Clustering Survival Data using Random Forest and Persistent Homology

by

Berhanu Anagaw Wubie

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Biostatistics

Department of Mathematical and Statistical Sciences
University of Alberta

Abstract

Survival data is mostly analyzed using Cox proportional hazards model to identify factors associated with survival time of patients. However recently random survival forest (RSF), a non-parametric method for ensemble estimation constructed by bagging of classification trees for survival data, is used as an alternative method for better survival prediction and ranking the importance of covariates associated with it. In addition to identification of variable importance for survival prediction, exploring clusters in survival data using the variables identified as important in RSF analysis were applied.

Clustering survival data (patients) to assess their survival experience was investigated using random forest clustering based on partitioning around the medoids and persistent homology(PH), a topological data analysis (TDA) technique for cluster identification in lower dimension (dimension zero). In both methods, we were able to identify different groups of patients possessing different survival experience accounting for those covariates most important in determining survival experience. The clusters formed were assessed for significant difference in their survival experience (log-rank test) and were found to have difference in survival experience between them. Further investigation was applied using PH to explore more detailed characteristic features of patients at higher dimension (dimension one). Both clustering methods result in a promising exploration of groups within patients that will give insight into to patient handling and give valuable information in providing quality service to patients who need more attention. All analysis procedures in this thesis were done using two datasets: the kidney and liver dataset.

Acknowledgements

First of all, I would like to thank Almighty God for his savior, without whom none of this would have been possible and who has given me the opportunity to go through this way and reach this success.

My heartfelt gratitude goes to Prof. Giseon Heo, my supervisor, for her immense and invaluable contribution interms of constant guidance and advice in the best way of handling this work and tireless efforts to make it a reality. My sincere appreciation and thanks also go to Department of Mathematical and Statistical Sciences academic and administrative staff members, particularly to Prof. Jochen Kuttler and Ms. Tara Schuetz-Zawaduk for their welcoming and assistance whenever needed. I want to thank University of California (UCLA), Department of Biostatistics, Scientific Registry of Transplant Receipts (SRTR): Prof. Russ Greiner, Prof. Andres Axel and Prof. Aldo Montano-Loza for giving us permission to use their data. I also would like to thank professors in the Department of Statistics for their unreserved knowledge sharing and cooperation, especially to Prof. Linglong Kong and Prof. Keumhee C. Chough. My appreciation also goes to Prof. Heo's research group: Matthew Pietrosanu and Steven Luoma for their support. To my friends and officemates Birtukan and Box, thanks a lot.

At last but not least, it is my deepest and warmest gratitude to my parents and siblings abroad for their prayer, encouragement and giving me critical and constructive advice to complete my study.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Survival data has traditionally been analyzed using semi-parametric Cox proportional hazards model, a method most widely used to study time-to-event data with censoring, or parametric survival regression methods. The parametric, semi-parametric and non-parametric methods are quite useful as they are very simple to make inference and to interpret the effects of the covariates. Despite its simplicity, these models need some specified link function that associates the dependent variable with the covariates in the model. The analytic techniques used with these methods entirely depend on model assumptions and the survival data should satisfy these assumptions. The model development process in survival analysis applies variable selection procedures like stepwise methods and also considers interaction terms. However, in model building process it is difficult to identify which interaction terms to include and is left for the investigator that needs subject matter knowledge.

Recently, however, a number of advanced and more powerful techniques incorporating machine learning have become relevant to high dimensional data. When the investigator does not want to apply modeling techniques with their respective assumptions, these methods are alternative options to do so. Machine learning techniques reduce potential risk to the misspecification of model, which leads to inconsistent

estimators and invalidity of results. These new techniques mark an improvement over traditional methods in both precision and robustness by identifying prognostic factors and higher-order interactions among features and an alternative way to build a good exploratory and prediction model. As a result, survival data can be assessed using these machine learning methods. Machine learning techniques are non-parametric techniques that can deal with classification and clustering of high dimensional data. Machine learning techniques used for classification and clustering are collectively known as supervised and unsupervised machine learning techniques, respectively.

Random forest is one of those techniques used for classification and clustering of high dimensional data. It is highly applicable for classifying data into groups of similar character. Compared to the standard methods, it also has the power of detecting automatically which interaction terms are sufficient in classifying data into groups. Besides, high dimensional data can be further dealt on most recent and expanding technique known as computational topology. Persistent homology, a special technique developed in computational topology to identify connected components, which have similar character and will last long as a group forming different geometrical shapes. It also helps us to compare the patterns forming clusters. Persistent homology is also important in discerning true features from noise in data and helps in formulating and testing a hypothesis so as to make valid inference.

1.1 Motivation of the Study

Standard statistical methods, such as Kaplan-Meier, Cox PH, Exponential, Weibull, etc., are the most widely used techniques to deal with survival data analysis in the previous years, however, recently there are other fascinating techniques which lead us better understand survival data. Therefore, this study is motivated to investigate, explore and better understand survival data through one of those machine learning

techniques, random forest and extend the application of persistent homology in identifying group of clusters so as to explore and understand survival experience of patients across the different groups. Moreover, random forest and computational topology clustering investigation will be applied using the following two datasets: renal cell carcinoma cancer and liver transplant for alcoholic patients datasets.

1.2 The Datasets

The following two datasets will be used throughout the thesis to illustrate some of the methods and applications proposed in this work.

1.2.1 The Kidney Data - Renal Cell Carcinoma Cancer

Cancer starts when cells in the body begin to grow out of control. Kidney cancer is one of the most widely known types of cancers and affects kidneys. Cells in nearly any part of the body can become cancerous, and can spread to other areas of the body. Renal cell carcinoma (RCC), also known as renal cell cancer or renal cell adenocarcinoma, is a kidney cancer, which is most common type of cancer and has the highest in occurrence among different kidney cancers. About 9 out of 10 kidney cancers are of this type. There are several subtypes of RCC, based mainly on how the cancer cells look under a microscope. Knowing the subtype of RCC can be a factor in deciding treatment and can also help doctors determine if patient's kidney cancer case might be due to an inherited genetic syndrome [1][4].

The American Cancer Society's most recent estimates for kidney cancer in the United States are for 2016: About 62,700 new cases of kidney cancer (39,650 in men and 23,050 in women) will occur. About 14,240 people (9,240 men and 5,000 women) will die from this disease. These numbers include all types of kidney and renal pelvis cancers. Most people diagnosed with kidney cancer are older people. The

average age of people when they are diagnosed is 64. Kidney cancer is very uncommon in people younger than age 45. Kidney cancer is among the 10 most common cancers in both men and women. Overall, the lifetime risk for developing kidney cancer is about 1 in 63 (1.6%) [1]. This risk is higher in men than in women. There are several DNA microarrays used to investigate the gene expression and molecular tumor markers of RCC and this help in identifying the diagnosis and prognosis of the cancer [2][3].

The data used in this study assessed eight tumor markers associated with renal cell carcinoma. These eight markers that are classified as molecular properties: tumor proliferation, cell cycle abnormalities, cell mobility and hypoxia pathway are associated with survival of renal cell carcinoma [2][4]. For this study a sample of 366 patients who underwent a radical or partial nephrectomy for renal cell carcinoma at University of California (UCLA) in the period 1989 to 2000 were considered [3]. The average age of the patients were reported as 60 years and female patients are found to be half of that of the males. In the study we group RCC patients in to different groups using random forest and persistent homology clustering techniques, that will result in clinically and biologically meaningful class of patient groups with different state of renal cell carcinoma cancer[3].

1.2.2 The Liver Data - Liver Transplant Failure Data

Liver is the second most commonly transplanted organ next to kidney. Human body needs a healthy liver so as to give the proper function to our body and is known as the powerhouse that produces various substances that our body needs highly such as glucose, proteins, blood-clotting substances to heal wound and for the production of an important fluid to absorb fats, minerals and vitamins. A liver also functions as a filter in removing impurities from blood and detoxify harmful substances produced in our body. Liver disease occurs when these essential functions are disrupted and

unable to work properly. In this case the patient needs to have liver transplant when the disease damage the liver severely impairs a person's health and quality of life [7].

As reported in 2014, 6729 liver transplants were performed in adults. These included 6449 organs from deceased donors and 280 from living donors, which shows an increase in need for transplant compared to previous years. Waitlist mortality remained a concern; in 2014, 1821 patients died while waiting for a transplant and another 1290 were removed from the list due to being too sick to undergo transplant. The increase in the need of liver transplant is due to increase in liver disease such as Hepatitis C virus (HCV) infection and hepatocellular carcinoma (HCC), were the dominant indicators for liver transplant in 2014. According to the report nearly 72,000 adults were living with a functioning liver graft. However, not all patients live good quality life due to post transplant complications. In 2014, about 35.7% of transplant recipients had been hospitalized, including 15.7% in the intensive care unit (ICU) due to complications and this reflect severity of liver disease in many liver transplant recipients [5][6][7].

In our work we considered a total of 8361 alcoholic patients who received liver transplant and are under follow-up to assess and classify into groups based on their quality of life post transplant. To do so, we considered different clinical and demographic factors associated with failure of the transplant in alcoholic patients. Although there are many clinical, demographic and socio-economic factors associated to transplant failure (death of a patient) in this study, we considered those factors associated with time to death of patients after transplant. The demographic factors considered are recipient's gender, age, blood group, height and weight and the clinical factors creatinine, bilirubin, albumin levels and the transplant factor cold ischemic time of the organ (Cold_isch) (defined in the appendix). In addition to recipient's characteristics we include donor's characteristics such as blood type, age, weight, height, gender

and donor type (whether the donor is deceased or living).

1.3 Thesis Outline

The organization of the thesis is as follows. In chapter 2, we discuss the standard Cox proportional hazards methods, random survival forest and persistent homology and literatures related to model prediction. In chapter 3, we apply the proposed methods in chapter 2: standard Cox and random survival forest, for the variable selection and prediction of survival using the two datasets. In chapter 4, we will discuss the implementation of data clustering using random forest and persistent homology. Chapter 5 will conclude the overall thesis with its discussion of the previous results, limitations and suggestions for future work.

Chapter 2

Review of Methodology

2.1 Survival Analysis

Survival analysis is a statistical method designed to study the amount of time an experimental unit survives, or the study of time between entry into observation and a subsequent event. The statistical approach to be used in this study is the analysis of time-to-event data, which are related with individual time elapse in certain situation or state. As the uses of survival analysis grew, parametric models gave way to nonparametric and semi-parametric approaches for their appeal in dealing with the ever-growing field of clinical trials in medical research. Survival analysis consists of a set of specialized statistical techniques used to study response time data. In analyzing such data, the main objects are to determine the length of time interval spent in a state and the transition probabilities from the current state to the entered state. The interest of this statistical tool is mainly focused on two distinguishing features of time to event data. Primarily, duration times are non-negative values usually exhibiting highly skewed distribution and therefore assumption of normality may be violated. Secondly, censoring may occur or the true duration is not always observed or known, that is, some subjects potentially being unobserved for the full

time to failure [9].

The main characteristic of these data is the issue of censoring which occurs when the periods of time of event occurrence for some individuals cannot be completely observed. The process of censoring makes these data unsuitable to analyze with traditional regression method and hence, the appropriate technique and analysis procedure usually called Survival Analysis helps in handling this condition. Details on various estimation methods developed in survival data analysis taken censoring into account can be obtained in Hosmer and Lemeshow (2008) [8]. As mentioned before, censoring is said to occur when the end-point of interest has not been observed at end of data collection. It occurs, for example, when some patients survive to the end of the trial investigating time to death; when a certain type of cancer does not occur again after surgical removal; when a patient has died from an unrelated cause to the one being investigated; and when a patient is lost to follow-up.

Censoring

The time period confinement for survival data gives rise to considerations specific to survival analysis, censoring. A censored observation is one whose value is incomplete due to random factors for each subject. Censoring can appear in various forms and the most common form is explained below:

Right Censoring: The most common form of incomplete data is right censoring. An observation is said to be right censored if it is recorded from its beginning until a well defined time before its end time. For instance, if our study objective is to assess the time to failure of organ for patients who received transplant, then a patient is said to be right censored, if the transplanted organ is functioning well without experiencing this scenario until the end the study period. In other words, an observation is said to be right censored if follow up of the study begins at time $t = 0$ and terminate before the outcome of interest is observed on the patient.

2.1.1 Descriptive Methods for Survival Data

In any applied setting, a statistical analysis should begin with a thoughtful and thorough univariate description of the data. And this description includes life table and Kaplan-Meier survival function estimation that are used for the estimation of the distribution of survival time from all observations available.

The Survival Function

The cumulative distribution function (cdf) is very useful in describing the continuous probability distribution of a random variable, such as time, in a survival analysis. The cdf of a random variable T, denoted $F(t)$, is defined by $F(t) = P(T \leq t)$. The survival function is defined as the probability of a subject at risk surviving beyond time t. Let $T \geq 0$ have a pdf $f(t)$ and cdf $F(t)$. Then the survival function takes on the following form,

$$S(t) = P\{T > t\} = 1 - F(t).$$

That is, the survival function gives the probability of surviving or being event-free beyond time t. Because $S(t)$ is a probability, it is non-negative and ranges from 0 to 1. It is defined as $S(0) = 1$ and as t approaches ∞, $S(t)$ approaches 0.

Median Survival Time: Median survival time m is defined as the quantity satisfying $S(m) = 0.5$. Sometimes denoted by $t_{0.5}$. If S(t) is not strictly decreasing, m is the smallest one such that $S(m) = 0.5$ or $t_{0.5} = S^{-1}(0.5)$.

The Hazard Function

The hazard function is also known as failure rate, force of mortality, conditional failure rate or simply hazard rate and it is defined as the probability that an individual fails at time t, conditioned on the fact that he or she has survived to that time. It therefore, represents the instantaneous failure rate for an individual surviving to

time t. For $h(t) \geq 0$, the hazard function $h(t)$ is given by the following:

$$
\begin{aligned}
h(t) &= lim_{\Delta t \to 0} \frac{p\{an \quad individual \quad fails \quad in \quad the \quad time \quad interval \quad (t, t+\Delta t | alive \quad at \quad t)\}}{\Delta t} \\
&= lim_{\Delta t \to 0} \frac{p\{t \leq T \leq t+\Delta t | T \geq t\}}{\Delta t} \\
&= P\{t < T < (t+\Delta t) | T > t\} \\
&= \frac{f(t)}{(1-F(t))} \\
&= \frac{f(t)}{S(t)}.
\end{aligned}
$$

The hazard function describes the concept of the risk of an outcome (e.g., death, failure, hospitalization) in an interval after time t, condition on the subject having survived to time t. The hazard function seems to be more intuitive to use in survival analysis than the pdf because it attempts to quantify the instantaneous risk that an event will take place at time t given that the subject survived to time t.

2.1.2 Non-parametric Methods in Survival Data

Kaplan-Meier Survival Function

The Kaplan-Meier (KM) estimator, or product limit estimator, is the estimator used by most software packages. The KM estimator incorporates information from all of the observations available, both censored and uncensored, by considering any point in time as a series of steps defined by the observed survival and censored times. When there is no censoring, the estimator is simply the sample proportion of observations with event times greater than t. The technique becomes a little more complicated but still manageable when censored times are included.

The KM estimator is a nonparametric estimator of the survivor function $S(t)$.

$$\hat{S}(t) = \prod_{t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j}\right), \tag{2.1.1}$$

where $t_{(j)}$ is the ordered event or failure times, d_j is the number of individuals who experience the event or failure at time $t_{(j)}$, and n_j is the number of individuals who have not yet experienced the event at that time and are therefore still at risk for experiencing it. The Kaplan-Meier estimator (2.1.1) is a step function with jumps at the observed event times. The size of the jump at a certain event time $t_{(j)}$ depends on the number of events observed at $t_{(j)}$, as well as on the pattern of the censored event times before $t_{(j)}$.

Comparison of Survival Curves

In clinical research one is concerned not only with estimating the survival function but, more often, with the comparison of the life experience of two or more groups of subjects differing for a given characteristic or randomly allocated to different treatments. After providing a description of the overall survival experience in the study, we usually turn our attention to a comparison of the survivorship experience in key subjects in the data. The simplest way of comparing the survival times obtained from two or more groups is to plot the Kaplan-Meier curves for these groups on the same graph. However, this graph does not allow us to say, with any confidence, whether or not there is a real difference between the groups. The observed difference may be a true difference, but alternatively, it could also be due merely to chance variation. Assessing whether or not there is a real difference between groups can only be done, with any degree of confidence, by utilizing statistical tests.

Since survival data are typically right skewed, we would likely use rank-based non-parametric tests followed by estimates and confidence intervals of the medians or other quantiles within groups. Modifications of these procedures are required when censored observations are present in the data. When we compare groups of subjects, it is good to begin with a graphical display of the data in each group. Among the various non-parametric tests one can find in the statistical literature, the Mantel-Haenzel (1959) test, commonly called the "log-rank" test will be used. Nowadays

the Kaplan-Meier method for estimating survival curves and the log-rank test for comparing two estimated survival curves are the most frequently used statistical tools in medical reports on survival data [8][9].

Log-rank Test

The log-rank test, developed by Mantel and Haenszel, is a non-parametric test for comparing two or more independent survival curves. Since it is a non-parametric test, no assumptions about the distributional form of the data need to be made. This test is however most powerful when used for non-overlapping survival curves. This test can be generalized to accommodate other tests that are alternatively used sometime in practice such as Generalized Wilcoxon test, Tarone-Ware test, and Peto-Peto Prentice test. Each of these tests uses different weight to adjust for censoring that is often encountered in survival data. The log-rank test statistic for comparing two groups is given by:

$$L = \frac{\left[\sum_{i=1}^{m} (d_{1i} - \hat{e}_{1i})\right]^2}{\sum_{i=1}^{m} \hat{v}_{1i}}$$

where:

m is the number of rank ordered event times.

d_{1i} is the observed number of events in group 1 at event time t_i.

n_{1i} is the number of individuals at risk in group 1 just prior to event time t_i.

d_i is the observed number of events in both group 1 and group 2 at event time t_i.

n_i is the number of individuals at risk in both group 1 and group 2 just prior to event time t_i.

$\hat{e}_{1i} = \dfrac{n_{1i} d_i}{n_i}$ is the expected number of events corresponding to d_i.

$\hat{v}_{1i} = \dfrac{n_{1i} n_{2i} d_i \left(n_i - d_i\right)}{n_i^2 \left(n_i - 1\right)}$ is the variance of the number of events d_{1i} at time t_i.

The log-rank test statistic L has an approximation of chi-square distribution with one degree of freedom for large samples. The null hypothesis of equality of survival functions will be rejected for large values of L. The log-rank test can be extended for comparing three or more groups of survival experience [8][9].

2.1.3 Regression Models for Survival Data

In most medical studies that give rise to survival data, supplementary information is collected on each individual so that the relationship between the survival experience of individuals and various explanatory variables may be investigated. A variety of models and methods have been developed for doing this sort of survival analysis using either parametric or semi-parametric approaches. Semi-parametric models are models that parametrically specify the functional relationship between the lifetime of an individual and its characteristics (demographic, socio-economic, etc.) but leave the actual distribution of lifetimes arbitrary. The most popular of the semi-parametric models is the Proportional Hazards model. It has the property that the ratio of the hazards depends on the values of their explanatory variables, say X_1, X_2, . . . , but does not depend on time t.

The Semi-Parametric Cox-Proportional Hazards Model

We can specify the density function of a parametric distribution or we can specify the hazard function. The advantage of the latter approach is that we directly address the aging process, but as shown previously, it does not easily lend to itself to the use of scatter plots to motivate regression models [11]. The latter approach may also be preferred in a setting where the end products of the statistical analysis are estimated parameters that compare the survival experience of the selected subgroups. By specifying a model through the hazard function, we may address specific questions such

as how survival is related to the subject's characteristics or the covariates.

Cox's (1972) paper took a different approach to standard parametric survival analysis and extended the methods of the non-parametric Kaplan-Meier estimates to regression type arguments for life-table analyses. Cox advanced to prediction of survival time in individual subjects by only utilizing variables that co-vary with survival and ignoring the baseline hazard of individuals. He did this by making no assumptions about the baseline hazard of individuals and only assumed that the hazard functions of different individuals remained proportional and constant over time. When there are several explanatory variables, and in particular when some of these are continuous, it is much more useful to use a regression method such as Cox rather than a KM approach [8][11].

Cox introduced the semi-parametric proportional hazards model to account for covariate effects for single event times. This model is valid under the assumption of proportional hazards. Cox (1972) observed that if proportional hazards assumption holds then it is possible to estimate the effect parameter(s) without any consideration of the hazard function. There are several reasons in which Cox's proportional hazards modelling was chosen to explain the effect of covariates on time until event. They are discussed below and include: the relative risk, no parametric assumptions, hazard function, the use of the partial likelihood function, and the estimates of survivor function.

The Hazard Function

The data in survival analysis based on the sample size n, consists of (t_i, δ_i, X_i), $i = 1, 2, \ldots, n$, where t_i is the time on the study for the i^{th} individual, δ_i is the event indicator ($\delta_i = 1$ if the event has occurred and $\delta_i = 0$ if it is censored (the lifetime may be right, left or interval censored), and X_i is the vector of covariates or the risk factors for the i^{th} individual that may affect for instance the time to full functioning state of transplanted organ [8][9]. The Cox proportional hazards model is generally

given by:

$$h\left(t, X_i, \beta\right) = h_0\left(t\right) exp\left(\ \beta' X_i\right), \tag{2.1.2}$$

where $h_0\left(t\right)$ is the baseline hazard function at time t, $X_i{}' = (X_{1i}, X_{2i}, ..., X_{ki})$ for $i = 1, 2, \ldots, n$ is a vector of measured covariates for the i^{th} individual at time t, and β' is a $1 \times k$ vector of unknown regression parameters that are assumed to be the same for all individuals in the study, which measures the influence of the covariate on the survival experience with β_i representing the increase in the log hazards as X_i increases one unit relative to the baseline hazard function. This model is referred to in the literature by a variety of terms, such as the Cox model, the Cox proportional hazards model or simply the proportional hazards model. The hazard function in equation (2.1.2) depends on both time and the associated covariates, but through two separate factors: the first is a function of time only, which is left arbitrary, but is assumed to be the same for all the subjects, the second is a quantity that depends on the individual covariates.

From the representation in equation (2.1.2) one can notice a couple of features. First, if $X_i = 0$ then the hazard function for the i^{th} individual is the baseline hazard function. It is the hazard function in the absence of covariates or when all of the coefficients of the covariates are assumed to be zero. Second, if we divide both sides by $h_0\left(t\right)$, we get equation:

$$\frac{h_i\left(t, X_i\right)}{h_0\left(t, 0\right)} = \frac{h_0\left(t\right) exp\left(\beta' X_i\right)}{h_0\left(t\right)} = e^{\beta' X_i}, \tag{2.1.3}$$

which shows where the term proportional comes from. Since for each individual, $e^{(\ X_i{}'\beta)}$ is constant across time, equation (2.1.4) shows that at every value of t, the i^{th} individual's log hazard function is constant proportion of the baseline hazard. Very loosely speaking, this implies that each individual's hazard function is "parallel" to

the $h_0(t)$.

The Cox model is often called proportional hazards model because, if we look at two independent subjects with covariate values X_1 and X_2, the ratio of their hazard functions at time t is:

$$\frac{h(t, X_1)}{h(t, X_2)} = \frac{h_0(t) exp(\beta' X_1)}{h_0(t) exp(\beta' X_2)} = exp\left[\beta'(X_1 - X_2)\right], \tag{2.1.4}$$

which is constant and does not vary over time, that is, the ratio does not depend on t and the hazard rates are proportional. The Cox proportional hazards model can equally be regarded a linear model, as a linear combination of the covariates for the logarithm transformation of the hazard ratio given by:

$$\log\left\{\frac{h(t, X)}{h_0(t)}\right\} = \beta' X \tag{2.1.5}$$

Note that the cumulative hazard function is given by:

$$H(t) = H_0(t) exp\left(\beta' X\right) \tag{2.1.6}$$

Consequently, from the proportional hazard function, we obtained the survivor function given by:

$$S(t, X, \beta) = [S_0(t)]^{exp(\beta' X)}, \tag{2.1.7}$$

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

2.1.4 Estimation of Parameters using Partial Likelihood

Since $h_0(t)$ is not specified parametrically, it is not possible to use an ordinary likelihood to estimate the regression coefficients β. The arbitrary function $h_0(t)$ is a nuisance function, and the aim is to estimate β on the basis of the information

conveyed by the observed data without having to involve $h_0(t)$. Cox (1972) argued conditionally on the set of observed failures and described the data with a function depending on β only. Consider a sample of n subjects and suppose a total of m failures occur, with m generally smaller than n, due to the presence of censoring. Let $t_1 < t_2 < ... < t_m$ be the m distinct ordered failure times observed and let $R(t)$ be the set of subjects, at risk at time t, who are not failed and under observation just before t. With a slight change of notation, we indicate with j the label of the subject who fails at t_j so that its vector of covariates is X_j. In general, X_i the vectors of covariates for the i^{th} subject and the covariates have a constant value in time. The probability that an individual with covariates X fails in the small interval $(t + \Delta t)$, given the set at risk at t, is:

$$\frac{h\left(t_{(j)}, X_j\right)\Delta t}{\sum_{i\epsilon R\left(t_{(j)}\right)} h\left(t_{(j)}, X_i\right)\Delta t}$$

It follows that the function describing the failure pattern is the product of m terms, one for each observed failure time.

$$L\left(h_0(t), \beta\right) = \prod_{j=1}^{m} \frac{h\left(t_{(j)}, X_j\right)\Delta t}{\sum_{i\epsilon R_j} h\left(t_{(j)}, X_i\right)\Delta t}.$$

Where the hazard function is defined by (2.1.2) and R_j=$R\left(t_{(j)}\right)$. Given expression (2.1.2), the baseline function $h_0(t)\Delta t$ cancels out and the product above simplifies to:

$$L = L(\beta) = \prod_{j=1}^{m} \frac{exp\left(\beta' X_j\right)}{\sum_{i\epsilon R_j} exp\left(\beta' X_i\right)} \tag{2.1.8}$$

where $L(\beta)$ in equation (2.1.8) depends on the unknown parameters β is referred to as the partial likelihood.

The partial likelihood given by equation (2.1.8), although it describes only part of

the data, could be regarded as a likelihood function allowing the estimation of β with standard procedures. In general, large sample properties like normality and consistency of maximum likelihood estimators of β based on partial likelihood have been shown to be the same as those of any estimator from complete likelihood [10][11].

The asymptomatic theory of maximum likelihood estimation requires that the likelihood function satisfies some "regularity conditions" which are met in most applications. The regression coefficients β are estimated by the values $\hat{\beta}$, which maximize the partial likelihood $L\left(\hat{\beta}\right)$ or $LL\left(\hat{\beta}\right)$ equivalently its logarithm $LL\left(\beta\right)$:

$$LL\left(\beta\right) = \sum_{j=1}^{m} \left\{ \beta' X_j - ln \left[\sum_{i \epsilon R_j} exp\left(\beta' X_i\right) \right] \right\} = \sum_{j=1}^{m} l_j$$

Where l_j is the contribution of the log-likelihood corresponding to the failure time $t_{(j)}$. The values $\hat{\beta} = \left(\hat{\beta}_1, ..., \hat{\beta}_k\right)$ are obtained by equating to zero the K first derivatives of log likelihood function with respect to $\beta_k (k = 1, ..., K)$.

In this study to estimate the survival function, we will use the non-parametric Kaplan-Meier or product limit estimation method and for comparisons of survival estimates the Log-rank test will be considered. Moreover, to investigate the effect of factors or covariates on the time to an event in the follow-up study, we consider the Cox's proportional hazards regression model.

2.1.5 Model Building or Model Development

In performing proportional hazards regression analysis for survival data requires a number of critical decisions. It is likely that we will have data on more covariates than we can reasonably expect to include in the model, so we must decide on a method to select a subset of the total number of covariates. When selecting a subset of the covariates, we must consider such issues as clinical importance and statistical

significance.

Before any model could be fitted, it is important to investigate which variable(s) goes into the model by using conventional selection procedure. The methods available to select a subset of the covariates to include in a proportional hazards regression model are essentially the same as those used in the other regression models, like purposeful selection, stepwise (forward selection and backward elimination) and best subsets selection.

In this study, model building starts from univariate analysis as suggested by Collet (1994), Collect [12] recommended the approach of first performing a univariate analysis to "screen" out potentially significant variables for consideration in the multivariate model in order to identify the importance of each predictor. All variables that are significant at 25% level, the modest level of significance for bivariate regression from one explanatory univariate regression model are taken into multivariable model where backward selection approach is used with 10% significant level of stay in the model. Variables that are selected at this stage are taken to stage three of the analysis where variables that are not significant in stage one are added one at a time and forward selection procedure is used with 5% significant level of entry into the model. The fourth stage involves combination of all variables that are significant at stage three in addition with their possible interactions using stepwise selection procedure with 10% significant level of entry and stay in the model and if the interaction is significant, but not the main effect of the covariate, we include both the interaction and the main effect in the final model even if the main effect is not significant. According to the hierarchical principle, if a model contains interaction terms, the corresponding lower order terms should also be included in the model. The final variables selected at this stage are then pruned to have the final model.

2.2 Random Forest

Random forest is a recently developed machine learning technique that deals with classification and clustering of data non-parametrically [13][14]. It is an ensemble method that combines a number of trees by taking the same number of bootstrap samples from the original data, and growing a tree on each bootstrap sample. Tree implementations are very simple and user-friendly and require fewer techniques from the investigator. The individual trees in a random forest are not pruned and used for decision in classification or clustering. Random forest uses a randomly selected subset of predictors for splitting the root nodes in to new daughter nodes for each split. From all trees grown in this process based on the bootstrap samples, we generate a forest. From the complete forest, the response variable for an instance is predicted as an average or majority vote of the predictions of all trees. Random forest can highly increase the prediction accuracy compared to an individual tree, as the ensemble reduces the variance [13][21][22].

Comparing random forest with other standard methods, it has several attractive features. It is highly data adaptive and virtually model assumption free, compared to standard analysis techniques, like the Cox PH model, which often rely on restrictive assumptions. With traditional regression methods, there is always the concern whether association between predictors and the outcome have been modeled appropriately, and whether or not non-linear effects or higher order interactions for predictors should be included. In contrast, such problems are handled automatically with trees and random forest. Furthermore, random forest is known for its better performance for prediction than other methods [15]. However, the drawback of using random forest is its lack of meaningful interpretability compared to the standard methods, which makes it crucial to have reliable variable importance measures derived from random forest. In addition, it is not easy to conduct hypothesis testing with random forest.

Random forest is constructed based on complete randomization techniques starting from selecting bootstrap samples to grow the tree using a random sample of covariates, $mtry$, at each splitting stage. Recently, Guerts *et al.* (2006) [19] proposed a variation of random forest that introduces even more randomization by randomly selecting both splitting variables and cut-points at each node. The algorithm they implemented has both very good accuracy and computational efficiency. However, they did not mention about the variable measures used to grow the tree in their algorithm, as all variables do not have equal importance [15][17][19]. Random forest explanation and its interpretation from statistical point of view is not straightforward like what we know before as it considers different trees for each bootstrap samples. The variable importance measures used at each step are also not simple to understand like in the standard models instead it's a black box, except for the overall forest. Consequently, one of the limitations noted for tree-structured methods is in terms of variable selection; trees give preference to predictors that have more levels or values [17][19]. CART (classification and regression trees; Breiman *et al.*, 1984 [18]) and random forest may favor variables with more categories because of the way the cut-points are chosen. In an effort to select the best split at each node, all possible cut-points of a candidate splitting variable are considered. Variables with more potential cut-points or splitting points are more likely to produce a good splitting score by chance, as in a multiple testing situation [19][22].

2.2.1 Classification and Regression Trees (CART)

A binary tree is an input-output model represented by a tree structure T, from a random input of variables X_1, X_2, . . . , X_p taking its values in $X_1 \times X_2 \times ... \times X_p$ to the output Y. Any node t in the tree represents a subset of the space X, with the root node being itself. Internal nodes t are labeled with a binary test or split $S_t = (X_m < c)$ dividing their subset in to two subsets corresponding to their two

daughters t_L and t_R, while the terminal nodes are labeled with a best guess value of the output variable. The predicted output for a new instance is the label of the leaf reached by the instance when it is propagated through the tree. A tree is built from a learning sample of size N drawn from $P(X_1, X_2, ..., X_p, Y)$ using a recursive procedure, which identifies at each node t the split $s_t = s^*$ for which the partition of N_t node samples into t_L and t_R maximizes the decrease [33]:

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R),$$

of some impurity measure (e.g, the Gini index, entropy of the variance of Y), where $p_L = N_{tL}/N_t$, N_{tL} left node samples and $p_R = N_{tR}/N_t$, N_{tR} right node samples. The construction of the trees stops when the nodes become pure in terms of Y, that when there is no more decrease in impurity of nodes.

2.2.2 Algorithm Description

The algorithm used in random forest proposed by Breiman (1996)[20], is a method that improves its prediction accuracy by decreasing the prediction error over CART based on bagging (bootstrap aggregation) [14]. Random forest based on extremely randomized trees builds an ensemble of trees according to the classic top-down procedure. Its main difference with the standard random forest procedure is that it considers only randomly selected cut-points from each of the randomly selected variables at each internal node [13][15]. The split that provides the best within group similarity will be considered. It needs the following three parameters: m, the number of randomly selected variables used for splitting a tree at each node: d, the number of cut-points randomly chosen for each one of the m selected variables and n_{min}, the minimum threshold number of subjects remain at each node after splitting. The random selection of m covariates and d cut-points at each split not only helps improve computational efficiency [13][14][20] but also selection bias in covariates.

The algorithm can be summarized as [14],

1. For $b = 1$ to B.

   (a) Draw a bootstrap sample of $Z*$ of size N from the training data.

   (b) Grow a random forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.

      i. Select m variables at random from the p variables.

      ii. Pick the best variable/split point among m.

      iii. Split the node into two daughter nodes.

2. Output the ensemble of trees $\{T_b\}_1^B$

To make a prediction at a new point x:

Regression: $\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$

Classification: Let $\hat{C}_b(x)$ be the class prediction of the b^{th} random forest tree. Then $\hat{C}_{rf}^B(x) = majority \quad vote\left\{\hat{C}(x)\right\}_1^B$.

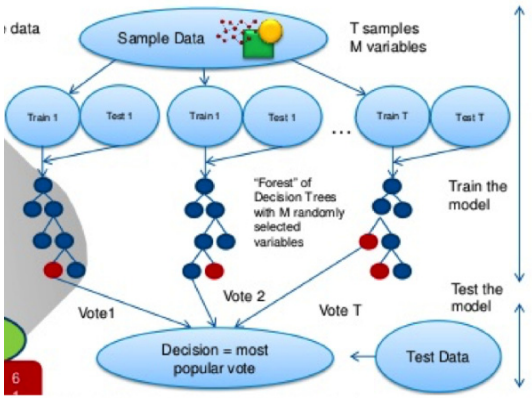


Figure 2.1: A schematic illustration of how random forest classification works: source from internet

In taking the bootstrap sample, some observations are considered more than once to grow the tree while the others remain unused, what is called OOB sample. Ap-

proximately 63% of the observations are used to grow the tree and the remaining 37% of them are not used for the growing the trees. Hence, the data excluded from the bootstrap samples or subsamples are sent down the tree to derive the predicted group membership [14][15][18].

2.2.3 Variable Importance (VIMP)

In random forest, variable importance is very interesting and obvious to asses the relative importance of a bunch of variables used at each step of splitting the data in to groups. In splitting the data into meaningful classes, we have to rely on the predictive importance of variables so as to have better result in assigning predicted class of out of bag (OOB) samples. This importance shows the mean decrease in prediction error that results from randomly permuting an explanatory variable. Random forest uses OOB samples to construct variable importance measure to evaluate the prediction strength of each variable. When the b^{th} tree is grown, the OOB samples are passed down the tree and the prediction accuracy is recorded. Then the values for the j^{th} variable are randomly permuted in the OOB samples and the accuracy is again computed. The decrease in accuracy as a result of this permuting is averaged over all trees and is used as a measure of the variable importance of variable j in the random forest which can be described as follows:

Random forest is constructed using ensembles of randomized trees, Breiman (2001, 2002) [13] proposed to evaluate the importance of the variable X_m for predicting Y by adding up the weighted impurity decreases $p(t)\Delta i\,(s,t)$ for all nodes t where X_m is used and averaged overall N_T trees in the forest, $p(t)$ is the proportion N_t/N of samples reaching t and $v(s_t)$ is the variable used in split s_t;

$$Imp\,(X_m) = \frac{1}{N_T}\sum_{T}\sum_{t\epsilon T: v(s_t)=X_m} p\,(t)\,\Delta i(s_t,t),$$

2.3 Random Survival Forests

Random Survival Forests is an extension of random forest, which is an ensemble tree method for the analysis of right-censored survival data. As is well known, constructing ensembles from base learners, such as trees, can significantly improve learning performance [16][23]. Random Survival Forests (RSF) closely modeled after Breiman's approach. Random Survival Forests naturally inherits many of its good properties from RF. It is user-friendly; fairly robust and the parameters need to grow mature trees are the same as RF. Moreover, it relies on the data and also its derivation is based on data and model assumption free unlike the standard methods used for survival data [16]. RSF an ensemble of forest method known for its consistency that the survival function converges uniformly to the true population survival trend [19][22].

Currently, random forest package for classification and regression problems, randomForestSRC is available for analyzing survival data for supervised and unsupervised forests [21]. The need for a Random Forest procedure separate from one that handles classification and regression problems is well motivated as the survival data is characterized by features which is unique and not handled in CART technique. In particular, the notion of what constitutes a good node split for growing a tree requires extensive coding on the users part. The splitting rule used as weighted or un-weighted for identifying noise and signal is important not to have a bad branch of a tree [19][22].

2.3.1 Random Survival Forest Algorithm

The algorithm used for a random survival forest is similar to the algorithm used in random forest except it focuses for survival data. To fill this need a randomForestSRC, an R software package for implementing Random Forest for survival, regression and classification is introduced [21][22][23]. The algorithm used by ran-

domForestSRC for survival is broadly described as follows:

1. Draw B bootstrap samples from the original data, B = number of trees, ($ntree$).

2. Grow a tree for each bootstrapped data set. At each node of the tree randomly select predictors (covariates) for splitting on $mtry$. Split on a predictor using a survival splitting criterion. A node is split on that predictor which maximizes survival differences across daughter nodes.

3. Grow the tree to full size under the constraint that a terminal node should have no less than node size unique events (deaths).

4. Calculate an ensemble cumulative hazard estimate by combining information from the $ntree$ trees. One estimate for each individual in the data is calculated.

5. Compute an out-of-bag (OOB) error rate for the ensemble derived using the first b trees, where $b = 1, ..., ntree$.

2.3.2 Splitting Rules

Node splits are a crucial ingredient to the algorithm. The package randomSurvival-Forest provides four different survival-splitting rules for the user. These are: (i) a log-rank splitting rule, the default splitting rule, invoked by the option splitrule= "log-rank"; (ii) a conservation of events splitting rule, splitrule="conserve"; (iii) a log-rank score rule, splitrule="log-rankscore"; (iv) and a fast approximation to the log-rank splitting rule, splitrule="log-rankapprox". However, randomForestSRC package apply the random log-rank as a default splitting rule [23].

Notation

Assume we are at node h of a tree during its growth and that we seek to split h into two daughter nodes. We introduce some notation to help discuss how the

various splitting rules work to determine the best split. Assume that within h there are n individuals. Denote their survival times and 0-1 censoring information by $, (T_1, \delta_1), ..., (T_n, \delta_n)$. An individual l is said to be right censored at time T_l if $\delta_l = 0$, otherwise the individual is said to have died at T_l if $\delta_l = 1$. In the case of death, T_l will be referred to as an event time (death time). An individual l who is known to have been alive at T_l, but the exact time of death is unknown is called as right censored.

A proposed split at node h on a given predictor x is always of the form $x \leq c$ and $x > c$. Such a split forms two daughter nodes (a left and a right daughter) and two new sets of survival data. A good split maximizes survival differences across the two sets of data. Let $t_1 < t_2 < ... < t_N$ be distinct death times in the parent node h and let d_{ij} and Y_{ij} equal the number of deaths and number of individuals at risk at time t_i in the daughter nodes $j = 1, 2$. Note that Y_{ij} is the number of individuals in daughter node j who are alive at time t_i or who have an event (death) at time t_i. More precisely,

$$Y_{i1} = \#\{l : T_l \geq t_i, x_l \leq c\}, Y_{i2} = \#\{l : T_l \geq t_i, x_l > c\}$$

where x_l is the value of x for individual $l = 1, 2, ..., n$. Finally, define $Y_i = Y_{i1} + Y_{i2}$ and $d_i = d_{i1} + d_{i2}$. Let n_j be the total number of observations in daughter j, thus, $n = n_1 + n_2$. Note that $n_1 = \#\{l : x_l \leq c\}$ and $n_2 = \#\{l : x_l > c\}$.

Log-rank Splitting

The log-rank test for a split at the value c for predictor is:

$$L(x, c) = \frac{\sum_{i=1}^{N} \left(d_{i1} - Y_{i1} \frac{d_i}{Y_i} \right)}{\sqrt{\sum_{i=1}^{N} \frac{Y_{i1}}{Y_i} \left(1 - \frac{Y_{i1}}{Y_i} \right) \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i}}$$

The value $|L(x, c)|$ is the measure of node separation. The larger the value for

$|L(x,c)|$, the greater the difference between the two groups, and the better the split is. In particular, the best split at node h is determined by finding the predictor x^* and split value c^* such that $|L(x^*,c^*)| \geq |L(x,c)|$ for all x and c.

Random Log-rank Splitting

A random log-rank test can be used in place of $L(x,c)$ to greatly reduce computations. To derive the approximation, first rewrite the numerator of $L(x,c)$ in a form that uses the Nelson-Aalen estimator for the parent node, where Nelson-Aalen estimator is:

$$\hat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{Y_i}$$

as shown in LeBlanc and Crowley (1993) one can write:

$$\sum_{i=1}^{N} \left(d_{1i} - Y_{i1} \frac{d_i}{Y_i} \right) = D_1 - \sum_{l=1}^{n} I\{x_l \leq c\} \hat{H}(T_l),$$

where $D_j = \sum_{i=1}^{N} d_{ij}$ for $j = 1, 2$. Because the Nelson-Aalen estimator is computed on the parent node, and not daughter nodes, this yields an efficient way to compute the numerator of $L(x,c)$.

Now to simplify the denominator, we approximate the variance of the numerator of $L(x,c)$ as in of Cox and Oakes (1988) (this approximation was suggested to us by Michael LeBlanc in personal communication) as cited in [23]. Setting $D = \sum_{i=1}^{N} d_i$, we obtain the following approximation to the log-rank test $L(x,c)$:

$$\frac{D^{1/2} \left(D_1 - \sum_{l=1}^{n} I\{x_l \leq c\} \hat{H}(T_l) \right)}{\sqrt{\left\{ \sum_{l=1}^{n} I\{x_l \leq c\} \hat{H}(T_l) \right\} \left\{ D - \sum_{l=1}^{n} I\{x_l \leq c\} \hat{H}(T_l) \right\}}}.$$

2.3.3 Ensemble Estimation

The randomForestSRC package produces an ensemble estimate for the cumulative hazard function. Cumulative hazard function is the predictor and main input for

the computation of performance error in random survival forest. The ensemble estimation is derived as follows. First, for each tree grown from a bootstrap data set we estimate the cumulative hazard function for the tree. This is accomplished by grouping hazard estimates by terminal nodes. Consider a specific node h. Let $\{t_{ih}\}$ be the distinct death times in h and let d_{ih} and Y_{ih} equal the number of deaths and individuals at risk at time t_{ih}. The cumulative hazard estimate for node h is defined as,

$$\hat{H}_h(t) = \sum_{t_{ih} \leq t} \frac{d_{ih}}{Y_{ih}}.$$

Each tree provides a sequence of such estimates, $\hat{H}_h(t)$. If there are M terminal nodes in the tree, then there are M such estimates. To compute $\hat{H}(t|x_l)$ for an individual l with predictor x_l, simply drop x_l down the tree, and then the terminal node for l yields the desired estimator [16][23]. More precisely,

$$\hat{H}(t|x_l) = \hat{H}_h(t), if \quad x_l \epsilon h. \tag{2.3.1}$$

Note this value is computed for all individuals l in the data.

The estimate (2.3.1) is based on one tree. To produce our ensemble we average (2.3.1) over all $ntree$ trees. Let $\hat{H}_h(t|x_l)$ denote the cumulative hazard estimate (2.3.1) for tree $b = 1, ..., ntree$. Define $I_{l,b} = 1$ if l is an OOB point for b, otherwise set $I_{l,b} = 0$. The OOB ensemble cumulative hazard estimator for l is [23]:

$$\hat{H}_e^*(t|x_l) = \frac{\sum_{b=1}^{ntree} I_{l,b} \hat{H}_b(t|x_l)}{\sum_{b=1}^{ntree} I_{l,b}}.$$

Observe that the estimator is obtained by averaging over only those bootstrap samples in which l is excluded (i.e., those datasets in which l is an OOB value). The OOB estimator in contrast to the OOB bootstrap ensemble cumulative hazard estimator

that uses all samples is:

$$\hat{H}_e\left(t|x_l\right) = \frac{1}{ntree}\sum_{b=1}^{ntree}\hat{H}\left(t|x_l\right) = \frac{1}{B}\sum_{b=1}^{B}\hat{H}\left(t|x_l\right). \tag{2.3.2}$$

2.3.4 Prediction Error

To compute the error rate, we need to have an OOB ensemble estimator $\hat{H}_e^*\left(t|x\right)$ and then using this estimator to the Harrell's concordance index, we can measure the performance of survival prediction by taking into account censoring of subjects [16][23]. Before computing concordance index, we must define what constitutes a worse predicted outcome. Let $t_1^*, ..., t_N^*$ denote all unique event times in the data. Individual i is said to have a worse predicted survival experience than j if;

$$\sum_{k=1}^{N}\hat{H}_e^*\left(t_k^*|x_i\right) > \sum_{k=1}^{N}\hat{H}_e^*\left(t_k^*|x_j\right).$$

Then the procedure for computing concordance error rate is as follows:

1. Form all possible pairs of observations over all the data.
2. Omit those pairs where the shorter event time is censored. Also, omit pairs i and j if $T_i = T_j$ unless $\delta_i = 1$ and $\delta_j = 0$ or $\delta_i = 0$ and $\delta_j = 1$. The last restriction only allows ties if one of the observations is a death and the other a censored observation. Let Permissible denote the total number of permissible pairs.
3. Count 1 for each permissible pair in which the shorter event time had the worse predicted outcome. Count 0.5 if the predicted outcomes are tied. Let Concordance denote the total sum over all permissible pairs.
4. Define the concordance index: $C = \frac{Concordance}{Permissible}$.
5. The error rate is $Error = 1 - C$. Note that $0 \leq Error \leq 1$ and that $Error =$

0.5 corresponds to a procedure doing no better than random guessing, where as $Error = 0$ indicates perfect accuracy.

2.4 Persistent Homology

Topological data analysis (TDA) is an emerging field whose goal is to provide mathematical and algorithmic tools to understand the topological and geometric structure of data. Topological structures underlying data often appear to be of higher dimension and much more complex than smooth manifolds [26]. TDA applies different techniques, like persistent homology in particular to analyze the structure of high dimensional datasets. Persistent homology is a method used in TDA to identify the fundamental property and structure of geometrical objects. More often large datasets come as point clouds embedded in high dimensional Euclidean spaces, or in a general metric spaces and contain information in exploring relevant structures and properties associated with it [26][29].

In persistent homology, to identify the geometric structures underlying data uses different measures to detect their similarity or dissimilarity between objects using correlation structures or distance measures. Specifically in this study since our main motivation is to identify relations between points in the data that possess similar structure using persistent based clustering, we used distance measure to identify patients that have the same characteristics and form clusters based on persistent based filtration techniques used in Betti numbers [25].

2.4.1 Dissimilarity Measure for Cluster Analysis

In data clustering using persistent homology we used an input data the proximity index or dissimilarity measure and hence before applying persistent based clustering we have to have a distance measure for the dataset. Distance measure computation

procedure for cluster analysis described by Kaufman and Rousseeuw book, 2005 [31], can be summarized as follows:

First the dataset used for clustering can be arranged in the following structures of n number of items with the corresponding attributes, that is the dimension of the data should be $n \times p$ objects-by-attributes matrix, where rows represent items or objects and columns represent the variables associated with each object. If the data is arranged in this way we can compute the $n \times n$ dissimilarity matrix, using the "DAISY" auxiliary program in the cluster R package. This program computes the dissimilarity matrix for items using attributes measured from them. The dissimilarity matrix computed from the dataset, $d(X_i, X_j) = d(X_j, X_i)$ measures the difference or dissimilarity between the objects X_i and X_j, $X_i, X_j \in \Re^p$. The advantage of using this program is that it can handle all attribute types like nominal, ordinal, asymmetric binary and ratio-scaled variables in the computation of the dissimilarity measure between objects. In calculating the distance measure the daisy program have three options, the Euclidean (default), Manhattan both for dataset with all metric attributes and the third one which is an extension of the above two developed by Gower (1971) which incorporates and handles attributes of mixed type in a dataset for the calculation of proximity measure. The resulting dissimilarity measure $d(X_i, X_j) = d(i, j)$ developed by Gower can be defined as [31]:

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}},$$

where $d_{ij}^{(f)}$ is the contribution of variable f to $d(i, j)$, which depends on its type:

- f binary or nominal: $d_{ij}^{(f)} = 0$ if $X_{if} = X_{jf}$, and $d_{ij}^{(f)} = 1$ otherwise,
- f interval-scaled: $d_{ij}^{(f)} = \frac{|X_{if} - X_{jf}|}{max_h X_{hf} - min_h X_{hf}}$
- f ordinal or ratio-scaled: compute ranks r_{if} and $z_{if} = \frac{1 - r_{if}}{max_h r_{hf} - 1}$ and treat

these z_{if} as interval-scaled and,

$\delta_{ij}^{(f)}$= weight of variable f:

- $\delta_{ij}^{(f)} = 0$ if X_{if} or X_{jf} is missing,
- $\delta_{ij}^{(f)} = 0$ if $X_{if} = X_{jf} = 0$ and variable f is asymmetric binary,
- $\delta_{ij}^{(f)} = 1$, otherwise.

The output from daisy is an object of the class dissimilarity, and can be used as input for several of the clustering functions.

2.4.2 Homology

Homology is a mathematical prescription that calculates the algebraic properties of objects called chain complexes. When these chain complexes consist of objects called simplexes, the homology that is calculated is a topological invariant of the space. It is thus a way to define isomorphisms of groups rather than homeomorphisms of spaces [28]. This turns out to simplify the question of whether two spaces are fundamentally put together the same way or not. Formally, simplicial homology is defined as follows.

A simplicial complex is a set K, together with a collection S of subsets of K called simplices such that for all $v \in K, \{v\} \in S$, and if $\tau \subseteq \sigma \in S$, then $\tau \in S$. We call the sets $\{v\}$ the *vertices* of K. When it is clear from context what S is, we refer to set K as a complex [27][30].

A simplicial $K - chain\,(c_k)$ is a sum of $K - simplices\,(\sigma_k)$:

$$c_k = \sum_i \alpha_i \sigma_k^i, \alpha \in F,$$

where F is some field. Each $K - simplex$ can be thought of as a $K - dimensional$ polytope. Thus, a 0-simplex is a single point, a 1-simplex represents a line segment,

a 2-simplex represents a triangle, a 3-simplex represents a tetrahedron, etc. Various $K - chains$ define a free Abelian group, which is denoted as C_k, that is $c_k \in C_K$. The boundary operator $\partial_k : C_k \to C_{k-1}$, is a linear homomorphism defined to act on $\partial_k = [v_0, v_1, ..., v_k]$:

$$\partial_k \sigma_k = \sum_i (-1)^i [v_0, v_1, ..., \hat{v_i}, ..., v_k] \in C_{k-1},$$

where "$\hat{v_i}$" means this element is removed from the simplex. The boundary homomorphisms connect the chain groups. A sequence of abelian chain groups connected with their boundary homomorphisms is known as a *chain complex*. This gives the condition used to compute homology group. This definition allows a flow of information in the various chain groups:

$$... \to C_{k+1} \to C_k \to C_{k-1} \to ...$$

Various subgroups of this map can be defined. In particular, the cycle group $Z_k = ker\partial_k$ and the boundary group $B_k = im\partial_{k+1}$. Because $\partial^2 \equiv 0$, this implies $B_k \subseteq Z_k \subseteq C_k$. This condition is necessary so the homology group can be defined as the quotient group,

$$H_k = Z_k / B_k = ker\partial_k / im\partial_{k+1}.$$

Each homology group, H_k, contains information about the existence of k-dimensional holes in the space. For instance, the torus has $H_0 = Z, H_1 = Z \bigoplus Z, H_2 = Z$ and all the remaining homology groups vanish [26][28]. Persistent homology requires the spaces to be triangulable, that can be thought of as a sum of k-simplexes. For an arbitrary data set, there is no fundamental procedure to triangulate this space. Various ways do however exist, each with their own distinct set of rules that can be used to construct simplexes from data. For each of these procedures, we choose the coefficients in equation to be in Z_2.

Let us consider $X \in \Re^p$ to denote the point cloud and $d_{ij} = \|X_i - X_j\|^2$ denoted the Euclidean distance in a metric space between points X_i and X_j. Construction of simplicial complexes in persistent homology from data can be done either by considering Vietoris-Rips complex or Cech complex procedure [25][26]. Formation of complexes can be done and visualized by considering and drawing a disk with radius $\epsilon/2$ centering each $X_i's$ in the data set and then data points which intersect with each other within the neighborhood will connect each other to form complexes with different dimensions starting from the simplest one edge, triangles, tetrahedron to higher order topological features or complexes. The construction of simplexes depends on the choice of ϵ and choosing different ϵ value yields different complexes for the same dataset. There is no cut-off value for choosing ϵ, instead we choose the best value that gives appropriate and valid simplexes that last long and explain some features of data [27][28].

2.4.3 Vietoris-Rips Complex

Given a point cloud, the Vietoris-Rips (VR) Complex of a point cloud Z at filtration value of $\epsilon, (R_\epsilon)$ defines k-simplexes as being determined by (k+1)-tuples of points whose balls of radius $\epsilon/2$ pairwise intersect [26]. The balls are drawn around each point in the point of cloud and the radius can be computed with an arbitrary metric. Hence, to construct $R\,(Z, \epsilon)$:

- The vertex set is Z.
- Edge $[a, b]$ is formed in $R(Z, \epsilon)$ iff $d\,(a, b) \leq \epsilon$
- Higher dimensional simplexes are in $R(Z, \epsilon)$ if all of its edges are in $R(Z, \epsilon)$.

The driving force for the construction is that the union of the balls that forms different topological features from point clouds or dataset and which we interpret

as being fundamentally representative of whatever topology the points came from, has a homotopy type that is closely related to the homotopy type of $R(Z, \epsilon)$ [29]. Each of the simplexes σ has vertices that are pairwise within distance ϵ. The VR complex is computed up to a maximum filtration value ϵ'. The complex can then be extracted at any $\epsilon < \epsilon'$. The evolution of the simplicial complexes over increasing values of ϵ can be tracked using persistence diagrams or barcodes.

Barcode is a graphical representation of $R(Z, \epsilon)$ as a collection of horizontal line segments in a plane whose x-axis corresponds to the parameter or the filtration value and whose y-axis represents an (arbitrary) ordering of homology generators as shown in Figure 2.2.

Betti Numbers are integers that count how many generators of a specific dimension exist at a specific filtration value. For instance, $|H_1\left(R(Z, \epsilon), \epsilon = 4\right)| = 2$ means the first dimensional homology group for a Vietoris-Rips complex at filtration of $\epsilon = 4$ has Betti number equal to 2. In other words, it has 2 one-dimensional loops at this specific filtration value. In a similar procedure the k^{th} Betti number denoted by β_k describe the topological properties of objects, hence [24][25]

- counts the number of connected components of a complex K,
- counts for instance the lower dimensions,
- counts the number of voids in K, which is an empty space enclosed by k and
- counts the number of k-dimensional holes in K.

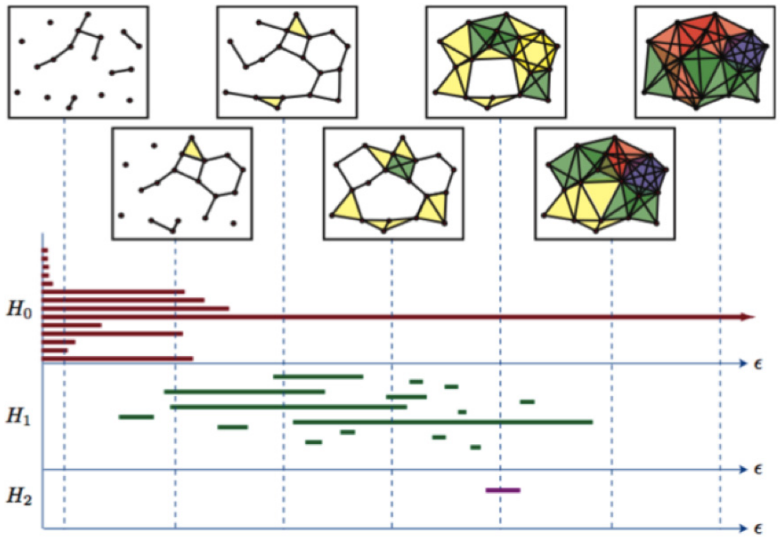


Figure 2.2: An example of a sequence of Rips complexes for a point cloud data set representing an annulus. Upon increasing ϵ (top) and the barcode representation of simplexes at different filtration value, ϵ with their representation in homology group zero, one and two (bottom). Source: BARCODES: The Persistent Topology of Data by ROBERT GHRIST [26]

Chapter 3

Data Analysis using Cox PH model and Random Survival Forest

3.1 Standard Cox PH Analysis: The Kidney Data

The data comprised of 366 patients who underwent a radical or partial nephrectomy for renal cell carcinoma at UCLA between 1989 and 2000. Of these 366 patients, we excluded 6 from the analysis due to missing information on their time to death. The response is time to death after performing nephrectomy. Among the patients 44.7% (161 out of 360 patients) died after performing kidney surgery and about 13.9% (50 out of 366) patients have clear cell renal cell carcinoma. The median survival time for patients was 5.12 years after surgery.

For the Cox Proportional Hazards model analysis we used eight different protein markers associated with prediction of renal cell carcinoma in patients. The Cox proportional hazards model is used to identify which protein markers are highly associated with the prediction of renal cell carcinoma. Hence, based on the results from Cox PH model as in Table 3.1, the most significant protein markers important in identifying kidney patients with renal cell carcinoma were Protein marker 3,

marker 5, marker 1 and marker 6. Therefore, based on the Cox PH model fit, the variables most important in identifying and predicting survival experience of patients who perform nephrectomy are respectively: Marker3, Marker5, Marker1, Marker6 and Marker7 (at 10% level of significance).

Variables	**Parameter Estimate**	**Standard Error**	**p-value**	**Hazard Ratio (HR)**	**95% CI for HR**	
Marker1	-0.00446	0.00265	0.0919	0.9955	0.9904	1.0007
Marker2	0.00145	0.00307	0.6255	1.0015	0.9955	1.0075
Marker3	0.04100	0.00758	<0.0001	1.0419	1.0265	1.0574
Marker4	-0.00319	0.00257	0.2145	0.9968	0.9918	1.0018
Marker5	0.01403	0.00448	0.0018	1.0141	1.0053	1.0231
Marker6	-0.00868	0.00361	0.0163	0.9914	0.9844	0.9984
Marker7	-0.00476	0.00280	0.0896	0.9953	0.9898	1.0007
Marker8	0.00181	0.00279	0.5178	1.0018	0.9963	1.0073

Table 3.1: Parameter estimates, 95 % confidence interval and corresponding p-values of the covariates in the study using Cox PH model for renal cell carcinoma cell data.

3.1.1 Checking the Proportionality of Covariates in the Model

One of the main assumptions of the Cox proportional hazard model is proportionality of hazards. The adequacy of the model was checked for the validity of proportional hazards assumption using a test based on the interaction of the covariates with the log of time. PH diagnostics plots for coefficients were also used to check for the trend against time. From Table 3.2, we can see that the proportionality test for all the covariates and the global test support the validity of the proportionality assumption at 5% level of significance. None of the covariates appear to be time dependent. Additionally, the global fit test shows that all the covariates were not significant, which justifies that PH assumption holds. Figure 3.1 depicts that the residuals are random without any systematic pattern and the smoothed plot looks straight without any departure from the horizontal line. This also implies that there is no violation of the proportional hazards assumption by the fitted model.

Variables	Rho-estimate	Chi-Square	p-value
Marker1	0.12834	2.6397	0.1042
Marker2	-0.00884	0.0114	0.9148
Marker3	-0.03154	0.13915	0.7091
Marker4	0.06375	0.82250	0.3645
Marker5	-0.00321	0.00186	0.9656
Marker6	-0.11376	2.27728	0.1313
Marker7	-0.12220	2.19915	0.1381
Marker8	-0.13619	2.88036	0.0897
GLOBAL	NA	9.71553	0.2856

Table 3.2: Proportional hazards assumption (PH) test for covariates included in the estimated Cox PH survival model for renal cell carcinoma data.

3.1.2 Checking Overall Significance of Cox PH Model

One method of checking goodness of fit of the model is to use R^2. In a proportional hazards regression model, as in all regression analyses, is to use there is no single, simple method of calculating and interpreting R^2, because in Cox proportional hazards model, R^2 depends on the proportion of the censored observations in the data. Therefore, for the model fitted in this study results of the Likelihood Ratio, Score and Wald tests for model goodness of fit are displayed in Table 3.3 and all of these tests, suggest that model is a good fit at a 5% level of significance. That is to say a model with all the covariates (protein markers) is adequate in explaining the survival experience of renal cell carcinoma patients.

Test	Chi-Square value	df	P-value
Likelihood Ratio	55.52	8	<0.0001
Wald	64.03	8	<0.0001
Score	67.28	8	<0.0001

Table 3.3: The Likelihood Ratio, Wald and Score tests for overall significance of covariates in the fitted Cox PH model for the renal cell carcinoma data.

3.2 Random Survival Forest Analysis: The Kidney Data

Previously we applied the standard Cox PH model for the analysis of the renal cell carcinoma data. Now we used the random survival forest algorithm developed by Ishwaran *et al* (2008) specifically designed for survival data that uses randomly selected bootstrap samples from the data to grow a tree. Random survival forest, unlike other analysis methods, provides an ensemble estimate for the cumulative hazard function. RSF is constructed based on trees grown from a sample of 366 renal cell carcinoma measures with 8 different protein markers for the prediction of survival of patients who underwent surgery.

In this study we grow a random sample of 1000 survival trees with the minimum node size set at 3, which is the minimum number of patients in a terminal node used to stop further splitting as shown below in Table 3.4. The number of variables tried in each split (as an input for splitting a node) is $mtry = 3$, the square root of 8, from a total of 8 covariates considered in the data suggested by Brieman (2001). For the splitting procedure we used the log-rank random splitting criteria developed in RSF and is known for its best splitting criteria with low prediction error and fast computational speed (Ishwaran *et al.*, 2008). The overall prediction error rate for the random survival forest is estimated to be 35.88% see Figure 3.3 (a). Based on the prediction error we can evaluate the performance of the model to predict the out of bag samples (around 37% of the data). Therefore, the model obtained by the RSF is fairly good to use for prediction of the out of bag samples. As the overall error rate is smaller than 50%, there is no strong evidence to conclude that the model is no longer important for prediction of survival probability of patients after surgery [16].

Sample size	366
Number of deaths	162
Number of trees	1000
Minimum terminal node size	3
Average number of terminal nodes	93.524
No. of covariates tried at each split	3
Total no of covariates used	8
Analysis	RSF
Family	surv
No of random splitting points	10
Error rate	35.88

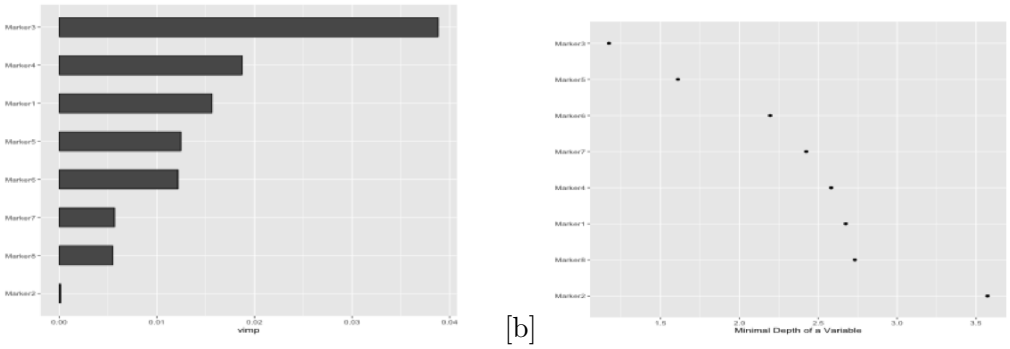
Table 3.4: The Random Survival Forest (RSF) algorithm result using the random Log-rank splitting criteria for the renal cell carcinoma data.

3.2.1 Variable Importance (VIMP) in Random Survival Forest

In a real dataset we cannot identify which variable is important in predicting the survival of a patient, and this is unknown before doing analysis. In RSF we can identify which variables are important in growing a tree (like in the standard Cox PH variables selection). The VIMP computation in a dataset can be determined by subtracting the prediction error of the ensemble obtained by random assignment of covariate X into the in-bag survival tree, to the prediction error of the original ensemble (Ishwaran *et al.*, 2008). Therefore based on the algorithm of those covariates considered in the study, those with positive VIMP values are predictive factors for survival rate. As we can see below from Table 3.5 and Figure 3.1 protein markers: Marker3, Marker4, Marker1, Marker6, Marker5, Marker7, Marker8 and Marker2 were the potential predictive factors, as they all had positive VIMP values. From those predictive factors Marker4 and Marker1 were found to be the most important factors worse than Marker3 in predicting the survival time of patients.

Variable	Depth Value	Depth Rank	VIMP value	VIMP Rank
Marker3	1.174	1	0.03912	1
Marker4	2.582	5	0.01940	2
Marker1	2.674	6	0.01494	3
Marker5	1.610	2	0.01378	4
Marker6	2.195	3	0.01202	5
Marker7	2.424	4	0.00574	6
Marker8	2.732	7	0.00406	7
Marker2	3.573	8	0.00029	8

Table 3.5: Variable Importance (VIMP) of the protein markers considered in the study using Random Survival Forest (RSF) for renal cell carcinoma data.



[a] [b]

Figure 3.1: (a) Variable Importance and (b) Minimal variable depth of the covariates using Random Survival Forest for the renal cell carcinoma data.

3.3 Discussion of Cox PH and Random Survival Forest Analysis

In this Chapter we have used the standard Cox PH model and Random Survival Forest to analyze and predict the survival time for renal cell carcinoma patients. The performance of both models was compared and the random survival forest is found to be better in predicting as compared to the standard Cox PH model (OOB concordance error of 35.98% vs 38.47%) in predicting the survival time of kidney cancer patients. Hence, based on the Harell's concordance index we can say that the RSF prediction is better than the standard Cox PH model.

The Cox PH model was built using the stepwise variable selection method. The covariates Marker3, Marker5, Marker6 and Marker7 were found to be significant in the contribution of model likelihood value. However using the variable importance techniques in the random survival forest, the covariates Marker3, Marker4, Marker1, Marker5 and Marker6 were found to be better in predicting patient survival compared to others with little contribution in patient prediction. In addition, none of the covariates' interaction terms were significantly associated in predicting survival time of patients in Cox PH model. However, in the random survival forest techniques, we identified different interactions terms associated in patient survival prediction as shown below in Figure 3.2.

From Figure 3.3 below, we can see that the estimated 5-year ensemble survival probability of patients using Marker5 conditioned on 3 groups with similar number of observations on Marker3. The pattern shows that at low value of Marker5, the predicted survival probability was found to be low, then after the survival probability increases with an increasing in Marker5. At some point, survival probability starts to decrease as the value of Marker5 increases for all of the three groups of Marker3. The survival plot also shows that the pattern is dependent on Marker3 level, that is the five year survival predicted probability for those patients whose marker level is below 6.88 is by far better than those patients with marker level higher than 17.5. Similarly, patients with Marker level less than 17.5 have a better survival probability than those patients with a higher level of Marker3.

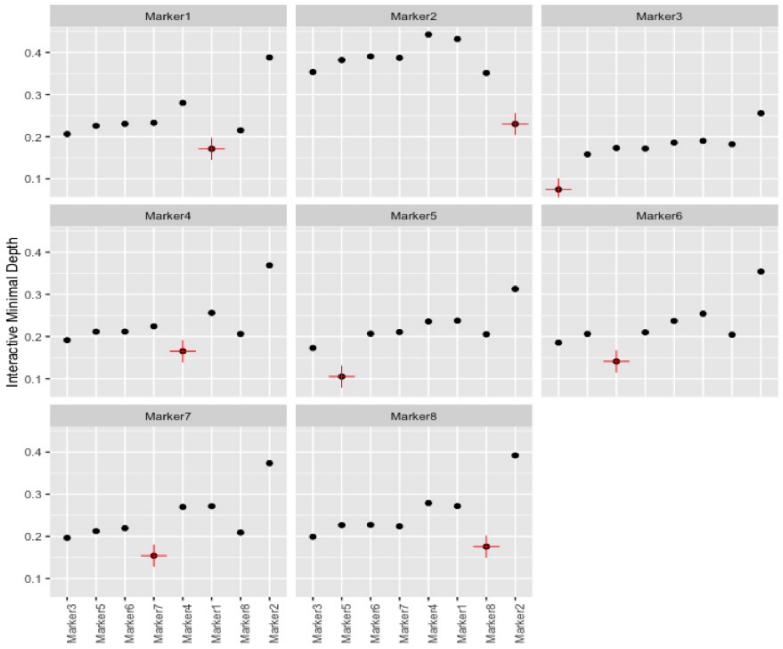
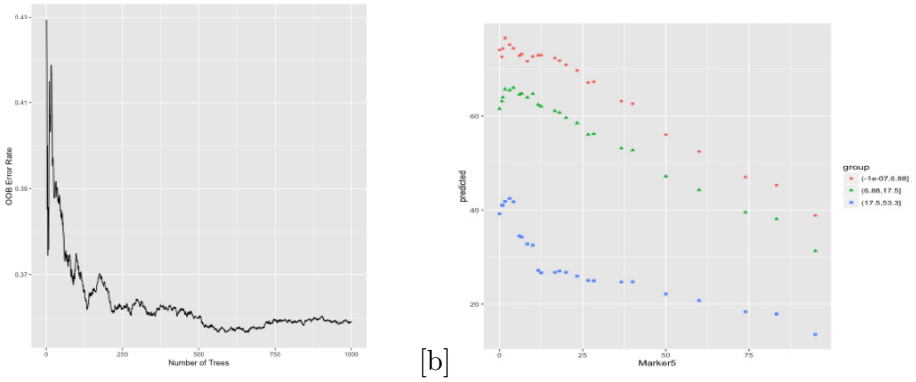


Figure 3.2: Minimal variable depth and importance for covariate interactions using Random Survival Forest for the renal cell carcinoma data.



[a] [b]

Figure 3.3: (a) The OOB error for RSF for 1000 trees (b) Predicted five-year survival probability versus Protein Marker5 conditioned on three groups of Marker3 using Random Survival Forest for the renal cell carcinoma data.

3.4 Standard Cox PH Analysis: The Liver Data

The liver transplant data contains 8361 alcoholic patients who received a liver transplant. The number of alcoholic patients with full a record and used to fit standard Cox PH model was 7143. The remaining patients are excluded from the study due to missing values in one or more of the variables considered for the study. From the 7143 patients included in the study, 67.16% (4797 of the 7143) of the patients deceased in different times after the transplant. The overall median survival time for all patients under study was 2191 days with a 95% confidence interval of 2179 and 2205 days after receiving their transplant.

(a) Standard Cox PH Analysis using Patient Characteristics Only

In our first attempt to fit the standard cox model, we used all the covariates associated with liver transplant obtained only from patient characteristics. In the second attempt, we consider all the covariates measured from both patient and donor characteristics. Hence, for fitting the cox model, we used nine variables taken from patient characteristics and found that most of the patient characteristics such as: gender, age, bilirubin level, creatinine level, albumin level, weight and cold_isch (cold ischematic time for the organ) of patients were found to be strongly associated with the survival time or time to death after transplant. Of these variables, age, weight, bilirubin level, creatinine level, albumin level, cold_isch and being blood group O are patient characteristics associated with risk of having poor survival time after transplant; as the measurements of these variables increase, the hazard of having early death after transplant increases (see Table 3.6). However, being male and having an increased level of cold_isch are identified as protective factors to experience early death after transplant. This result seems contradict with the conclusion made by Sibulesky *et al.* [34] and Michal *et al.* [35] and needs further investigation. A male patient and a patient who have increased cold_isch level have a better survival probability than others.

Variables	Parameter Estimate	Standard Error	p-value	Hazard Ratio (HR)	95% CI for HR	
Gender Male	-0.14686	0.04286	0.0006	0.8634	0.7939	0.9391
Female(R)						
Age	0.00414	0.00173	0.0164	1.0042	1.0008	1.0076
Blood Type AB	0.08552	0.06734	0.2041	1.0893	0.9546	1.2430
B	0.04253	0.04642	0.3596	1.0435	0.9527	1.1429
O	0.05629	0.03210	0.0795	1.0579	0.9934	1.1266
A(R)						
Bilirubin Level (LN)	0.20937	0.01499	<0.0001	1.2329	1.1971	1.2697
Creatinine Level (LN)	0.28996	0.02728	<0.0001	1.3364	1.2668	1.4098
Albumin Level (LN)	0.23720	0.06257	0.0001	1.2678	1.1215	1.4332
Height	-0.00061	0.00184	0.7396	0.9994	0.9958	1.0030
Weight	0.00377	0.00087	<0.0001	1.0038	1.0021	1.0055
Cold Isch (LN)	-0.41308	0.02694	<0.0001	0.6616	0.6276	0.6975

Table 3.6: Parameter estimates, 95 % confidence interval and corresponding p-values of the covariates in the study using Cox PH model using patient's characteristics only for liver transplant in alcoholic patients data.

Checking the Proportional Hazards Assumption of the Covariates in the Model

To check whether the PH assumption is satisfied by the covariates included in the model, we used a proportionality test as shown below in Table 3.7. As for the correlation with time, from the table we can see that the log transformed bilirubin and albumin levels of patients are found to varying with time, but the proportionality diagnostics for coefficients plots in the appendix show there is no much variation across time, that is the change in the covariate value with time remains proportional. Moreover, from Table 3.8 the overall significance of the standard Cox PH model is found to be significant, the variables included in the final model are adequate in predicting the survival experience of a patient. Hence, we used the standard cox model as our final model to predict the survival experience of patients after liver transplant.

Variables		Rho-estimate	Chi-Square	p-value
Gender	Male	0.01831	1.64875	0.1990
Age		0.00694	0.23161	0.6300
Blood Type	AB	-0.02757	3.67823	0.0551
	B	-0.01332	0.85349	0.3560
	O	0.00469	0.10614	0.7451
Bilirubin Level (LN)		-0.06212	17.8662	0.0003
Creatinine Level (LN)		-0.00823	0.32523	0.5681
Albumin Level (LN)		-0.04223	8.7337	0.0312
Height		-0.01046	0.4988	0.4801
Weight		-0.00138	0.0090	0.9250
Cold_isch (LN)		0.02910	3.2297	0.0723
GLOBAL		NA	39.6021	0.0001

Table 3.7: Proportional hazards assumption (PH) test for covariates included in the estimated Cox PH fit using only patient's characteristics for liver transplant in alcoholic patients data.

Test	Chi-Square value	df	P-value
Likelihood Ratio	645.8	11	<0.0001
Wald	689.1	11	<0.0001
Score	692.5	11	<0.0001

Table 3.8: The Likelihood Ratio, Wald and Score tests for overall significance of covariates in the fitted Cox PH model using patient's characteristics only for liver transplant data.

(b) Standard Cox PH Analysis using Patient and Donor Characteristics

In this section we attempted to fit the Cox PH model using both the characteristics of the patients and donors. There were 14 covariates considered to predict the survival experience of patients after liver transplant, 9 patient characteristics and 5 donors characteristics. For the analysis of these 16 covariates, we used 7120 liver patients from which 67.22% (about 4786 of 7120) died after the transplant. Based on the analysis we found most of the covariates included in the study were associated with the survival time of patients after liver transplant.

According to the result from the cox analysis, the covariates found to be risk factors

for survival time of patients after transplant are: bilirubin level, creatinine level, albumin level, recipient weight, donor age, donor weight and donor height (see, Table 3.9). As the measurements in these covariates increase, the hazard of dying after transplant will also increase, that is, if the measurements of these covariates increase for a patient after liver transplant then this will cause the patient to die earlier than the anticipated survival time. On the other hand, the covariates gender of a patient (male as compared to female) and cold_isch were different. A patient who received an organ from a male donor and patient who received an organ from a donor whose blood type is O were found to be protective factors from dying early after transplant, that is a male patient, a patient with increased level of cold_isch and a patient who received an organ from male donor and from a donor whose blood type is O will have a better survival time that those who do not have these characteristics.

Variables	Parameter Estimate	Standard Error	p-value	Hazard Ratio (HR)	95% CI for HR	
Gender Male	-0.1586	0.0431	0.0002	0.8534	0.7843	0.9285
Female(R)						
Age	0.0030	0.0017	0.0953	1.003	0.9995	1.0060
Blood Type AB	0.0581	0.1258	0.6442	1.0601	0.8282	1.3561
B	0.2212	0.1198	0.0649	1.2481	0.9864	1.5781
O	0.3980	0.0965	<0.0001	1.4890	1.232	1.417
A(R)						
Bilirubin Level (LN)	0.2131	0.0153	<0.0001	1.237	1.201	1.275
Creatinine Level (LN)	0.2948	0.0273	<0.0001	1.343	1.273	1.417
Albumin Level (LN)	0.2102	0.0626	0.0007	1.234	1.091	1.395
Height	-0.0012	0.0018	0.5256	0.998	0.9952	1.002
Weight	0.0027	0.0009	0.0024	1.003	1.001	1.004
Cold_isch (LN)	-0.4778	0.0323	<0.0001	0.6201	0.5821	0.6607
Don_Age	0.0051	0.0009	<0.0001	1.005	1.003	1.007
Don_Gender Male	-0.0760	0.0345	0.0275	0.9268	0.8662	0.9916
Female(R)						
Don_Height	0.0029	0.0011	0.0071	1.003	1.001	1.005
Don_Weight	0.0046	0.0008	<0.0001	1.005	1.003	1.006
Don_Blood Type AB	0.1527	0.1476	0.3010	1.165	0.8723	1.556
B	-0.1365	0.1233	0.2684	0.8724	0.6851	1.111
O	-0.3610	0.0958	0.0002	0.6970	0.5777	0.8409
A(R)						

Table 3.9: Parameter estimates, 95% confidence interval and corresponding p-values of the covariates in the study using Cox PH model, using patient's and donor's characteristics for liver transplant in alcoholic patients data.

Checking the Proportional Hazards Assumption of the Covariates in the Model

The above fitted Cox PH model was checked for the validity of the basic PH assumption. As we can see from the table below, Table 3.11, the covariates bilirubin level, albumin level of the patient and donors height look to violate the PH assumption and vary differently with time, however the pairwise correlation of each covariates with time does not look very high in absolute value and from the proportionality diagnostics plots we see there is no strong evidence of a time trend scatter plot. Hence we consider this model as a final model for predicting the survival time of patients after transplant. In addition using the likelihood ratio, Wald and Score tests, the model was checked for its overall significance with the covariates. See able 3.10.

Test	Chi-Square value	df	P-value
Likelihood Ratio	806.7	18	<0.0001
Wald	835.1	18	<0.0001
Score	839.6	18	<0.0001

Table 3.10: The Likelihood Ratio, Wald and Score tests for overall significance of covariates in the fitted Cox PH model using patient characteristics only for liver transplant in alcoholic patients data.

Variables		**Rho-estimate**	**Chi-Square**	**P-value**
Gender	Male	0.0164	1.32	0.250
Age		0.0074	0.264	0.607
Blood Type	AB	-0.0009	0.0039	0.950
	B	0.0088	0.370	0.543
	O	0.0024	0.0244	0.876
Bilirubin Level (LN)		-0.0604	0.169	<0.0001
Creatinine Level (LN)		-0.0101	0.481	0.488
Albumin Level (LN)		-0.0446	9.750	0.0018
Height		-0.0104	0.495	0.482
Weight		-0.0071	0.238	0.626
Cold_isch (LN)		0.0317	4.251	0.0392
Don_Age		0.0051	0.125	0.723
Don_Gender	Male	-0.0166	1.27	0.259
Don_Height		0.0575	8.70	0.0032
Don_Weight		-0.0027	0.0339	0.854
Don_Blood Type	AB	-0.0139	0.880	0.348
	B	-0.0171	1.41	0.235
	O	-0.0001	0.0001	0.994
GLOBAL		NA	54.40	<0.0001

Table 3.11: Proportional hazards (PH) assumption test for covariates included in the estimated Cox PH model using patient and donor characteristics for liver transplant in alcoholic patients data.

3.5 Random Survival Forest Analysis: The Liver Data

(a) Random Survival Forest Analysis using Patient Characteristics Only

Like in the previous data set, we did the analysis for the liver transplant data using the random survival forest. We first apply RSF for the entire data set using only patient characteristics (without incorporating the donor characteristics) and grow trees based on the nine patient covariates. To run the RSF algorithm we grow 1000 random trees of large size and 3 covariates taken randomly for splitting a parent node into different daughter nodes of similar survival experience until the minimum terminal node size 3 patients is reached, a point in which we would stop further splitting a node. See Table 3.12.

Sample size	8361
Number of deaths	5397
Number of trees	1000
Minimum terminal node size	3
Average number of terminal nodes	2739.33
No. of covariates tried at each split	3
Total no of covariates used	9
Analysis	RSF
Family	surv
No of random splitting points	10
Error rate	35.71

Table 3.12: Random Survival Forest (RSF) algorithm result using the random Log-rank splitting using patient's characteristics for liver transplant in alcoholic patients data.

The prediction error in using RSF for the overall grown trees is found to be 35.71% and from the OOB error rate we can see that as the number of random trees grown in a forest increases the OOB error slowly stabilizes and becomes closer to the mentioned overall error rate, (see figure 3.4 (b)). Hence, we can say that the forest grown based on the 8361 alcoholic patients has a good predictive ability of the survival experience (or the cumulative hazard of having death) after liver transplant

based on the behaviors and characteristics of a new patient receiving liver transplant.

Variable Importance (VIMP) in Random Survival Forest

The prominent advantage of using RSF over the standard cox model is that we can identify and rank the most important covariates used for growing trees in a forest so as to have reliable predictive ability of its survival for the OOB patients who receive a liver transplant or new patients waiting to undergo liver transplant. As a result, of those patient characteristics considered in the study cold_isch, creatinine level, bilirubin level, patients height and albumin level were found to be the five most important patient characteristics that give valid predictive hazard rate for alcoholic patients, (see Table 3.13). Consider the depth where the covariates used for splitting a node were almost associated with its VIMP. The first five patient characteristics were found on average work best in the first five splitting steps of the tree in the process of growing a huge size forest.

In the standard cox model none of two-way interactions between patient characteristics were found to be significant in predicting the survival time of patient after transplant. However, in employing RSF we can extract which variable interactions at what splitting step are very important in producing more or less homogenous daughter nodes with similar survival experience than the parent nodes. This variable interactive minimal depth plot is shown below and we can pick that the interaction between bilirubin and cold_isch, creatinine and cold_isch, creatinine and albumin, age and creatinine, height and cold_isch are some of the interactions terms which have the potential to give best splitting, at an early stage of splitting and growing trees in a forest, to result in a good predictive model for survival of alcoholic patients, Figure 3.5.

Variable	Depth Value	Depth Rank	VIMP value	VIMP Rank
Cold_Isch	1.294	2	0.01758	1
Creatinine Level	1.247	1	0.01633	2
Bilirubin Level	1.423	3	0.01395	3
Height	2.680	5	0.00379	4
Albumin Level	2.193	4	0.00360	5
Gender	4.357	9	0.00128	6
Weight	2.964	7	0.00122	7
Age	2.922	6	0.00027	8
Blood Type	3.413	8	0.00020	9

Table 3.13: Variable Importance (VIMP) of patient's characteristics considered in the study using Random Survival Forest (RSF) for liver transplant in alcoholic patients data.

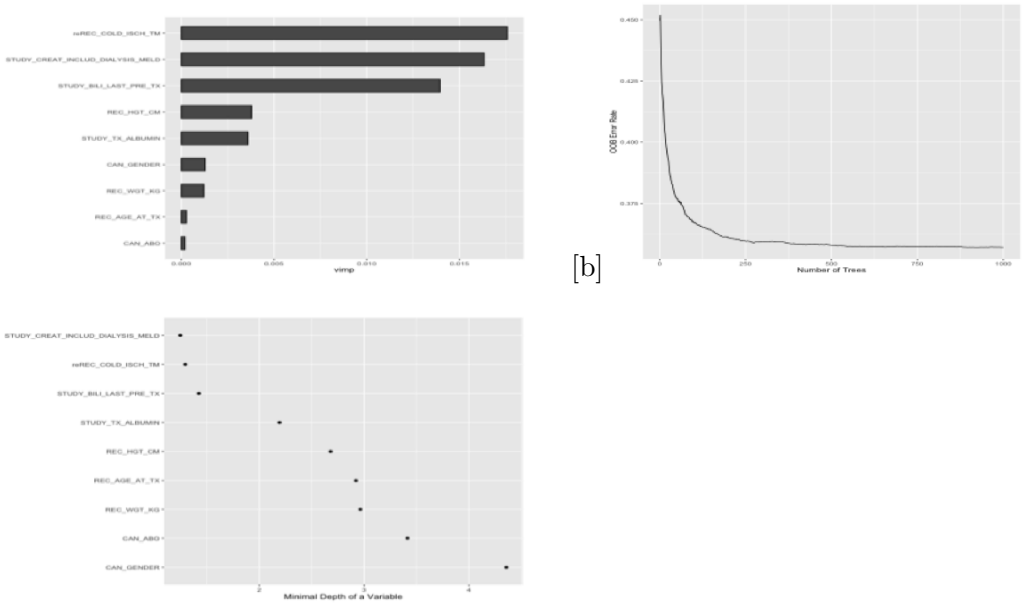


Figure 3.4: (a) Variable importance (b) The OOB error for RSF for 1000 trees and (c) Minimal variable depth of patient characteristics using Random Survival Forest in liver transplant for alcoholic patients.

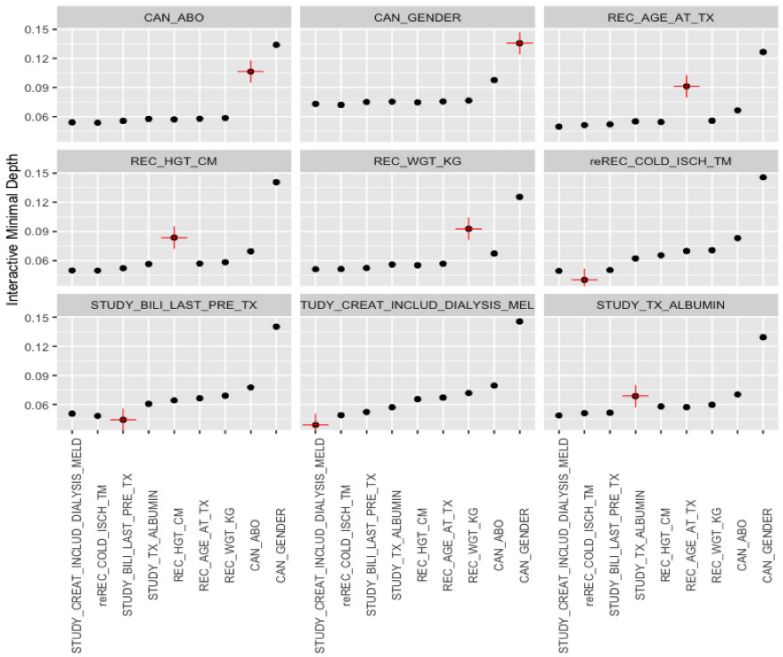


Figure 3.5: Minimal variable depth and importance for patient characteristics interactions using Random Survival Forest in liver transplant for alcoholic patients.

(b) Random Survival Forest Analysis using Patient and Donor Characteristics

RSF analysis is again applied in liver transplant for alcoholic patients by incorporating both patient and donor characteristics taken altogether to see if the donors characteristics have influence in predicting the survival experience of alcoholic patients. As usual the RSF algorithm was applied on the whole data set and resulted in an overall prediction error of about 34.89% in growing a forest of 1000 trees with minimal node size of 3 to stop further splitting. At each split a random sample of 4 out of 16 covariates was used to best split a parent node into daughter nodes to produce a better group of patients with similar survival time than the parent nodes. In growing trees to construct the forest about 1000 random trees were grown and the OOB error rate in growing all these trees is shown below in Figure 3.6 (b).

Sample size	8361
Number of deaths	5397
Number of trees	1000
Minimum terminal node size	3
Average number of terminal nodes	2760.458
No. of covariates tried at each split	4
Total no of covariates used	16
Analysis	RSF
Family	surv
No of random splitting points	10
Error rate	34.89

Table 3.14: Random Survival Forest (RSF) algorithm result using the random Log-rank splitting using both patient's and donor's characteristics for liver transplant in alcoholic patients data.

Variable Importance (VIMP) in Random Survival Forest

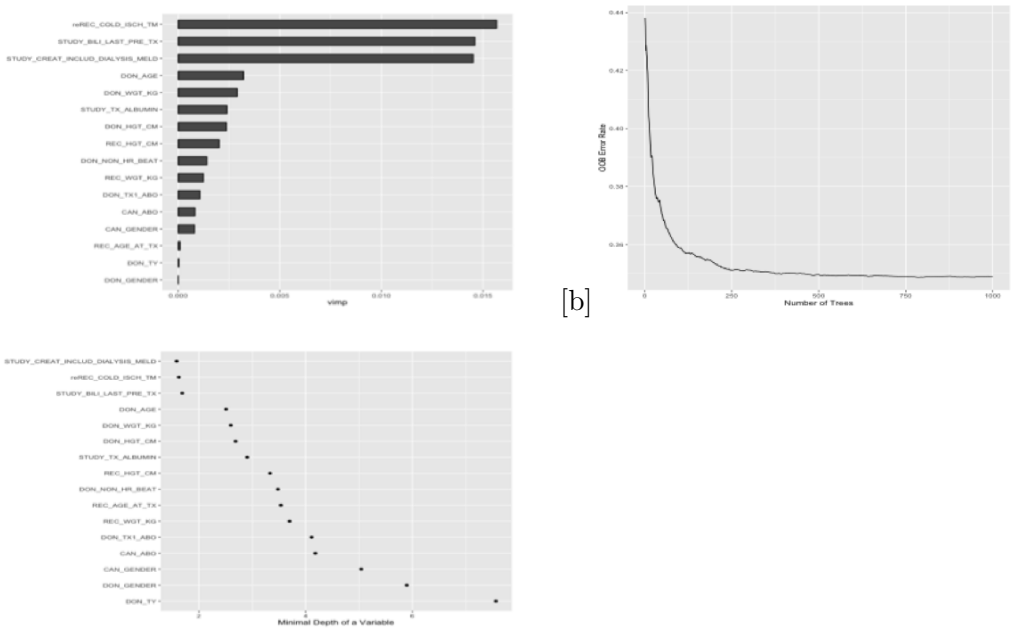
The variable importance assessment was also done on all 16 covariates and it was found that both patient and donor characteristics were very important in growing trees for the entire a forest which has the best splitting criteria. As we can see from Table 3.15 and Figure 3.6 (a) and (c), of those 16 patient and donor characteristics imputed for growing trees in a random forest, cold_isch, bilirubin level, creatinine level, donor age, donor weight, albumin level were found the top 6 important characteristics while patients' age, donor type and donor gender were found to be the least important characteristics used for splitting patients groups.

In addition, the top most important characteristics are used in splitting the nodes at, on average, an earlier stage than the less important ones. This shows that those important characteristics brought a good classification of alcoholic patients that underwent liver transplant and who experience similar survival trends close to the root nodes compared to classifying patients around the terminal nodes. The interactions of patients' characteristics were also investigated to achieve a better survival prediction for the OOB patients and for patients who are waiting to have liver transplant. It is shown that the interactions of patients' characteristics such as the interaction

of most of patient and donor characteristics (creatinine level, bilirubin, cold_isch, donor age and weight) were found to be very important at the root nodes in distinguishing patients with different survival time than at the terminal nodes. See Figure 3.7.

Variable	Depth Value	Depth Rank	VIMP value	VIMP Rank
Cold_isch	1.625	2	0.01567	1
Bilirubin Level (LN)	1.689	3	0.01460	2
Creatinine Level	1.582	1	0.01452	3
Don_Age	2.509	4	0.00321	4
Don_Weight	2.597	5	0.00291	5
Albumin Level (LN)	2.904	7	0.00261	6
Don_Height	2.689	6	0.00624	7
Height	3.333	8	0.00202	8
Don_Hrt_Beat	3.481	9	0.00141	9
Weight	3.699	11	0.00124	10
Don_Blood Type	4.112	12	0.00107	11
Blood Type	4.181	13	0.00083	12
Gender	5.041	14	0.00081	13
Age	3.537	10	0.00010	14
Don_Type	7.564	16	0.00003	15
Don_Gender	5.894	15	0.00001	16

Table 3.15: Variable Importance (VIMP) of patient and donor characteristics included in the study using Random Survival Forest (RSF) in liver transplant for alcoholic patients.



[a] [b] [c]

Figure 3.6: (a) Variable Importance (b) The OOB error for RSF using 1000 trees and (c) Minimal variable depth of patient and donor characteristics take together using Random Survival Forest in liver transplant for alcoholic patients.

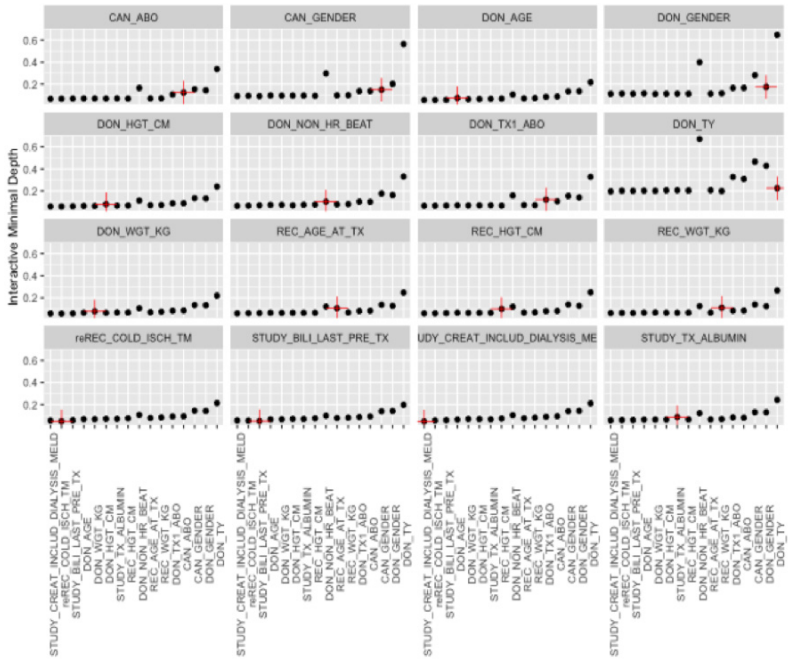


Figure 3.7: Minimal variable depth and importance for patient and donor characteristics interactions using Random Survival Forest in liver transplant for alcoholic patients.

Chapter 4

Data Clustering using Random Forest and Persistent Homology

4.1 Clustering using Random Forest: The Kidney Data

Prognosis and classifications of patients to deliver appropriate service needed within time were performed based on traditional methods using biological factors. Nowadays there are different methods applied not only to classifying patients but also to identifying groups (clusters) in high dimensional data by considering not only the biological factors but also others factors that help with identification of better and more similar groups. In this section, we applied the random forest for clustering technique so as to find a clinically meaningful group of patients with similar characteristics. To identify a group of patients with renal cell carcinoma we used different DNA expression profiles or protein expression patterns or protein markers that help us in identification group of renal cell carcinoma patients. Random forest for clustering is a method of clustering data into components using some distance measure, so as to partition a dataset into smaller classes with similar behavior. The distance measure employed in random forest for clustering to assess group of pa-

tients with similar behavior is a dissimilarity measure generated from the similarity matrix produced from the RF predictor using labeled data, based on observed and synthetic data.

Random Forest Dissimilarity for Clustering

Random forests clustering uses a dissimilarity measure for unlabeled data (Breiman and Cutler 2003). The dissimilarity measure is computed from the similarity matrix generated from the RF prediction process obtained from distinguishing the observed data to that of the synthetic data. The synthetic data is not the original data but it is generated from the reference distribution of the original data. Hence, RF classification is applied on the outcome obtained from these two data sets by labeling the observed original data as class one and the generated synthetic data as class two. The next step is restricting the resulting labeled similarity measure to the original observed data to obtain a similarity measure for the unlabeled observed data. However, the similarity measure generated is dependent on the process of generating synthetic data.

In the process of generating synthetic data, the generated data is added by randomly sampling from the product of empirical marginal distributions of the variables considered in the study from the observed data. The RF tree predictors are grown aiming to separate the synthesized data from the observed data; each grown tree will have potential variables associated with each other in splitting the synthetic and observed data. As a result the RF dissimilarity measure will be built based on these dependent variables and the dissimilarity measure for the observed data will be used as an input for the RF clustering process using partitioning around the medoids (PAM) (T. Shi and S. Horvath, 2006) [32].

In this study we used the RF dissimilarity measure as an input to cluster renal cell carcinoma patients using the eight protein markers. The dissimilarity measures generated using the RF algorithm described above is used to cluster the 366 renal cell

carcinoma patients into two clusters based on PAM (shown below in the following multidimensional scaling plot, Figure 4.1).

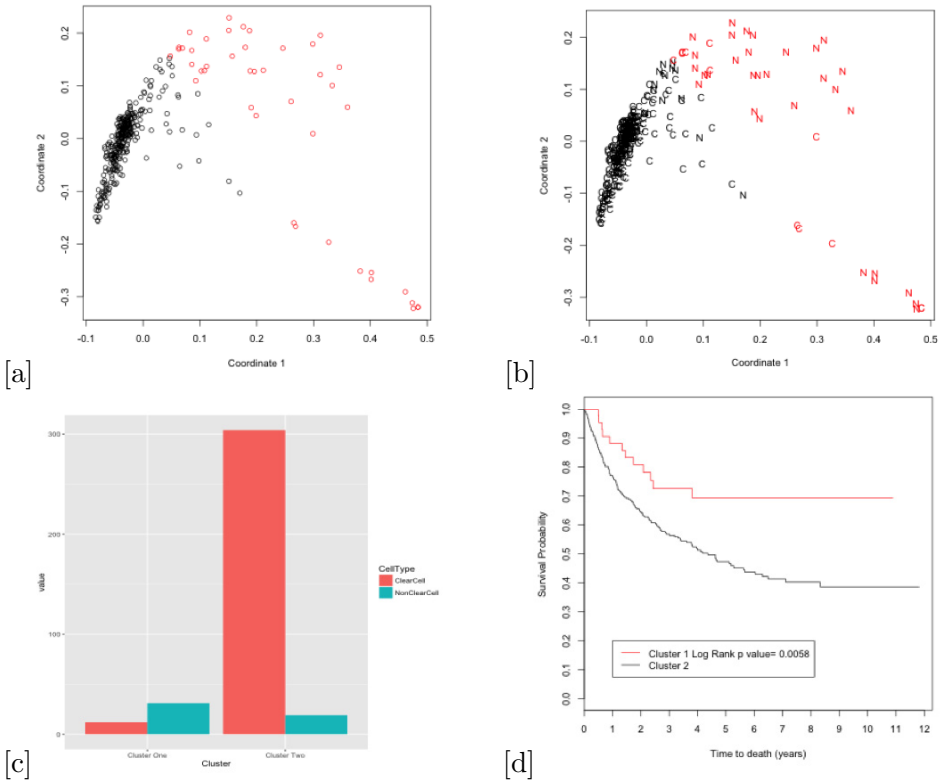


Figure 4.1: (a) RF two cluster multidimensional scaling scatter representation of the 366 renal cell carcinoma patients. (b) RF cluster representation of the 366 renal cell carcinoma patients with their tumor cell subtypes (C for clear tumor cell and N for non-clear tumor cell) and cluster membership: red for cluster 1 and black for cluster 2. (c) Histogram representation of composition of tumor cell types in cluster 1 and 2. (d) Predicted survival plot of renal cell carcinoma cell: red plot is for cluster 1 and black is for cluster 2.

From the resulting clusters, we can see that 11.7% (43 of 366) of renal cell carcinoma patients are in cluster one and of these patients 27.9% (12 of 43) patients have a clear tumor cell. The remaining 88.3% (323 of 366) are in cluster two and incorporates 94.1% of the patients with clear tumor cell. After construction of the cluster membership, we evaluate whether the predicted survival experience of those patients in cluster one is different from that of cluster two. A formal test for the difference in the survival experience between the two clusters was performed using

log-rank test and found to be significant (p-value = 0.0058). This implies that there is a significant difference in the predicted survival probability of patients with renal cell carcinoma in cluster 1 compared to cluster 2. Specifically, those patients in cluster one have a better survival probability than those patients in cluster two (see Figure 4.2).

In addition, we attempted to cluster those patients in to three groups using the RF dissimilarity measure generated from the algorithm using PAM, where about 8.5% (31 of 366) patients are classified as group one, of which 19.4% (6 of 31) patients are with clear tumor cell; about 62.8% (230 of 366) are in cluster two, out of which 89.6% (206 of 230) are with clear tumor cell; and the remaining 28.7% (105 of 366) of patients are in cluster three, with 99% (104 of 105) having clear tumor cells. To check whether these three clusters have different survival experiences a log-rank test was performed and found to be significant between groups (p-value-0.002). See Figure 4.2.

4.2 Clustering using Persistent Homology: The Kidney Data

High dimensional data often comes as a point cloud or in a matrix format embedded in a general metric spaces. Persistent homology is applicable to both point cloud and distance-based approaches generated from the metric spaces of the high dimensional data as an input to extract and visualize the topological and geometrical structure inferred from data. High dimensional datasets in a metric space are then analyzed using the applications of topological data analysis (TDA) including dimensionality reduction, visualization and simplification of data so as to explain the persistent features of the data.

Persistent homology results in different topological features of data at different Vietoris-Rips filtration levels, which uses radius of balls, to build the existing fea-

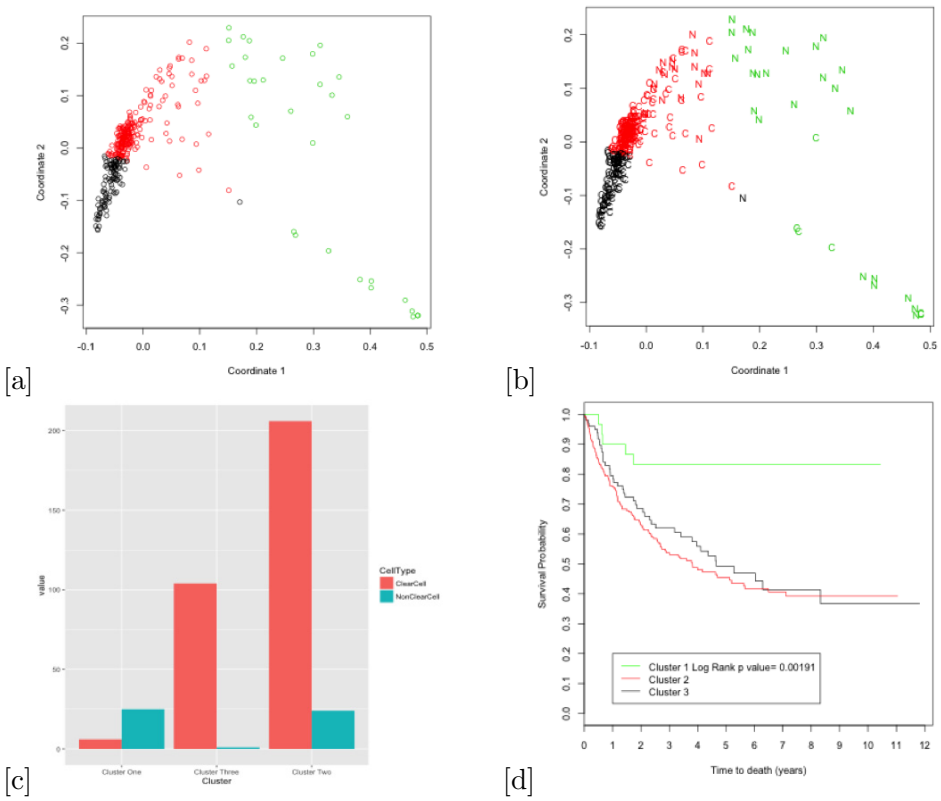


Figure 4.2: (a) RF three cluster MDS scatter representation of the 366 renal cell carcinoma patients. (b) RF cluster representation of the 366 renal cell carcinoma patients with their tumor cell subtypes (C for clear tumor cell and N for non-clear tumor cell) and cluster membership: green for cluster 1, red for cluster 2 and black for cluster 3. (c) Histogram representation of composition of tumor cell types in cluster one, two and three. (d) Predicted survival plot of renal cell carcinoma cell: green for cluster 1, red plot is for cluster 2 and black is for cluster 3.

tures. By changing the filtration levels (radius of balls), we can identify the connected components. The cycles and cavities appear, to form a more complex feature of the data. This process results in persistence diagrams which are used to reveal and characterize topological features for the purpose of classification, clustering and other explorations of data properties. Based on TDA techniques described in chapter two, we analyzed our dataset to explore and identify the persistent characteristics which are topologically invariant under homology of sublevel groups or Betti numbers obtained from Vietoris-Rips filtration of different dimensions. In section 4.2.1, the different connected components generated from Vietoris-Rips complexes with

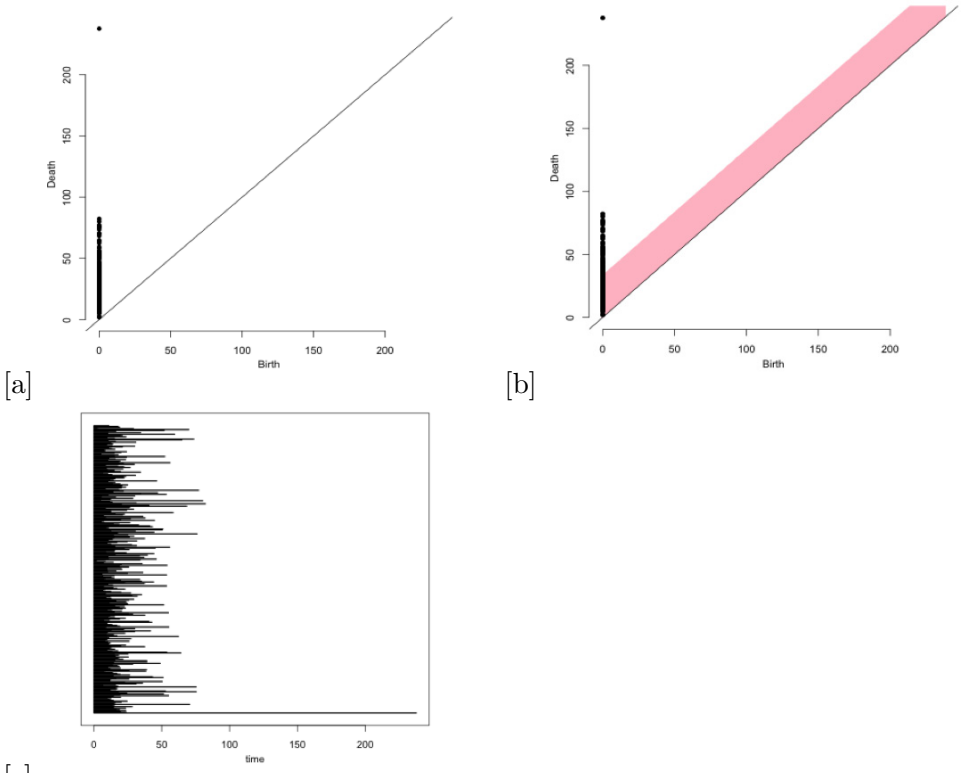
dimension zero, (in statistical language, the cluster) and in section 4.2.2, the persistent loops or cycles arising from the Vietoris-Rips complexes corresponding to the long lived one dimensional holes.

4.2.1 Cluster Extraction using Persistent Homology (Dimension zero)

Cluster identification using persistent homology uses the homology of Betti numbers obtained from Vietoris-Rips filtration levels of dimension zero. As we can see in Figure 4.3, we can visualize the persistent features, that is the clusters (topological features in dimension zero) by calculating the intervals of each component, in our case patients. This visualization can be done by either drawing all the renal cell carcinoma patients as a point using a persistence diagram (Figure 4.3(a)) with the corresponding $(1-\alpha)100\%$ confidence band (Figure 4.3(b)) or drawing all renal cell carcinoma patients as intervals in the plane, which are called as barcodes (Figure 4.3(c)). From the plots we can see that the points within the 95% confidence band are all noise points, that is they are born and die within short period of time. This means that these points have no contribution for the persistence of any feature which last a long time and form some topological feature.

In persistent homology a cluster is formed if points are connected and form a group at some threshold value of Vietoris-Rips filtration level (radius of balls, ϵ). The points that form a connected component with dimension zero or Betti zero using some threshold value of Vietoris-Rips filtration are called clusters. Different filtration value yields a different group of connected components in homology group dimension zero (Betti zero). In our data analysis we used two different threshold values which gave a significant connected component or cluster. The threshold value that we chose in our data setting is greater than 33.81217. A point that forms a connected component is said to be significant if the death time is outside of the 95% Confidence band [0, 33.81217]. If the death time is within the confidence band the point is considered to

be a noise point (Figure 4.3(b)).



[a] [b]

[c]

Figure 4.3: (a) Persistence diagram representations of the 366 renal cell carcinoma patients for features extraction at dimension zero. (b) A 95% confidence band for persistence diagram (dimension zero) of renal cell carcinoma patients. (c) Barcode representations of the 366 renal cell carcinoma patients for features extraction at dimension zero.

Hence, to identify the number of clusters in persistent homology that form the most persistent topological features that highlights those points which exist for a very long time and form connected components, we choose two different Vietoris-Rips filtration levels at death time $\epsilon = 51$ and death time $\epsilon = 55$, where we found two different types features, shown in Figure 4.4 and Figure 4.5. As we can see from Figure 4.4, using a filtration value $\epsilon = 55$, we found two clusters formed from the entire 366 renal cell carcinoma patients who underwent kidney surgery. The first cluster is formed from 35 patients for a connected component, which lasts longer

than the death time of 55, and the second cluster is formed by 331 of the patients that persist more than death time.

The identified clusters are based on persistent homology, as we can see from the table below or from the histogram in Figure 4.4(b). The first cluster is characterized by those patients whose tumor cell type is known to be non-clear, that is about 74.3% of those patients in cluster one are known to have non-clear tumor cell while the second cluster is characterized by majority of clear tumor cell type, which is 92.7%, a renal cell carcinoma subtype known to have relatively poor prognosis. This is confirmed from Figure 4.4(c), as cluster two has poor predicted survival experience when compared to cluster one and this survival difference is statistically significant (p-value < 0.001). Therefore, we can say that there is significant association between renal cell carcinoma subtype and the number of components formed by persistent homology (chi-square value of 114.98, p-value <0.001), that is the patients which form cluster one are those patients with non-clear tumor cell subtypes where as that of cluster two are those with clear tumor cell tumor subtypes.

Cluster	**Renal tumor cell subtypes**		**Total**
	Non-clear	Clear	
One	26(74.3)	9(25.7)	35
Two	24(7.3)	307(92.7)	331
Total	50	316	366

Table 4.1: Distribution of renal cell carcinoma subtype for each cluster formed by Persistent homology at Vietoris-Rips filtration value of $\epsilon = 55$

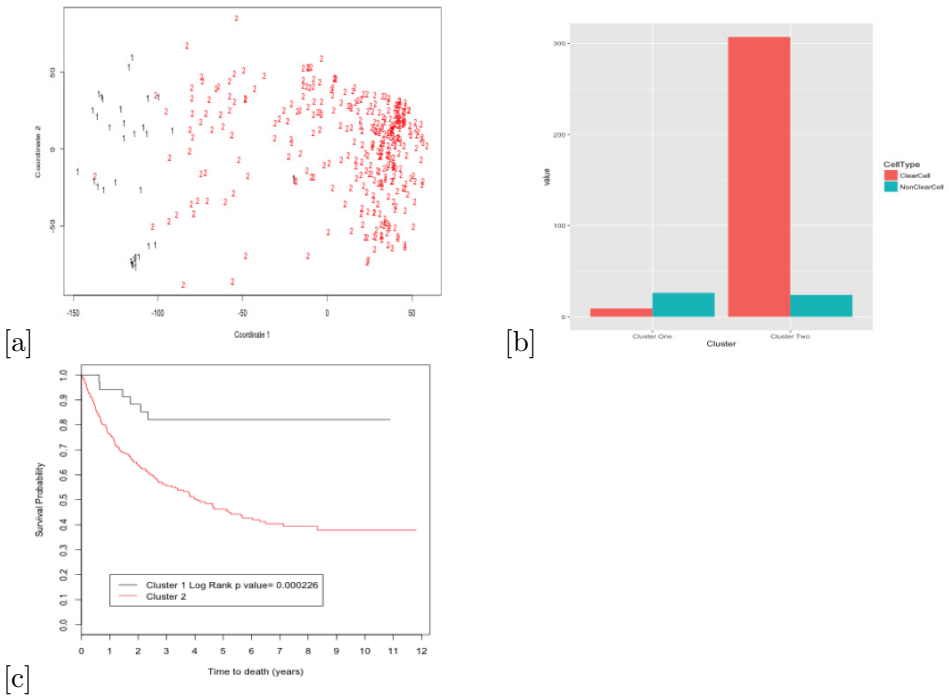


Figure 4.4: (a) Persistent homology cluster representation of the 366 renal cell carcinoma patients and cluster membership (at $\epsilon = 55$): black for cluster 1 and red for cluster 2. (b) Histogram representation of composition of tumor cell types in cluster one and two. (c) Predicted survival plot of renal cell carcinoma cell: black plot is for cluster 1 and red for cluster 2.

Similarly, we applied a different Vietoris-Rips filtration value of $\epsilon = 51$. As a result of shorter death time for each patient that forms persistent components, the number of points in persistent component one or cluster one increased. The number of patients which form the first persistent component is about 56, of which 67.9% are with non-clear renal cell carcinoma cell subtype, where as the second component has with 3.9% patients with non-clear cell renal carcinoma, seen in Table 4.2.

In addition, these persistent components formed by those 366 renal cell carcinoma patients who underwent nephrectomy were evaluated for their survival experience using a log-rank test and it was found that patients that form cluster one have a better survival experience than those patients that form the second cluster (p-value=0.0206), Figure 4.5(c). Hence, we can say that cluster one is characterized

as a persistent component formed by the majority of patients with non-clear renal tumor cell types and cluster two mainly by those patients with clear renal cell carcinoma cell subtype, seen in Figure 4.5(b).

Cluster	**Renal tumor cell subtype**		**Total**
	Non-clear	Clear	
One	38(67.9)	18(32.1)	56
Two	12(3.9)	298(94.3)	310
Total	50	316	366

Table 4.2: Distribution of renal cell carcinoma subtype for each cluster formed by Persistent homology at Vietoris-Rips filtration value of $\epsilon = 51$

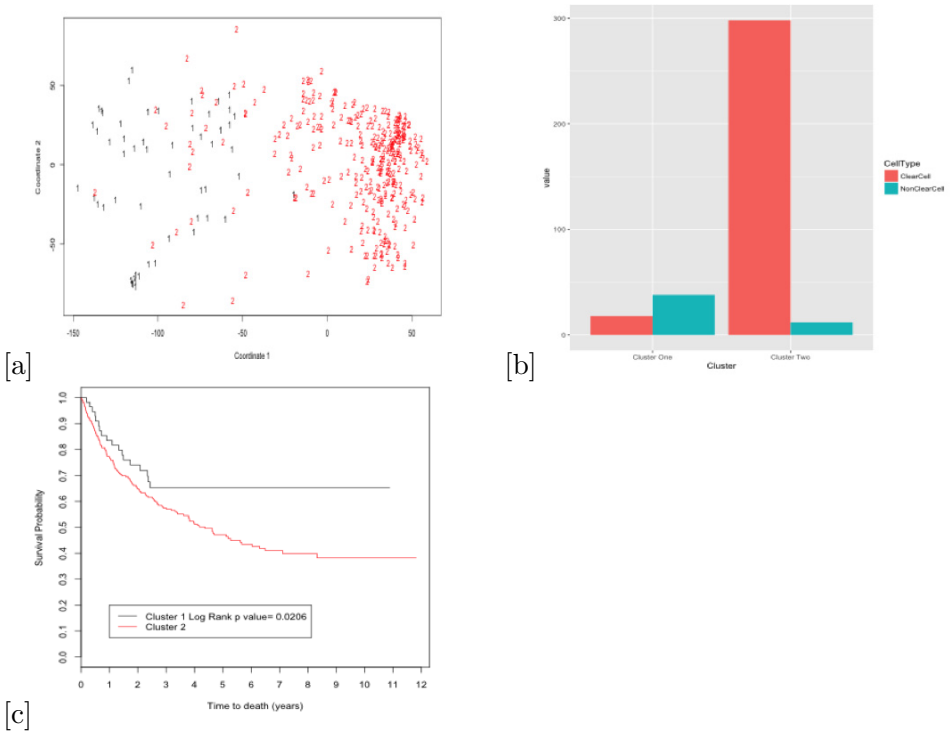
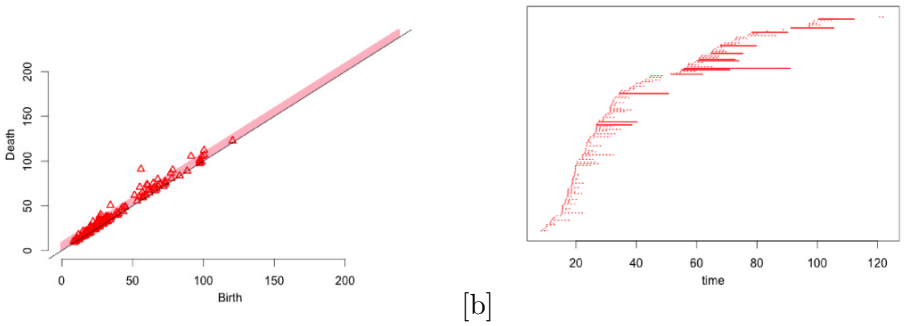


Figure 4.5: (a) Persistent homology cluster representation of the 366 renal cell carcinoma patients (at $\epsilon = 51$) and cluster membership: black for cluster 1 and red for cluster 2. (b) Histogram representation of composition of tumor cell types in cluster one and two. (c) Predicted survival plot of renal cell carcinoma patients: black plot is for cluster 1 and red for cluster 2.

4.2.2 Feature Extraction using Persistent Homology (Dimension one)

In this part we want to visualize and explore the most persistent components in homology groups of dimension one, or Betti one, obtained from Vietoris-Rips filtration. The persistent generators of the first homology group arising from Vietoris-Rips complexes correspond to long-lived loops formed by the set of renal cell carcinoma patients. To identify these most persistent loops, as shown in Figure 4.6, it is clear to visualize that there are some points forming loops that last a long time. As we can see from the persistence diagram, Figure 4.6 (a), not all points last long in forming the loops. Some of them are not significant and stay alive in forming the loops.



[a] [b]

Figure 4.6: (a) Persistence diagram representations of the 366 renal cell carcinoma patients for features extraction at dimension one (b) Barcode representation of the 366 renal cell carcinoma patients for features extraction at dimension one.

Even though, most of the bars looked visually insignificant, using Vietoris-Rips filtration of those persistent loops generated by renal cell carcinoma patients, we can identify the most significant points which forms dimension one loops (Betti one) features. The following table shows the patients who form the five most persistent loops. Of those 366 patients only 50 of them (13.7%) form the first five significant loops, shown in Table 4.3, with 8 patients identified to belong to more than one loop. As we can see from the patients forming the loops, after the third significant loop some of the patients become members of more than one loop, that is, the loops intersect with each other in one or more of its elements.

From Table 4.3 below, we can explore and identify the different characteristics possessed by the constructed loops so that we can explain the behavior of the renal cell carcinoma that form the loops. The first significant loop is more composed of those patients with non-clear cell renal carcinoma (64.7%), than patients with clear cell renal carcinoma, while the second loop constitutes about 6.7% of those patients with non-clear cell renal carcinoma, the third and fourth loops include an almost equal amount of clear and non-clear cell renal carcinoma, 60% and 55.6% respectively. Similarly, the fifth significant loop includes 25% of those non-clear cell carcinoma patients.

Persistent loops	**Renal tumor cell subtype**		**Total**
	Non-clear	**Clear**	
One	11(64.7)	6(35.3)	17
Two	1(6.7)	14(93.3)	15
Three	3(60)	2(40)	5
Four	5(55.6)	4(44.4)	9
Five	3(25)	9(75)	12
Total	23	22	58

Table 4.3: Distribution of renal cell carcinoma subtype for the first five persistent loops formed by persistent homology with dimension one (Betti one).

From Figure 4.7, we can examine that the first five most persistent loops: loop one, three and four are constructed more from those patients with non-clear renal cell carcinoma subtypes that is, 64.7%, 60% and 55.6% respectively. On the other hand persistent loop two and five are characterized with very few non-clear cell renal carcinoma subtypes, which are about 6.7% and 25% respectively.

In addition we applied a chi-square test of association to see if there is some association between the loops and renal cell carcinoma subtypes and found that significant association between loops and renal cell carcinoma sub types exists (X-squared = 14.17, df = 4, p-value = 0.0068). Hence we can describe the characteristics behavior of the most significant loops constructed from these patients (loop one, three, and

four) as as dominated by who have better survival experience. However persistent loops two and five are highly dominated by patients with clear cell renal carcinoma, which means these loops are dominated by patients with poor survival experience.

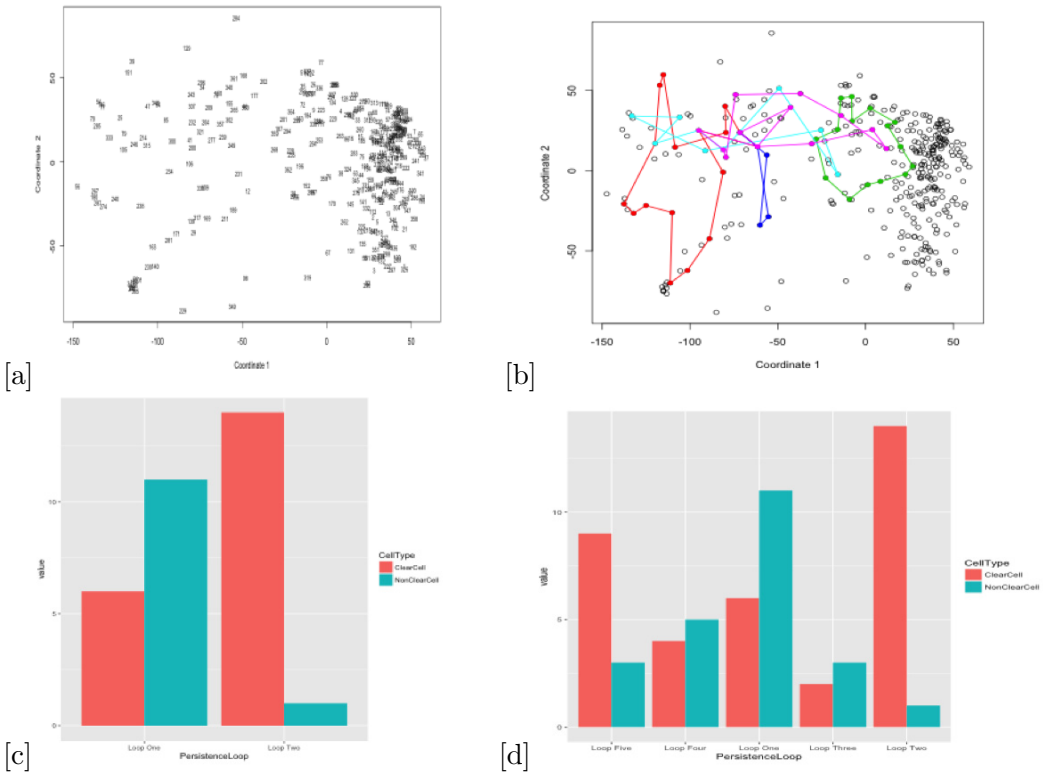


Figure 4.7: (a) 2D classical multidimensional scaling plot of 366 renal cancer cell patients. (b) Five most significant persistent features representation of the 366 renal cell carcinoma patients at dimension one: 1st, 2nd, 3rd, 4th, 5th, most significant features (clusters) are represented by red, green, blue, blue-green and purple loops respectively. (c) Histogram representation of composition of renal cell carcinoma subtypes in persistent loops one and two. (d) Histogram representation of composition of renal cell carcinoma subtypes in persistent loops one, two, three, four and five.

4.3 Clustering using Persistent Homology: The Liver Data

4.3.1 Cluster Extraction using Persistent Homology (Dimension zero)

(a) Cluster Extraction using Patient Characteristics Only

In this section, we applied persistent homology to find patient clusters based on their characteristics. Like what we did in the standard Cox and RSF section, we divide our analysis into two parts, first using only patient characteristics and then both patient and donor characteristics to see if donor characteristics have a significant impact in forming clusters in dimension zero and thus a significant effect in identifying the most persistent loops in dimension one.

As we can see from Figure 4.8, most of the topological features are born and die early except few of them that are born early and persist a little bit longer to form connected components or meaningful and persistent clusters. In persistent homology, cluster formation is entirely dependent on the Vietoris-Rips (VR) filtration value or the radius of the balls used to form the connected components in the neighborhood. Hence, we pick two different threshold values and form clusters, see Figure 4.9. The first extracted cluster using a VR filtration value of $\epsilon = 10$ consists of 22 alcoholic patients in cluster one, of which 59% died with a median survival time of 4376 days after liver transplant, cluster two consists of a total of 478 patients of which 40% died with a median survival time of 5712 days after the transplant. This shows that cluster one is mostly formed by those patients who have lower survival time than those patients in cluster two.

Assessment for its significant difference in the survival experience was performed using log-rank test and found that there is no significant difference (p-value=0.119) in the survival time between patients in cluster one vs two, see Figure 4.9(a). Similarly, by changing the VR filtration value to $\epsilon = 11$, we gain a second type of cluster which consists of 15 patients in cluster one out of which 66.7% (10 of them) of the patients died with a median survival time of about 3653 days after liver transplant and of those patients forming cluster two, about 40% of them died having a median survival time of 5712 days after transplant. This difference in survival time is confirmed by log-rank test (P-value = 0.0976) at 10% level of significance.

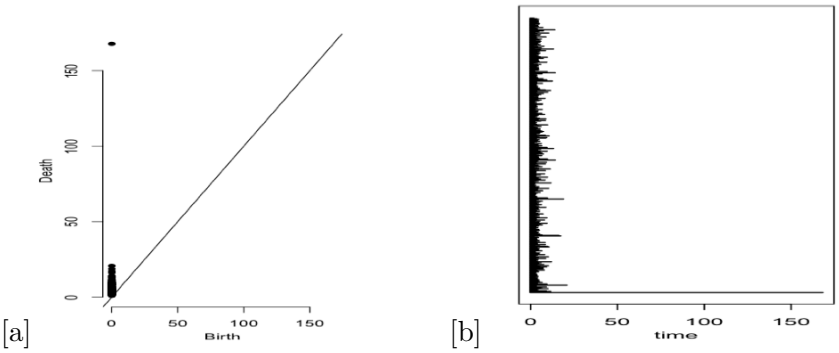


Figure 4.8: (a) Persistence diagram and (b) Barcode representations of the 500 alcoholic patients using patient characteristics only for features extraction at dimension zero.

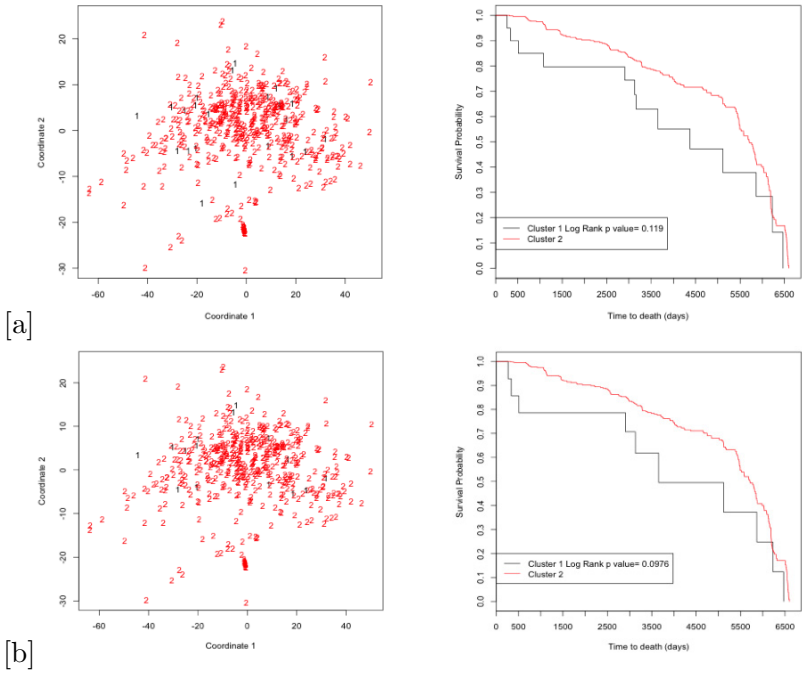


Figure 4.9: (a) Persistent homology cluster representation at $\epsilon = 10$ and (b) PH cluster representation at $\epsilon = 11$ for 500 sample patients and the corresponding survival curves of alcoholic patients receiving liver transplant: black plot is for cluster 1 and red for cluster 2.

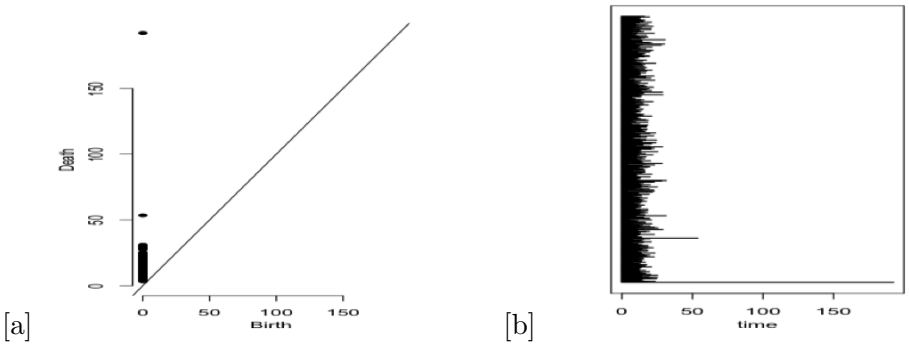
(b) Cluster Extraction using Patient and Donor Characteristics

Clustering using persistent homology was again applied on the sample data by incorporating donor characteristics to see the contribution of those characteristics in

forming cluster of patients with similar behavior. From Figure 4.10, we can see that of those of dimension zero that were born early, only a few of them persist to form connected components or clusters that last a long time to explore the features and characteristics of components in the cluster.

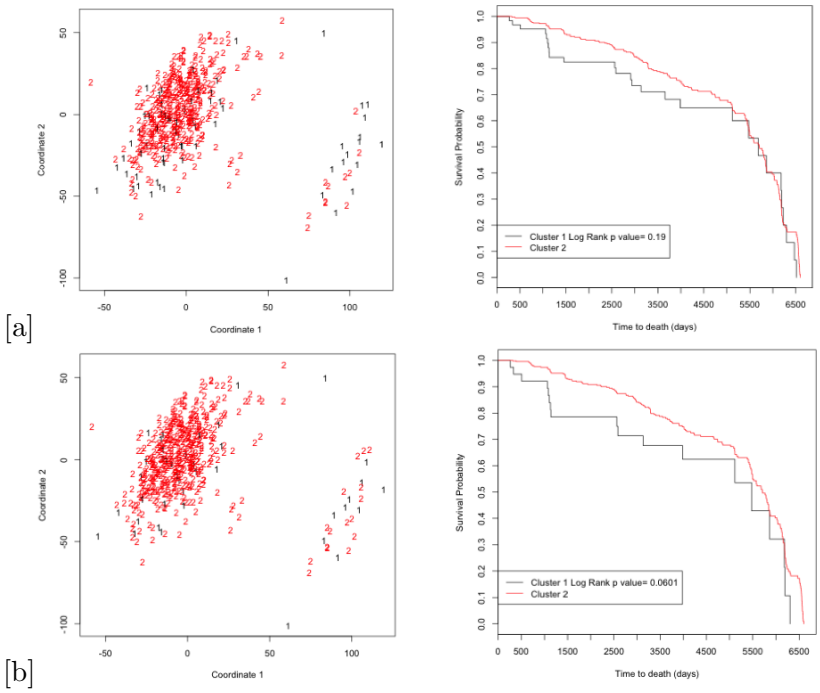
In investigating to explore the characteristics of patients which are significantly different, we used two different cut-points or VR filtration levels in finding clusters that lead us to a meaningful partition. In the first attempt we chose a filtration value of $\epsilon = 20$ in forming clusters, which are dissimilar in their survival experience, by taking into consideration both patient and donor characteristics and examining the effect of these characteristics in partitioning alcoholic patients who received transplant. The clusters formed using $\epsilon = 20$ consist of 66 patients in cluster one with 41% (27 of 66) reported as dead with a median survival time of 5682 days after transplant while cluster two consists of about 41% (177 of 434) of the patients reported as death after living a median survival time of 5712 days after receiving liver transplant.

From the result and Figure 4.11 (a), we can conclude that there are no significant differences in the survival experience, a log-rank p-value of 0.19, of patients who are in cluster one and cluster two. A similar attempt was made to form clusters by increasing the VR filtration level from, $\epsilon = 20$ to $\epsilon = 22$. The resulting clusters were assessed for features difference and found that of those patients forming cluster one, 45% (18 of 40) of the patients died after living a median survival time of 5474 days after liver transplant and that of patients in cluster two, 40.4% (186 of 460) patients died after living a median survival time of 5712 days after receiving transplant. This difference in survival time was evaluated using a log-rank test and found to have a significant difference in survival experience of patients in cluster one and cluster two at 10% level of significance (p-value = 0.0601), see Figure 4.11(b).



[a] [b]

Figure 4.10: (a) Persistence diagram and (b) Barcode representations of the 500 alcoholic patients for features extraction at dimension zero using both patient and donor characteristics.



[a]

[b]

Figure 4.11: (a) Persistent homology cluster representation of a sample of 500 alcoholic patients receiving liver transplant and cluster membership with corresponding survival curves $\epsilon = 20$ using patient characteristics only: black for cluster 1 and red for cluster 2. (b) PH cluster representation for 500 sample patients and the corresponding survival curves of alcoholic patients receiving liver transplant $\epsilon = 22$: black plot is for cluster 1 and red for cluster 2.

4.3.2 Feature Extraction using Persistent Homology (Dimension One)

(a) Feature Extraction using Patient Characteristics Only

Persistent homology is used to further investigate if there are some other characteristics that can be explored in higher order dimensional analysis. To do this we analyze the 500 sample data in dimension one (results in Figure 4.12). Figure 4.12 shows that in dimension one there are some characteristic features that persist longer than others and form some loops that are explained by patients who received liver transplant. Hence from the dimension one barcode representation of patients (Figure 4.12 (b)), we can identify that there are some significant features constructed, based on some alcoholic patients with similar characteristics to form long lasting persistent loops.

After identifying those patients that form the first five most significant persistent loops, an assessment for the difference in survival experience between patients forming these loops was conducted. As a result, except for the first loop all of them reveal that majority of their components come from alcoholic patients who receive transplant and have better survival experience (57.1%, 57.1%, 70% and 70% for persistent loop two, three, four and five respectively). While first persistent loop is characterized by incorporating a majority of patients who have poor survival experience after the transplant, about 54.5%, compared to the other persistent loops, seen in Table 4.4.

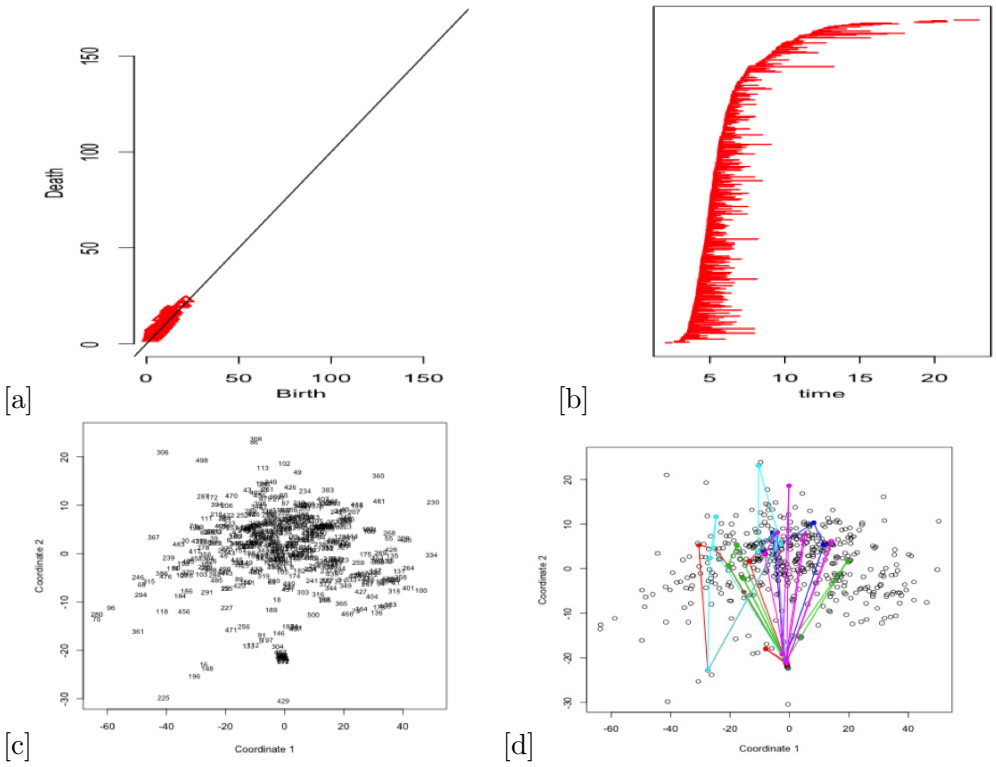


Figure 4.12: (a) Persistence diagram representations of the 500 sample alcoholic patients (b) Barcode representation of the 500 patients for features extraction at dimension one. (c) 2D multidimensional scaling plot of the 500 alcoholic patients (c) Five most significant persistent features representation of the 500 patients at dimension one: 1st, 2nd, 3rd, 4th, 5th, most significant features are represented by red, green, blue, blue-green and purple loops respectively using patient characteristics only.

Persistent loops	**Survival status**		**Total**
	Dead	**Alive**	
One	6(54.5)	5(45.5)	11
Two	3(42.9)	4(57.1)	7
Three	3(42.9)	4(57.1)	7
Four	3(30.0)	7(70.0)	10
Five	3(30.3)	7(70.0)	10
Total	18	27	45

Table 4.4: Distribution of alcoholic patients who receive liver transplant with the corresponding survival status for the first five persistent loops formed by persistent homology with dimension one (Betti one) using patient characteristics only.

(b) Feature Extraction using Patient and Donor Characteristics

Persistent homology was revisited for the analysis exploring persistent features formed by alcoholic patients possessing some characteristics and as we can see from the persistence barcode there are many components that form persistence loops with dimension one from patients, but not all of them form a significant feature that lasts a long time. As a result we considered the first five most significant persistent loops formed by alcoholic patients, Figure 4.13 (d).

The patients that form these five most persistent loops were identified and assessed for their survival experience and it was found that loops one, two and five were formed by patients who experience better survival time that those patients who formed the persistent loops three and four. Therefore, from this preliminary result, we can see that in addition to assessing the survival experience of patients, we also need to further investigate whether patients that form these persistent loops were affected not only by their characteristic features but also by their donors' features too. As we can see from Table 4.5 below, persistent loop one, two and five were formed by those patients with better survival time after transplant (67.7%, 75% and 71.4% respectively) compared to loops three and four, which comprises a little more than 50% of patients in each loop with poor survival experience. Therefore, taking into account donors' characteristics for patient who underwent liver transplant is more informative and helps clinicians in predicting survival experience and quality of life a recipient.

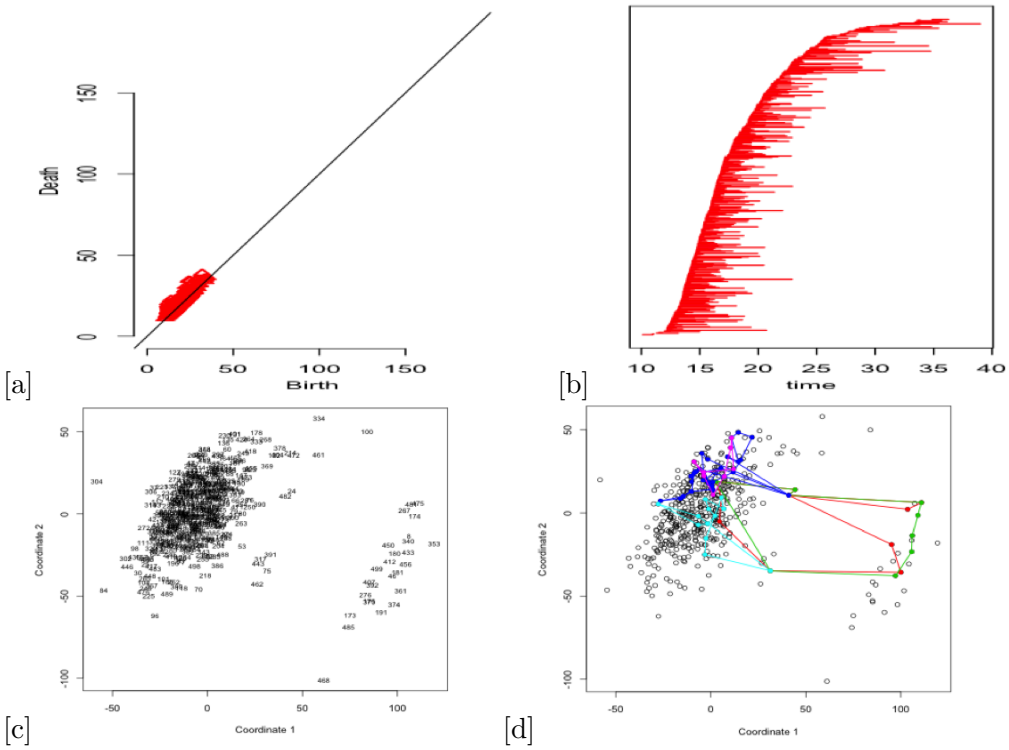


Figure 4.13: (a) Persistence diagram representations of the 500 alcoholic patients (b) Barcode representation of the 500 patients for features extraction at dimension one. (c) 2D multidimensional scaling plot of the 500 alcoholic patients (c) Five most significant persistent features representation of the 500 patients at dimension one: 1st, 2nd, 3rd, 4th, 5th, most significant features are represented by red, green, blue, blue-green and purple loops respectively using both patient and donor characteristics.

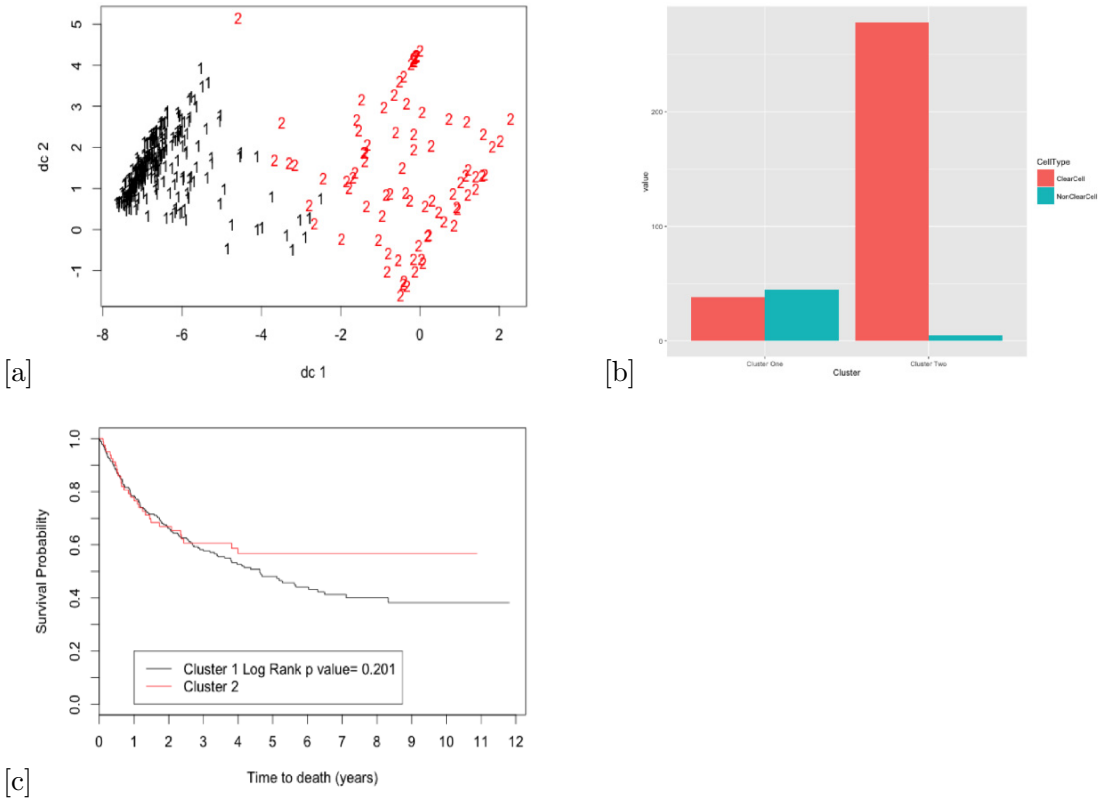
Persistent loops	**Survival status**		**Total**
	Dead	**Alive**	
One	3(33.3)	6(67.7)	9
Two	2(25.0)	6(75.0)	8
Three	13(52.0)	12(48.0)	25
Four	6(54.5)	5(45.5)	11
Five	2(28.6)	5(71.4)	7
Total	26	34	60

Table 4.5: Distribution of alcoholic patients who receive liver transplant with the corresponding survival status for the first five persistent loops formed by persistent homology with dimension one (Betti one) using both donor and patient characteristics.

4.4 Clustering Survival Data using K-Means

4.4.1 K-Means Clustering: The Kidney Data

The K-Means clustering, with two clusters (k = 2), was applied on the kidney data, as we can see from Table 4.6 below and from the histogram in Figure 4.14(b). The first cluster is composed of 83 patients of which 38 (45.8) are with clear tumor cell type and 45 (54.2) with non-clear tumor cell type. The second cluster consists of 283 patients with majority (98.2) of them are with clear cell carcinoma type. We can also see from Figure 4.14(c), that there is no statistically significant difference in survival experience of the two clusters (p-value = 0.201).



[a] [b] [c]

Figure 4.14: (a) K-Means (k=2) scatter plot cluster representation of the 366 renal cell carcinoma patients. (b) Histogram representation of composition of tumor cell types in cluster 1 and 2. (c) Predicted survival plot of renal cell carcinoma cell: black plot for cluster 1 and red for cluster 2.

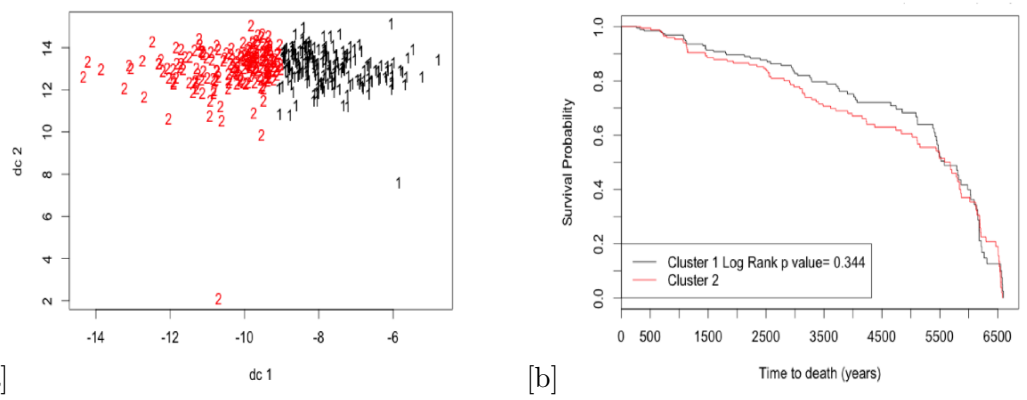
Cluster	Renal tumor cell subtype		Total
	Non-clear	Clear	
One	45(54.2)	38(45.8)	83
Two	5(1.8)	278(98.2)	283
Total	50	316	366

Table 4.6: Distribution of renal cell carcinoma subtype for each cluster formed by K-Means clustering with K=2.

4.4.2 K-Means Clustering: The Liver Data

(a) K-Means Clustering using Patient Characteristics Only

The liver data was analyzed using K-means clustering, with two clusters (k = 2), and found that, from 500 patients included in the study about 52.2% of them are in cluster one of which 39.5% of them are died with a median survival time of 5578 days after transplant. On the other hand, cluster two consists of 47.8% of the total patients with a median survival time of 5679 days, of which 48.4% of them are died after receiving the transplant. For the constructed clusters a Log-rank test was performed to test their survival experience and found that there is no statistically significant difference in the survival experience between the two clusters (p-value = 0.344), see Figure 4.15 (c).

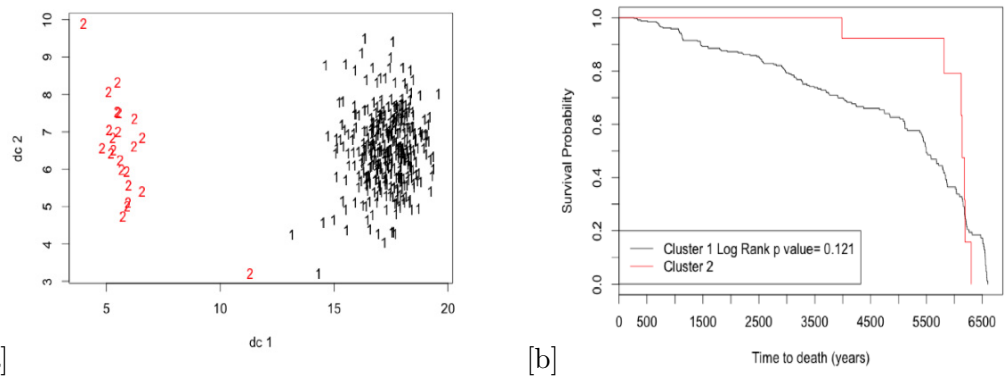


[a] [b]

Figure 4.15: (a) K-Means (k=2) scatter plot cluster representation of the 500 sample alcoholic patients and (b) Predicted survival plot of 500 sample alcoholic patients receiving liver transplant: black plot for cluster 1 and red for cluster 2.

(b) K-Means Clustering using Patient and Donor Characteristics

A similar analysis was applied on a sample of 500 liver data by considering both patient and donor characteristics. The result of K-means clustering shows that about 94.1% of them are cluster one and the remaining 5.9% in cluster two. In cluster one, about 30.4% (7 of 23) patients are died with a median survival time of 6133 days. In cluster two of those patients 44.5% (163 of 366) of them are died with a median survival time of about 5493 days. The Log-rank test for difference in survival experience in between these two clusters resulted in no statistically significant difference in the survival experience of the two clusters (p-value = 0.121), as shown in Figure 4.16.



[a] [b]

Figure 4.16: (a) K-Means (k=2) scatter plot cluster representation of the 500 sample alcoholic patients and (b) Predicted survival plot of 500 sample alcoholic patients receiving liver transplant: black plot for cluster 1 and red for cluster 2.

From the above discussions, we can see that clustering techniques, K-means, Random forest and Persistent homology were applied the kidney and liver datasets and showed that the proposed clustering methods: random forest and persistent homology techniques resulted in two different group of patients (clusters) with significantly different survival experience between groups. However, the K-Means clustering technique we are unable to identify group of patients or clusters with statistically significantly different survival time between clusters.

Chapter 5

Conclusions and Future Work

In this thesis we applied clustering techniques using random forest and persistent homology for survival data using renal cell carcinoma cancer and liver transplant data. Before conducting data clustering we analyzed these two datasets using standard Cox proportional hazards and Random Survival Forest to assess and identify the predictive factors strongly associated with survival time of patients under study. Both methods give similar conclusion in identifying the most important variables for the prediction of survival status. In RSF these significant factors associated with survival prediction were ranked based on their importance in identifying patient classification according to the trend in their survival time. RSF is found to have better power than standard Cox proportional hazards model in predicting the survival status of patients under study.

Moreover, these datasets were used in cluster analysis to identify groups with similar characteristics using random forest and persistent homology. Both methods identified a convincing cluster of patients with different behaviors between clusters. The constructed clusters identified by these two methods were evaluated using clinical and statistical aspects and are found to be reliable, that is, the survival experi-

ence for one cluster is better that the other. In persistent homology, in addition to identifying clusters at dimension zero, we applied persistent homology analysis in dimension one to explore further features extraction of the data. We considered the first five most persistent loops formed and these loops were evaluated using patients characteristics and found that patients forming these loops have something in common, that is they have similar survival experience. Some loops were formed by a majority of those patients whose survival experience is better while some others are formed by those majority patients with poor survival status. Hence, clustering high dimensional data using RF and persistent homology is more meaningful and flexible in extracting and capturing the underlying features that need further clinical attention to provide better medical services and improve quality of life for patients.

The limitation of this work is that data analysis to identify the variables associated with survival prediction was performed by standard Cox proportional hazards model and did not try the extended Cox model which does not need the proportional hazards assumption (variables time-dependent). Some of the variables considered in the study did appear to violate the proportional hazards assumption. In the topological data analysis, this study did not incorporate high dimensional persistent features that might produce more informative and meaningful topological features of data. For future work, this thesis did not consider an extended Cox model for identifying variable importance and RF clustering for mixed covariate features and persistent homology using high dimensional features. It would be recommended to pursue these avenues to see if new information would come to light.

Bibliography

[1] American Cancer Society. *Cancer Facts and Figures 2016*. Atlanta, Ga: American Cancer Society, 2016.

[2] Higgins JPT, Shinghal R, Gill H, *et al.* :*Gene Expression Patterns in Renal Cell Carcinoma Assessed by Complementary DNA Microarray*. Am J Pathol., 2003, 162:925 – 932.

[3] Tao Shi, David Seligson, Arie Belldegrun,*et al.*: *Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma*. Modern Pathology, 2005, 18:547 – 557.

[4] Louis S Liou, Ting Shi, Zhong-Hui Duan, *et al.*: *Microarray Gene Expression Profiling and Analysis in Renal Cell Carcinoma*. BMC Urology, 2004, 4:9

[5] Feng S, Goodrich NP, Bragg-Gresham JL, *et al.*: *Characteristics Associated with Liver Graft Failure: The Concept of a Donor Risk Index*. Am. J Transplant, 2006, 6: 783 – 790.

[6] Luciana Haddad, Alex Jones Flores Cassenote, Wellington Andraus, *et at.*: *Factors Associated with Mortality and Graft Failure in Liver Transplants: A Hierarchical Approach*. PLOS ONE — DOI: 10.1371 / 2015.

[7] Scientific Registry of Transplant receipts: *http://www.srtr.org/default.aspx* and *e*MedicineHealth:thttp://www.emedicinehealth.com/liver/article_em.htm.

[8] Hosmer D.W., Lemeshow S. and May S.: *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. Wiley Series in Probability and Statistics. Canada, 2008.

[9] Mara Tableman, Jong Sung Kim: *Survival Analysis Using S: Analysis of Time-to-Event Data*. Chapman and Hall/CRC, London, UK, 2003.

[10] Anderson P.K. and Gill R.D.: *Cox's regression model for counting process: a large sample study*. Annals of statistics. 1982, 10:1100 − 1120.

[11] Cox D.R. *Partial likelihood: Biometrica*. 1975, 62:269 − 276

[12] David Collett: *Modelling Survival Data in Medical Research*. Second Edition, Chapman and Hall/CRC, London, UK, 2005.

[13] Breiman, L.: *Random forests. Machine Learning*. 2001, 45:5 − 32.

[14] Hastie T., Tibshirani R., Friedman JH.: *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Second edition. Springer-Verlag, New York, 2009.

[15] Biau, G.: *Analysis of a random forests model*. Journal of Machine Learning Research. 2012, 13:1063–1095.

[16] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., *et al.*: *Random survival forests*. The Annals of Applied Statistics. 2008, 2:841 − 860.

[17] Strobl C, Boulesteix AL, Kneib T, *et al.*: *Conditional variable importance for random forests*. BMC Bioinformatics. 2008, 9:307−.

[18] Breiman, L., Friedman, J. H., Olshen, R. A., *et al.*: *Classification and Regression Trees*. Wadsworth, Belmont, California, 1984.

[19] Ishwaran, H., Kogalur, U. B., Chen, X., *et al.*:*Random survival forests for high-dimensional data*. Statistical Analysis and Data Mining. 2011, 4:115 − 132.

[20] Breiman, L.: *Bagging predictors*. Machine Learning. 1996, 26:123 – 140.

[21] Ishwaran, H., Kogalur, U.B. *randomForestSRC: Random Forests for Survival, Regression and Classification (RF-SRC)*. R package version 2.2.0, 2016. http://cran.r-project.org.

[22] Ishwaran H.: *The effect of splitting on random forests*. Machine Learning. 2015, 99:75 – 118.

[23] Ishwaran H, Kogalur UB. *Random survival forests for r. R News*. 2007, 7(2):25-31.

[24] G. Heo. *Topological and Statistical Data Analysis*. Notes for Math 600 course, Summer. 2013.

[25] G. Heo, J. Gamble, and P. Kim.: *Topological Analysis of Variance and the Maxillary Complex*, 2012, JASA, 107:477–492.

[26] R. Ghrist,: Barcodes: *The Persistent Topology of Data*, Bull. Amer. Math. Soc., 2008, 45(1):61 – 75.

[27] A. Zomorodian and G. Carlsson,: *Computing Persistent Homology*, 2004.

[28] A. de Silva and G. Carlsson,: *Topological Approximation by Small Simplicial Complexes*, 2003.

[29] G. Carlsson,: *Topology and Data,* Bull. Amer. Math. Soc. (N.S.). 2009, 46 no. 2, 255 – 308.

[30] W. Fedus, M. Gartner, A. Georges, *et al.*: *Persistent Homology for Mobile Phone Data Analysis*, 2015.

[31] Kaufman, L. and P.J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley Sons, New York, 2005.

[32] Tao Shi and Steve Horvath: *Unsupervised Learning with Random Forest Predictors*, Journal of Computational and Graphical Statistics. 2006, Volume 15, Number 1:118 − 138.

[33] Gilles Louppe, Louis Wehenkel, Antonio Sutera and Pierre Geurts: *Understanding variable importance in forests of randomized trees.* Dept. of EE CS, University of Liege, Belgium

[34] Sibulesky L, Li M, Hansen RN, Dick AA, *et al.*: *Impact of Cold Ischemia Time on Outcomes of Liver Transplantation: A Single Center Experience*, Ann. Transplant. 2016, 21:145 − 51.

[35] Michal Grat, Karolina M. Wronka, Waldemar Patkowski, *et al.*: *Effects of Donor Age and Cold Ischemia on Liver Transplantation Outcomes According to the Severity of Recipient Status*, Dig. Dis. Sci., 2016; 61: 626–635.

Appendix A

Appendix

Definition of Terms

Bilirubin is a yellow-brown substance formed when the liver breaks down old red blood cells. Too much bilirubin can be a sign that the liver cannot adequately remove bilirubin from the system due to blockage (e.g., gallstones, tumors), cirrhosis, or acute hepatitis. Elevated bilirubin can also indicate hemolytic anemia, a reduction in red blood cells due to abnormal breakdown of red blood cells (hemolysis).

Albumin is a small protein made in the liver that constitutes the major protein in blood serum. Albumin performs many functions in the body, including nourishing tissues, transporting various substances through the body (hormones, vitamins, drugs, and ions), and preventing fluid from leaking out of the blood vessels. Albumin concentration will drop if a person suffers from liver damage, kidney disease, malnourishment, serious inflammation, or shock.

Creatinine is produced by the muscles as they breakdown creatine, a substance involved in muscle contraction. Creatinine is formed at a constant rate in the body and excreted by the kidneys, so by evaluating the amount of creatinine in the blood, the concentration in the blood is compared to a some standard amount for a specific age and sex. Increased blood creatinine levels may indicate an increase in lupus

involvement of the kidney.

Cold Ischemic Time (Cold_Isch) is the time between the chilling of a tissue or organ after its blood supply has been reduced or cut off and the time it is warmed by having its blood supply restored.

Proportional Hazards Assumption Diagnostics Plot: Kidney Data

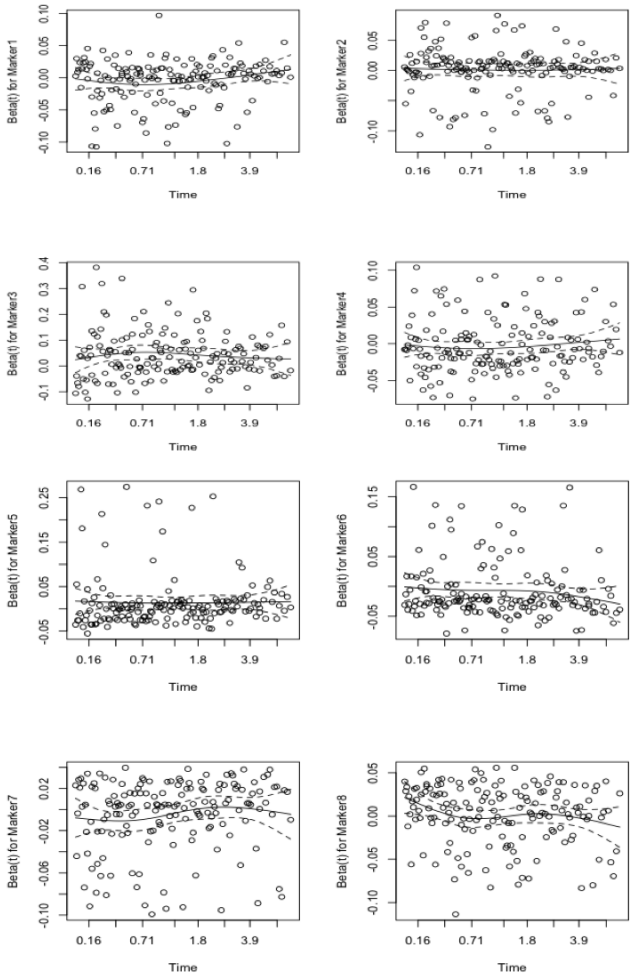


Figure 1.1: Diagnostic plots of checking the PH assumption of the coefficients for RCC data. Each plot is of a component of $\beta(t)$ against ordered time. A spline smoother is shown, together with 2 standard deviation bands.

Proportional Hazards Assumption Diagnostics Plot: Liver Data

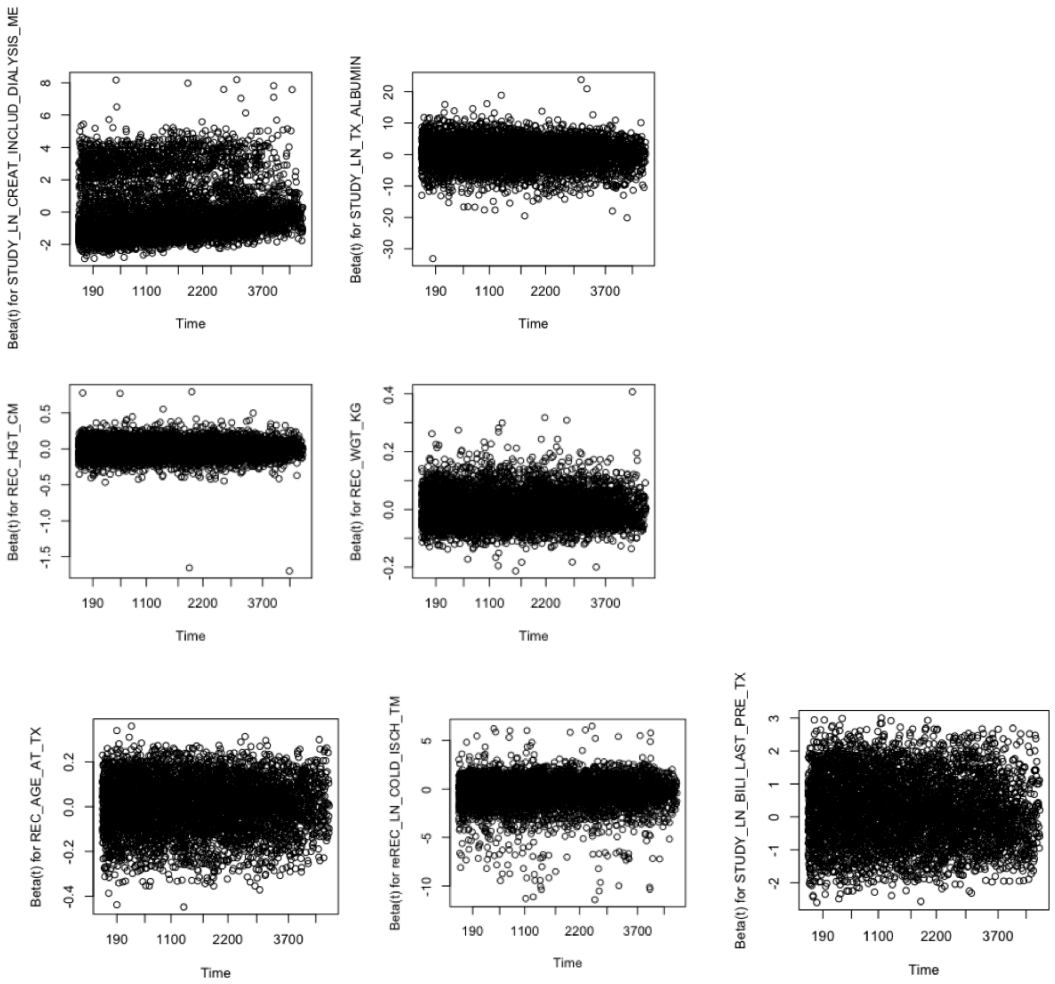


Figure 1.2: Diagnostic plots of checking the PH assumption of the coefficients for Liver Transplant data. Each plot is of a component of $\beta(t)$ against ordered time. A spline smoother is shown, together with 2 standard deviation bands.

Persistence Features at Dimension One: The Kidney Data

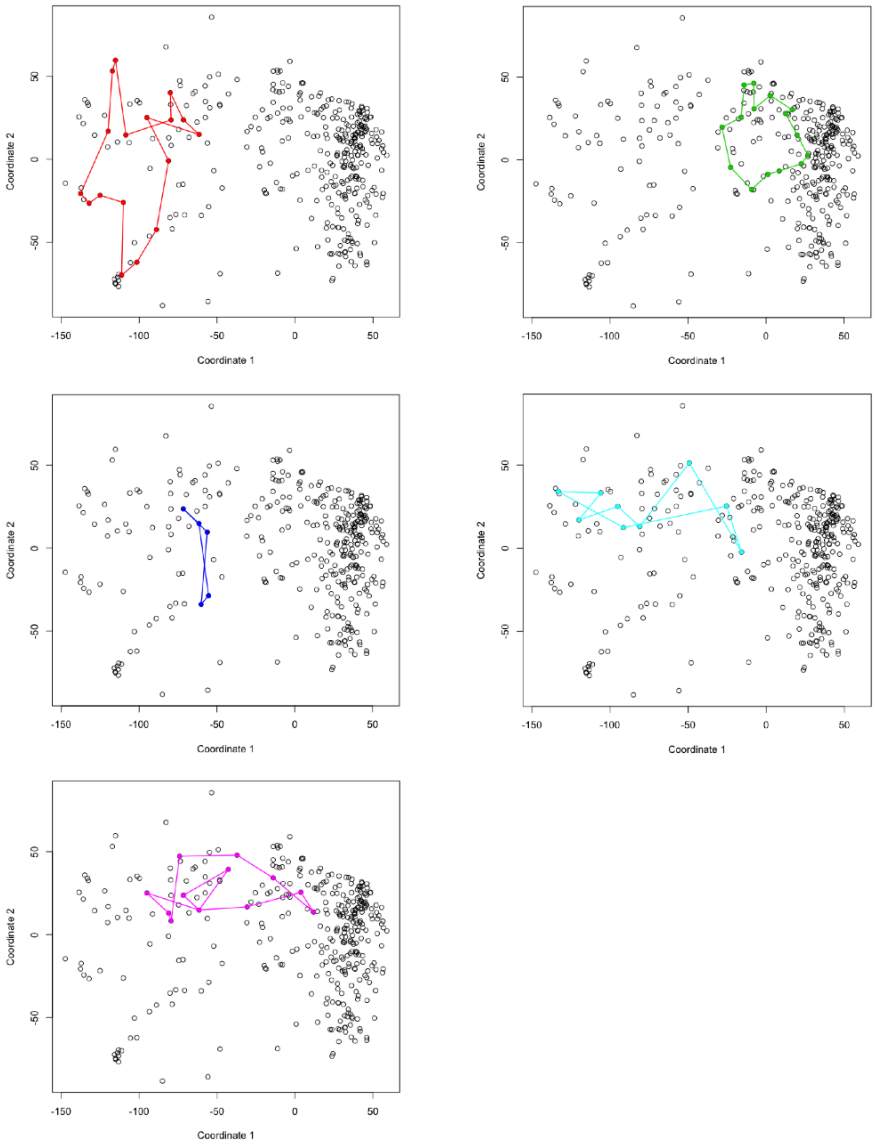


Figure 1.3: The first five most significant persistent features representation of the 366 RCC patients at dimension one: 1st, 2nd, 3rd, 4th, 5th, are represented by red, green, blue, blue-green and purple loops respectively

Persistence Features at Dimension One: The Liver Data

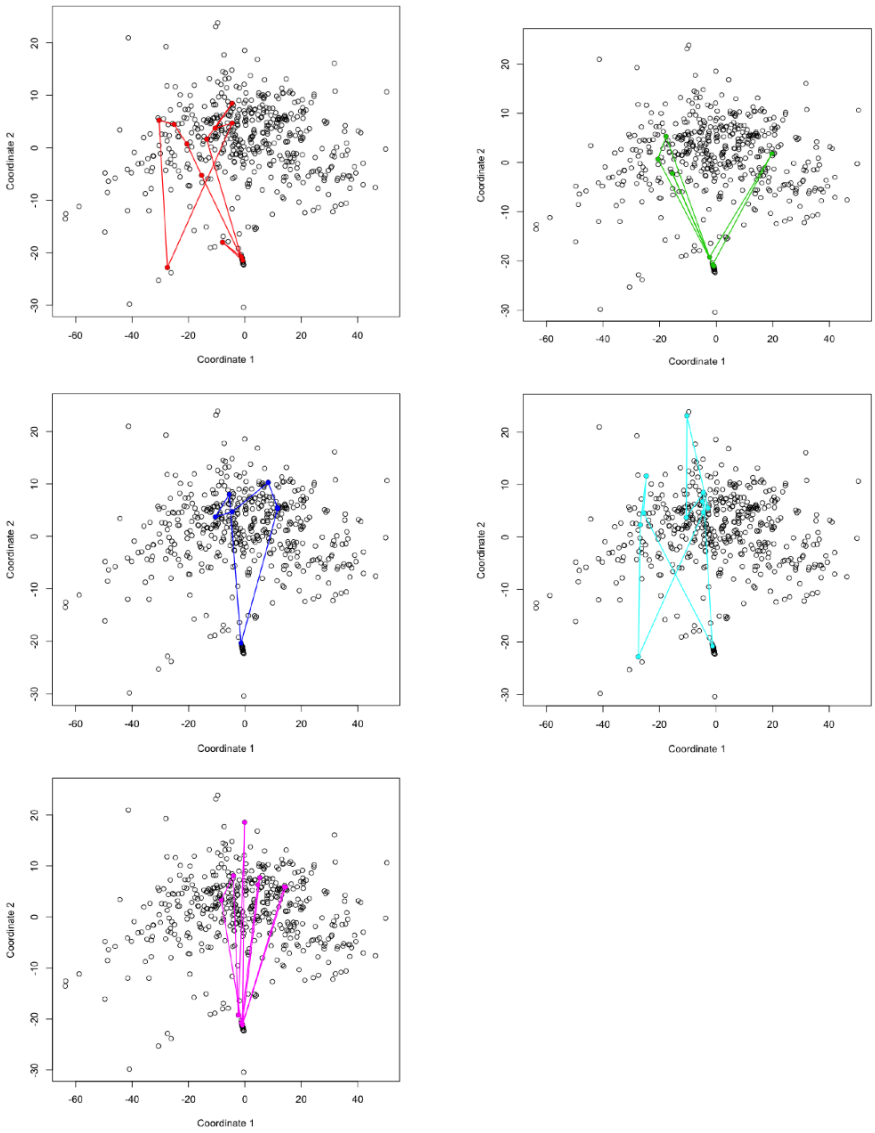


Figure 1.4: The first five most significant persistent features representation of the 500 alcoholic patients using patient characteristics only at dimension one: 1st, 2nd, 3rd, 4th, 5th, are represented by red, green, blue, blue-green and purple loops respectively

Persistence Features at Dimension One: The Liver Data

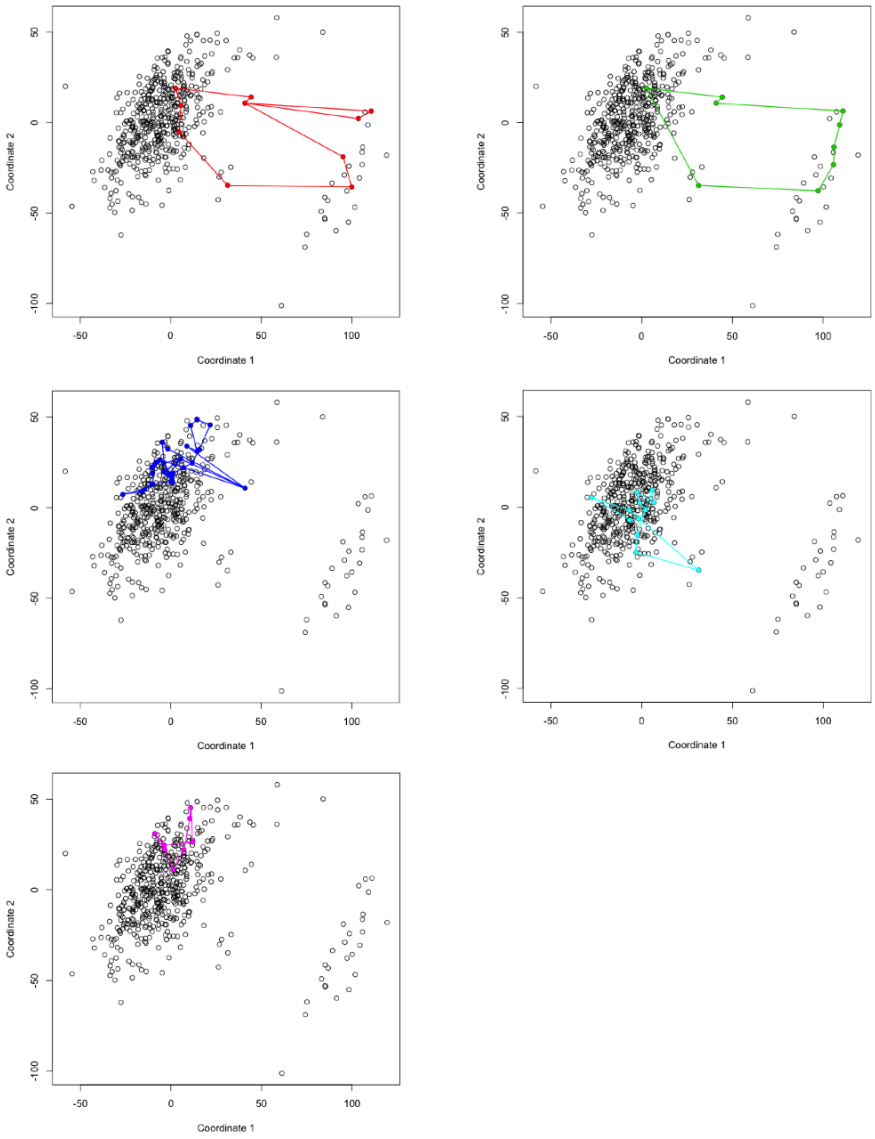


Figure 1.5: The first five most significant persistent features representation of the 500 alcoholic patients using both patient and donor characteristics at dimension one: 1st, 2nd, 3rd, 4th, 5th, are represented by red, green, blue, blue-green and purple loops respectively