19, no. 1, 2019, doi:10.1186/s12885-019-5746-6.

[18] Chai, Hua, Yong Liang, et al. “A Novel Logistic Regression Model Combining Semi-Supervised Learning and Active Learning for Disease Classification.” Scientific Reports, vol. 8, no. 1, 2018, pp. 2–3, doi:10.1038/s41598-018-31395-5.

[19] R. Saikia and M. Pratim Barman, “A Review on Accelerated Failure Time Models,” Int. J. Stat. Syst., vol. 12, no. 2, pp. 311–322, 2017.

[20] S. Choi and H. Cho, “Accelerated failure time models for the analysis of competing risks,” J. Korean Stat. Soc., vol. 48, no. 3, pp. 315–326, 2019.

[21] Chai, Hua, Zi Na Li, et al. “A New Semi-Supervised Learning Model Combined with Cox and SP-AFT Models in Cancer Survival Analysis.” Scientific Reports, vol. 7, no. 1, 2017, doi:10.1038/s41598-017-13133-5.

[22] D. N. Nawumbeni, A. Luguterah, and T. Adampah, “Performance of Cox Proportional Hazard and Accelerated Failure Time Models in the Analysis of HIV/TB Co-infection Survival Data,” vol. 4, no. 21, pp. 2225-0484, 2014.

[23] Shen, Haiwei, et al. “Robust Sparse Accelerated Failure Time Model for Survival Analysis.” Technology and Health Care, vol. 26, no. S1, 2018, pp. S55–63, doi:10.3233/THC-174141.

[24] Gui, Jiang, and Hongzhe Li. “Penalized Cox Regression Analysis in the High-Dimensional and Low-Sample Size Settings, with Applications to Microarray Gene Expression Data.” Bioinformatics, vol. 21, no. 13, 2005, pp. 3001–08, doi:10.1093/bioinformatics/bti422.

[25] T. Dey, A. Mukherjee, and S. Chakraborty, “A Practical Overview and Reporting Strategies for Statistical Analysis of Survival Studies,” Chest, vol. 158, no. 1, pp. S39–S48, 2020.

[26] Shih, Jia Han, and Takeshi Emura. “Penalized Cox Regression with a Five-Parameter Spline Model.” Communications in Statistics - Theory and Methods, 2020, doi:10.1080/03610926.2020.1772305.

[27] K. Urbanska, J. Sokolowska, M. Szmidt, and P. Sysa, “Glioblastoma multiforme - An overview,” Wspolczesna Onkol., vol. 18, no. 5, pp. 307–312, 2014.

[28] Yamaguchi, Masayoshi, and View Affiliations. Of Molecular Medicine. Vol. 1, no. I, 2013, pp. 1–2.

[29] A. Zare, M. Hosseini, M. Mahmoodi, K. Mohammad, H. Zeraati, and K. Holakouie Naieni, "A comparison between accelerated failure-time and cox proportional hazard models in analyzing the survival of gastric cancer patients," Iran. J. Public Health, vol. 44, no. 8, pp. 1095–1102, 2015.

[30] Rosenwald A, et al. (2003). The proliferation gene expression signature is aquantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. Cancer Cell. 3:185–97. DLBCL (2003) dataset: http://llmpp.nih.gov/mcl.

[31] AML dataset: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?view=data&acc=GSE425&id=21974&db=GeoDb_blob01.

[32] Pancreas dataset: https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE34153.

[33] Bender R, Augustin T, Blettner M. (2005). Generating survival times to simulate Cox proportional hazards models. 24:1713–23.

[34] Cox, D.R (1972). : Regression Models and Life Tables, Journal of the Royal Statistical Society, Series B, (Methodological),Vol 34, No. 2, 187-220.

[35] Xu ZB, et al. (2010). L1/2 regularization. Sci China. 40(3):1–11. series F.

[36] Kaplan, E.L. and Meier, P. (1958) Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association, 53, 457-481. http://dx.doi.org/10.1080/01621459.1958.10501452.

[37] A. El-Gohary, Ahmad Alshamrani, Adel Naif Al-Otaibi. The generalized Gompertz distribution.