CHAPTER 1

1.1 INTRODUCTION

Survival analysis is a statistical method used to analyze the time until an event of interest occurs. This method is commonly used in medical research, social sciences, engineering, and other fields where the time to an event is important. The event of interest can be anything from the onset of a disease to the failure of a machine or the occurrence of a specific behavior.

It accounts for censoring, which occurs when the event of interest has not yet occurred for some participants at the end of the study, and allows for the estimation of the probability of the event occurring over time.

It is typically involving the use of survival curves, which show the proportion of participants who have not experienced the event of interest over time. Other common techniques used in survival analysis include hazard functions, which describe the instantaneous rate at which the event occurs, and Cox proportional hazards models, which allow for the examination of the effect of multiple predictor variables on survival time

1.2 History of survival analysis

The history of survival analysis can be traced back to the work of 17th-century mathematicians, including John Graunt and Edmund Halley, who developed life tables to estimate the probability of survival and mortality rates for different age groups. In the 19th century, life tables were used to study mortality rates associated with different diseases and conditions, and the field of actuarial science emerged to develop methods for calculating life insurance premiums based on survival probabilities. In the mid-20th century, survival analysis began to be used more widely in medical research, particularly in the field of cancer research, where the goal was to estimate the survival time of patients with different types of cancer. In the 1970s, the development of the Kaplan-Meier estimator allowed for the estimation of survival curves in the presence of censoring, and the Cox proportional hazards model was developed to allow for the examination of the effect of multiple predictor variables on survival time. Since then, survival analysis has been applied to a wide range of fields, including engineering, social sciences, and economics. The development of more advanced statistical methods, such as parametric survival models and competing risk models, has expanded the scope of survival analysis to more complex data structures and research questions.

1.3 BASIC CONCEPT OF SURVIVAL ANALYSIS

1.3.1 Survival time

Survival time refers to the duration of time from a specific starting point (such as the time of diagnosis, initiation of treatment, or entry into a study) until the occurrence of an event of interest, such as death or disease progression. Survival time is also sometimes referred to as time-to-event, failure time, or time-to-failure.

It can be censored if the event of interest has not occurred for a participant by the end of the study period or if the participant is lost to follow-up before the event occurs. In these cases, the exact time to event is unknown, but it is known that the event has not occurred up to a certain point in time. Censored survival times are an important consideration in survival analysis, as they can affect the estimation of survival probabilities and hazard rates.

It is a key outcome measure in many medical and epidemiological studies, as it provides important information on the timing and probability of occurrence of events of interest, which can be used to inform clinical decision-making and the development of treatment strategies.

1.3.2 Study time

It generally refers to the duration of time over which participants are observed for the occurrence of an event of interest. The study time can vary depending on the research question and study design.

For example, in a clinical trial of a new cancer treatment, the study time might be defined as the time from treatment initiation to death or disease progression. In a study of the time to recovery from a particular illness, the study time might be defined as the time from symptom onset to recovery or until the end of the study period.

The study time can be fixed or flexible. A fixed study time means that all participants are observed for the same duration of time, regardless of whether they experience the event of interest or not. In contrast, a flexible study time means that participants are observed until they experience the event of interest, and the study ends when a certain number of events have occurred or a certain proportion of participants have experienced the event.

It is an important consideration in survival analysis because it determines the length of time over which the risk of the event of interest is being measured. Different study times can affect the estimated survival probabilities and hazard rates, and can impact the conclusions drawn from the study. It is important to carefully define the study time in advance and to use appropriate statistical methods to analyze the data.

1.3.3 Patient time

Patient time is a key concept in survival analysis because it allows for the calculation of survival probabilities and hazard rates, which describe the risk of experiencing the event of interest over time. The calculation of patient time takes into account any censoring that may occur during the study period, such as participants who are lost to follow-up or who do not experience the event of interest before the end of the study.

For example, if a study is investigating the time to relapse after treatment for a particular disease, the patient time for each participant would be calculated as the time from treatment initiation to relapse or the end of the study period, whichever comes first. If a participant is lost to follow-up before experiencing relapse, their patient time would be censored at the time of their last follow-up visit.

By calculating patient time and accounting for censoring, survival analysis can provide important information on the probability and timing of occurrence of events of interest, which can be used to inform clinical decision-making and the development of treatment strategies.

1.4 Censoring

Most of the studies, censoring refers to the situation where the exact time to an event of interest (such as death or failure) is not observed for some study subjects. This can occur when the study ends before all subjects have experienced the event, or when subjects are lost to follow-up before the event occurs.

Censoring is an important consideration in survival analysis because it affects the estimation of survival probabilities and hazard rates. The two most common types of censoring are right-censoring and left-censoring.

1.4.1 Types of censoring

➢ Type I censoring
➢ Type II censoring

Type I censoring

This occurs when all participants have experienced the event of interest by the end of the study period. This is relatively rare in survival analysis, but can occur in studies with a small sample size or a short follow-up period. Each type of censoring requires different statistical methods to properly analyze the data and obtain accurate estimates of survival probabilities and hazard rates.

Type II censoring

This occurs if an experiment has a set number of subjects or items and stops the experiment when a predetermined number are observed to have failed; the remaining subjects are then right-censored.

Right-censoring: This is the most common type of censoring in survival analysis. It occurs when the event of interest has not occurred for a participant by the end of the study period, or when the participant is lost to follow-up before the event occurs. The exact time to event is unknown, but it is known that the event has not occurred up to a certain point in time.

Left-censoring: This occurs when the event of interest has already occurred before the participant enters the study, and the exact time of the event is unknown. For example, if a study is investigating the time to diagnosis of a disease, participants who are already diagnosed with the disease at the beginning of the study are considered left-censored.

Interval-censoring: This occurs when the event of interest is only known to have occurred within a certain time interval, rather than at a specific time point. For example, if a study is investigating the time to recurrence of a cancer, and a participant's recurrence is detected at a routine follow-up appointment, the exact time of the recurrence is unknown, but it is known to have occurred within the interval between the previous appointment and the current appointment.

Informative censoring: This occurs when the likelihood of censoring is related to the outcome being studied. For example, if participants with more severe disease are more likely to drop out of a study, the censoring is considered informative and can bias the results if not properly accounted for.

1.4.2 Reason for censoring

Study design: The study may be designed to end at a certain point in time, regardless of whether all participants have experienced the event of interest. For example, a clinical trial may be designed to end after a certain number of years, even if some participants have not yet developed the disease being studied.

Loss to follow-up: Participants may drop out of the study or become lost to follow-up, meaning that their event status is unknown. This can happen for various reasons, such as moving away, refusing to continue participation, or being difficult to contact.

Event occurrence outside of study time frame: Some participants may experience the event of interest before the study begins or after it ends. In these cases, the exact time of the event is unknown and the data is said to be left-censored or right-censored, respectively.

Incomplete data: In some cases, data may be missing or incomplete, making it impossible to determine the exact time of the event or censoring. It is important to properly account for censoring in survival analysis to ensure accurate estimates of survival probabilities and hazard rates. Failure to do so can lead to biased results and incorrect conclusions.

1.3.3 Survival function

The survival function, also known as the survival probability function, is a fundamental concept in survival analysis. It is a function that describes the probability that an individual survives beyond a certain time t. In other words, it gives the proportion of individuals who have not experienced the event of interest (such as death or disease progression) up to time t.

The survival function is often denoted by the symbol S(t) and is defined as:

$$S(t) = P(T > t)$$

where $, \lim_{t \to 0} P(T > t),$

T is the survival time and P(T > t) is the probability that the survival time is greater than t.

The survival function can be estimated using statistical methods such as the Kaplan-Meier method or the Nelson-Aalen estimator. The survival function can be used to estimate the probability of survival at different time points, and to compare the survival probabilities between different groups of individuals or under different conditions. For example, in a clinical trial of a new cancer treatment, the survival function can be used to estimate the probability of survival at different time points for patients who receive the new treatment versus those who receive standard treatment.

The survival function can also be used to calculate other important measures in survival analysis, such as the median survival time (the time at which 50% of individuals have experienced the event of interest) and the hazard function (the rate at which individuals experience the event of interest).

Overall, the survival function is a crucial tool in survival analysis for describing and comparing the probability of survival over time, and for informing clinical decision making and the development of treatment strategies.

1.4.4 Hazard function

The hazard function, also known as the failure rate, is a concept in statistics and reliability theory that describes the probability of a failure occurring in a given time interval, given that the item being observed has survived up to that time. In other words, it is the instantaneous rate of failure at a given time, given that the item has not failed before that time. Mathematically, the hazard function h(t) is defined as the probability of failure occurring in the time interval (t, t + dt) given that the item has survived up to time t, divided by dt:

$$\mathbf{h(t)} = \lim_{dt \to 0} \frac{pr(t \leq T \leq t + dt | T \geq t)}{dt}$$

where T is the random variable representing the time of failure.

The hazard function is an important concept in survival analysis, which is the study of time-to-event data, such as the time until death, failure of a machine, or recurrence of a disease. It is used to model and analyze the risk of failure over time and can be used to compare the failure rates of different groups or to estimate the remaining useful life of a product or system.

1.4.5 Cumulative hazard function

The cumulative hazard function is a concept in statistics and reliability theory that is closely related to the hazard function. It is defined as the cumulative sum of the hazard function up to a given time t, and represents the total amount of risk or failure experienced up to that time.

Mathematically, the cumulative hazard function H(t) is defined as:

$$\mathbf{H(t) = \int[0,t] \ h(u) \ du,}$$

where h(u) is the hazard function at time u.

The cumulative hazard function can be interpreted as the expected value of the total number of failures up to time t. It is a monotonically increasing function, reflecting the fact that the amount of risk or failure experienced by a system increases over time.

The cumulative hazard function is commonly used in survival analysis to model and analyze time-to-event data. It can be used to estimate the probability of failure or survival at a given time, and to compare the survival or failure rates of different groups or systems. It is also useful for estimating the reliability or remaining useful life of a product or system.

1.5 Application of survival analysis

There are three primary goals of survival analysis, to estimate and interpret survival and or hazard functions from the survival data; to compare survival and or hazard functions, and to assess the relationship of explanatory variables to survival

time. Survival analysis provides a great tool for analyzing the time to an event type of data, which is very common in any clinical trial. Researchers are not using it frequently because they are not confident in the theory of its application and its interpretation.

1.6 Non parametric Methods

Non-parametric tests are commonly used in survival analysis to compare the survival functions of two or more groups. Survival analysis is a statistical method used to analyze time-to-event data, such as the time until death or failure of a machine. The goal of survival analysis is to estimate the probability of survival or failure as a function of time, and to compare the survival or failure rates of different groups or treatments.

Kaplan Meier Analysis:

The Kaplan-Meier analysis, also known as the Kaplan-Meier curve or survival analysis, is a statistical method used to estimate the probability of an event occurring over time. It is commonly used in medical research to analyze the survival time of patients with a certain disease or condition. It's curve is a graphical representation of the probability of survival or event-free survival over time. The curve is constructed by plotting the proportion of individuals surviving or event-free at each time point, with the time points determined by the occurrence of events, such as death or disease progression.

The Kaplan-Meier analysis takes into account censoring, which occurs when individuals are lost to follow-up or the study ends before all individuals have experienced the event of interest. Censoring is represented in the Kaplan-Meier curve by vertical lines at the censored time points. It can be used to compare the survival or event-free survival of different groups of individuals, such as those receiving different treatments or with different genetic profiles. The log-rank test is commonly used to compare the survival curves of two or more groups.

$$\widehat{S}(t) \ = \ \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

with t_i a time when at least one event happened, d_i the number of events (e.g., deaths) that happened at time t_i and n_i the individuals known to have survived (have not yet had an event or been censored) up to time t_i

Some common non-parametric tests used in survival analysis include:

Log-rank test: The log-rank test is used to compare the survival functions of two or more groups. It is a non-parametric test that is based on the difference between the observed and expected number of events in each group.

Wilcoxon test: The Wilcoxon test is used to compare the survival functions of two groups. It is a non-parametric test that is based on the difference between the median survival times of the two groups.

Tarone-Ware test: The Tarone-Ware test is used to compare the survival functions of two or more groups. It is a non-parametric test that is based on a weighted combination of the log-rank and Wilcoxon tests.

Gehan-Breslow-Wilcoxon test: The Gehan-Breslow-Wilcoxon test is used to compare the survival functions of two groups. It is a non-parametric test that is based on a weighted sum of the number of events at each time point.

Non-parametric tests are useful in survival analysis because they do not require assumptions about the distribution of the survival times. However, they may have less power than parametric tests when the assumptions of the latter are met. It is important to choose the appropriate test based on the type of data and the research question being investigated.

1.7 Semi parametric Methods

A semi-parametric test is a statistical test that combines non-parametric and parametric methods to analyze data. Semi-parametric methods are used when the data is partially specified by a parametric model and partially unspecified or unknown.

In the context of survival analysis, a popular semi-parametric method is the Cox proportional hazards model. The Cox proportional hazards model assumes that the hazard function for each group is proportional over time, but it does not specify the functional form of the hazard function. This allows for a flexible model that can handle various types of survival data, including censored data, while also incorporating covariates that affect the survival time. The Cox model is a semi-parametric method because it uses a parametric model for the hazard ratios but not for the baseline hazard function.

Other examples of semi-parametric tests include the generalized estimating equation (GEE) and the generalized linear mixed model (GLMM), which combine parametric and non-parametric methods to analyze correlated data.

Semi-parametric tests have the advantage of being more flexible than purely parametric methods and more efficient than purely non-parametric methods, especially when the parametric assumptions are only partially met. However, they may require more computational resources and more complex statistical models than purely parametric or non-parametric methods.

COX REGRESSION

Cox regression, also known as the Cox proportional hazards model, is a statistical method used to analyze the relationship between one or more predictor variables and a survival outcome.

The Cox regression model is a semi-parametric method that estimates the hazard ratio, which is a measure of the relative risk of an event occurring in one group compared to another group. The hazard ratio represents the ratio of the hazard rates, or the instantaneous probability of an event occurring at a particular time, between two groups. A hazard ratio of 1 indicates no difference in the hazard rates between the two groups, while a hazard ratio greater than 1 indicates a higher hazard rate in one group compared to the other.

This model assumes that the hazard ratio is constant over time, which is known as the proportional hazards assumption. This means that the effect of the predictor variable(s) on the outcome is constant over time, and does not change with the length of follow-up. The model also assumes that the hazard rate follows a specific distribution, such as the exponential or Weibull distribution.

$$h(t) = h_0(t) * exp(b_1 x_1 + b_2 x_2 + \ldots + b_n x_n)$$

Where,

t=survival time

h(t)=hazard function

$x_1 + x_2 + \ldots + x_n$ = covariates

$b_1 + b_2 + \ldots + b_n$ = measures of impact of covariates or biases for each covariates

The term h_0 is called the baseline hazard.

The Cox regression can be used with both continuous and categorical predictor variables, and can also handle time-dependent covariates. It is a useful method for studying the impact of multiple factors on survival outcomes and can help identify important prognostic factors for a given disease or condition.

The Cox proportional hazards model assumes that the hazard rate is proportional across all levels of the predictor variables. This means that the effect of the predictor variables on the hazard rate is constant over time. The model does not assume a particular distribution for the hazard rate, which makes it a semi-parametric model.

Significance of Cox-Ph model over the others

- Here, the interest is mostly to find which covariate has more significance effect on the survivability of a subject. Hence, the more important point of interest is the hazard ration. Hence, e^{b_i} the application of the cox-proportional model seems like a better fit. Any other models could have been chosen if main point of interest was on the prediction of an individual surviving beyond a particular time.

- The Cox-Ph model allows the HAZ to fluctuate along with time (it is a function of time). On the other hand, the Weibull model has the HAZ which is

proportional to time i.e., the HAZ either increases or decreases with proportion to time.

- While other models are used for comparing only one variable at a time and operating on categorical variables like gender, status etc., the Cox-Ph model can be used for non-categorical variables like age, weight, height etc.

Hazard Ratio - $exp(bi)$ or e^{b_i}:

The value of exp(bi) is called the Hazard Ratio (HR) and it has the following significance.

HR = 1: No effect

HR < 1: Reduction in Hazard

HR > 1:Increase in Hazard

RESIDUALS

Residuals refer to the differences between the observed and predicted values of the survival time. These residuals are used to assess the goodness of fit of the Cox regression model, as well as to identify outliers and influential observations.

There are three types of residuals namely,

- Schoenfeld residuals
- Deviance residual
- Martingale residual

Schoenfeld residuals

Let $r_i(t) = e^{\widehat{\beta}x_i(t)}$ be the estimated hazard ratio for the ith subject at t compared to $x(t) = 0$.

Then for $\bar{x}(\hat{\beta}, t) = \frac{\sum at\ risk\ at\ t^{r_i(t)x_i(t)}}{\sum at\ risk\ at\ t^{r_i(t)}}$, The Schoenfeld residual for the kth subject failing at time t is given by $x_k(t) - \bar{x}(\hat{\beta}, t)$

The scaled Schoenfeld residual is the Schoenfeld residual divided by a variance estimate. Grambsch and Therneau (1994) showed that the scaled Schoenfeld residual measures the deviation of a time-dependent log hazard ratio β(t) from time-constant $\hat{\beta}$. Can use linear regression comparing scaled Schoenfeld residuals to functions of time to examine evidence for lack of constant hazard ratio over time.

Diagnostics for the Cox model

The Cox proportional hazards model makes several assumptions. Thus, it is important to assess whether a fitted Cox regression model adequately describes the data.

Here, we'll discuss three types of diagnostics for the Cox model:

- Testing the proportional hazards assumption.
- Examining influential observations (or outliers).
- Detecting nonlinearity in relationship between the log hazard and the covariates.

In order to check these model assumptions, *Residuals* method are used. The common residuals for the Cox model include:

- Schoenfeld residuals to check the proportional hazards assumption
- Martingale residual to assess nonlinearity
- Deviance residual (symmetric transformation of the Martinguale residuals), to examine influential observations

Testing proportional Hazards assumption

The proportional hazards (PH) assumption can be checked using statistical tests and graphical diagnostics based on the *scaled Schoenfeld residuals*.

In principle, the Schoenfeld residuals are independent of time. A plot that shows a non-random pattern against time is evidence of violation of the PH assumption.

The function *cox.zph*() [in the *survival* package] provides a convenient solution to test the proportional hazards assumption for each covariate included in a Cox regression model fit.

For each covariate, the function *cox.zph*() correlates the corresponding set of scaled Schoenfeld residuals with time, to test for independence between residuals and time. Additionally, it performs a global test for the model as a whole.

The proportional hazard assumption is supported by a non-significant relationship between residuals and time, and refuted by a significant relationship.

To test influential observations or outliers, we can visualize either:

- the deviance residuals or
- the df beta values

Testing non linearity Often, we assume that continuous covariates have a linear form. However, this assumption should be checked.

Plotting the Martingale residuals against continuous covariates is a common approach used to detect nonlinearity or, in other words, to assess the functional form of a covariate. For a given continuous covariate, patterns in the plot may suggest that the variable is not properly fit. Nonlinearity is not an issue for categorical variables, so

we only examine plots of martingale residuals and partial residuals against a continuous variable.

Martingale residuals

A value of martingale residuals near 1 represents individuals that "died too soon", and large negative values correspond to individuals that "lived too long". To assess the functional form of a continuous variable in a Cox proportional hazards model, the function ggcox functional() [in the survminer R package].

The function ggcox functional() displays graphs of continuous covariates against martingale residuals of null cox proportional hazards model. This might help to properly choose the functional form of continuous variable in the Cox model. Fitted lines with lowness function should be linear to satisfy the Cox proportional hazards model assumptions.

Logistic regression

Logistic regression has been applied to numerous investigations that examine the relationship between risk factors and various disease events. Recently, the ability to consider the time element of event occurrences by proportional hazards models has meant that logistic regression has played a less important role in the analysis of survival data. This paper, however, shows that when event times are grouped into intervals, logistic regression can be adapted to the analysis of such data by modeling the interval when an event occurs. Furthermore, it is shown that results from such an adaptation will often lead to parameter estimates close to those obtained by the proportional hazards model in the grouped event time setting. An illustration of the application of logistic regression to survival analysis is based on data from the Framingham Heart Study his type of statistical model (also known as logit model) is often used for classification and predicative analytics. In logistic regression, a logit transformation is applied on the odds-that is , the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

$$\textbf{Logit(pi) = 1/(1+ exp(-pi))}$$

$$\textbf{ln(pi/(1-pi)) = Beta\_0 + Beta\_1*X\_1 + … + B\_k*K\_k}$$

In this logistic regression equation, logit(pi) is the dependent or response variable and x is the independent variable. The beta parameter, or coefficient, in this model is commonly estimated via maximum likelihood estimation (MLE). This method tests different values of beta through multiple iterations to optimize for the best fit of log odds. All of these iterations produce the log likelihood function, and logistic regression seeks to maximize this function to find the best parameter estimate. Once the optimal coefficient (or coefficients if there is more than one independent variable)

is found, the conditional probabilities for each observation can be calculated, logged, and summed together to yield a predicted probability. For binary classification, a probability less than .5 will predict 0 while a probability greater than 0 will predict 1. After the model has been computed, it's best practice to evaluate the how well the model predicts the dependent variable, which is called goodness of fit. The Hosmer–Lemeshow test is a popular method to assess model fit.

Interpreting logistic regression

Log odds can be difficult to make sense of within a logistic regression data analysis. As a result, exponentiating the beta estimates is common to transform the results into an odds ratio (OR), easing the interpretation of results. The OR represents the odds that an outcome will occur given a particular event, compared to the odds of the outcome occurring in the absence of that event. If the OR is greater than 1, then the event is associated with a higher odds of generating a specific outcome. Conversely, if the OR is less than 1, then the event is associated with a lower odds of that outcome occurring. Based on the equation from above, the interpretation of an odds ratio can be denoted as the following: the odds of a success changes by $\exp(cB\_1)$ times for every c-unit increase in x. To use an example, let's say that we were to estimate the odds of survival on the Titanic given that the person was male, and the odds ratio for males was .0810. We'd interpret the odds ratio as the odds of survival of males decreased by a factor of .0810 when compared to females, holding all other variables constant.

- **Binary logistic regression:** In this approach, the response or dependent variable is dichotomous in nature—i.e. it has only two possible outcomes (e.g. 0 or 1). Some popular examples of its use include predicting if an e-mail is spam or not spam or if a tumor is malignant or not malignant. Within logistic regression, this is the most commonly used approach, and more generally, it is one of the most common classifiers for binary classification.

1.8 OBJECTIVE OF THE STUDY

From the given data, we are interested in finding the estimate of survival function using non parametric method namely Kaplan Meier analysis. Also, we compare the survival experience of two or more groups using log rank test. we use semi parametric method namely Cox Proportional Hazard model and Cox Proportional assumptions using Residuals. By using software SPSS, Python, R Programming.

CHAPTER 2

2.1 AN OVERVIEW OF CANCER

Cancer is a term used to describe a group of diseases that involve abnormal cell growth and the potential for those cells to invade other tissues and organs. In normal longer needed. Sometimes this orderly process breaks down, and abnormal or damaged cells, growth and division is tightly controlled and cells die off when they are no cells grow and multiply when they shouldn't. These cells may form tumors, which are lumps of tissue. Tumors can be cancerous or not cancerous (benign).

However, in cancer cells, the normal controls over growth and division are disrupted, leading to uncontrolled growth and the potential for these cells to invade nearby tissues and spread to other parts of the body (a process called metastasis). Cancerous tumors may also be called malignant tumors. Many cancers form solid tumors, but cancers of the blood, such as leukemias, generally do not.

Benign tumors do not spread into, or invade, nearby tissues. When removed, benign tumors usually don't grow back, whereas cancerous tumors sometimes do. Benign tumors can sometimes be quite large, however. Some can cause serious symptoms or be life threatening, such as benign tumors in the brain.

CAUSES:

Cancer is caused by a combination of genetic, environmental, and lifestyle factors. Some of the known risk factors for cancer include:

Genetic mutations: Some people inherit gene mutations that increase their risk of developing certain types of cancer. These mutations may be passed down from one or both parents.

Environmental factors: Exposure to certain chemicals, radiation, or viruses can increase the risk of cancer. For example, exposure to ultraviolet (UV) radiation from the sun or tanning beds can increase the risk of skin cancer.

Lifestyle factors: Certain lifestyle choices, such as tobacco use, heavy alcohol consumption, a poor diet, and lack of physical activity, can increase the risk of cancer.

Age: The risk of cancer increases with age, as the body's cells are more likely to accumulate genetic mutations over time.

Family history: Some types of cancer, such as breast, ovarian, and colon cancer, can run in Families. Cancer can affect any part of the body and there are many different types of cancer. Some of the most common types of cancer include:

Lung cancer: Lung cancer is the leading cause of cancer death worldwide. It usually starts in the cells lining the bronchi and can spread to other parts of the body.

Breast cancer: Breast cancer is the most common cancer in women and can also occur in men. It usually begins in the cells of the breast ducts or lobules.

Colorectal cancer: Colorectal cancer affects the colon and rectum and usually develops from precancerous polyps in the colon or rectum.

Prostate cancer: Prostate cancer is the most common cancer in men and usually develops in the cells of the prostate gland.

Skin cancer: Skin cancer is the most common cancer in the United States and can occur anywhere on the skin. The most common types of skin cancer are basal cell carcinoma, squamous cell carcinoma, and melanoma.

Leukemia: Leukemia is a cancer of the blood and bone marrow, which produces abnormal white blood cells.

Lymphoma: Lymphoma is a cancer that affects the lymphatic system, which is part of the immune system.

These are just a few examples of the many types of cancer that can occur. The specific causes, symptoms, and treatments for each type of cancer can vary widely, and early detection and treatment are important for improving outcomes.

Treatments for cancer depend on the type and stage of the cancer, as well as other individual factors such as age and overall health. Treatments may include surgery, radiation therapy, chemotherapy, immunotherapy, and targeted therapy, among others. Early detection and treatment can improve outcomes and increase the chances of a cure, although cancer can be a difficult disease to treat and manage.

2.2 Structure of cancer cell
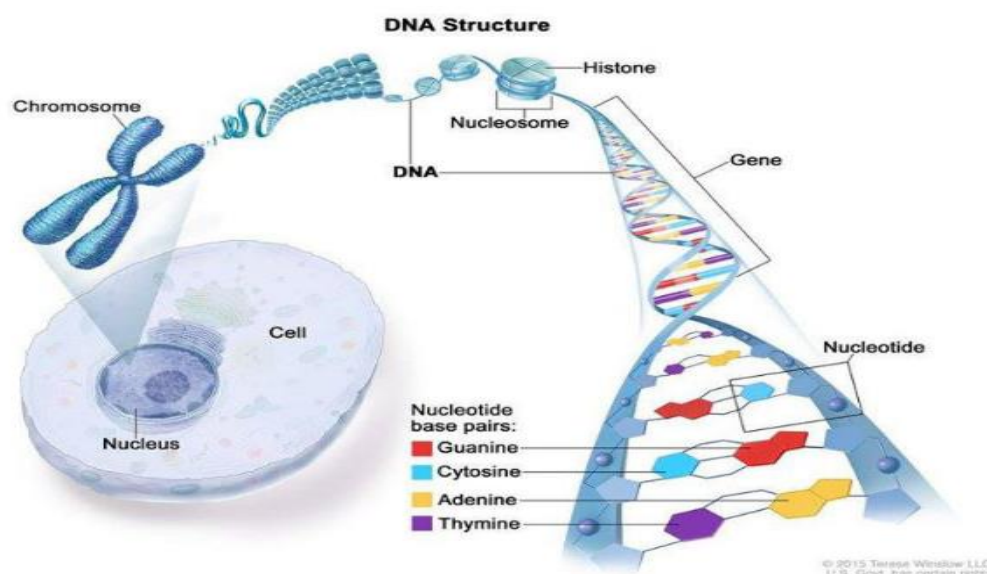
Cancer is caused by certain changes to genes, the basic physical units of inheritance. Genes are arranged in long strands of tightly packed DNA called chromosomes.

2.2.1 Characteristics of Cancer Cells

Cancer cells grow and divide at an abnormally rapid rate, are poorly differentiated, and have abnormal membranes, cytoskeletal proteins, and morphology. The abnormality in cells can be progressive with a slow transition from normal cells to benign tumors to malignant tumors.

They characteristic of cancer are as follows,

- ignore signals that normally tell cells to stop dividing or to die (a process known as <u>programmed cell death</u>, or <u>apoptosis</u>).

- invade into nearby areas and spread to other areas of the body. Normal cells stop growing when they encounter other cells, and most normal cells do not move around the body.

- tell blood vessels to grow toward tumors. These blood vessels supply tumors with oxygen and nutrients and remove waste products from tumors.

- hide from the immune system. The immune system normally eliminates damaged or abnormal cells.

- trick the immune system into helping cancer cells stay alive and grow. For instance, some cancer cells convince immune cells to protect the tumor instead of attacking it.

- accumulate multiple changes in their chromosomes, such as duplications and deletions of chromosome parts. Some cancer cells have double the normal number of chromosomes.

- rely on different kinds of nutrients than normal cells. In addition, some cancer cells make energy from nutrients in a different way than most normal cells. This lets cancer cells grow more quickly.
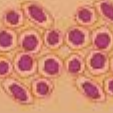
Many times, cancer cells rely so heavily on these abnormal behaviors that they can't survive without them. Researchers have taken advantage of this fact, developing therapies that target the abnormal features of cancer cells. For example, some cancer therapies prevent blood vessels from growing toward tumors, essentially starving the tumor of needed nutrients.

2.2.2 Development of Cancer

Cancer cells are always present in the human body, however they are typically recognized by the immune system and destroyed before they cause any problems. It is when cancer cells go unrecognized and begin to multiply that they become a burden. Cancer cells differ from healthy cells in critical ways that make them harmful to the body.
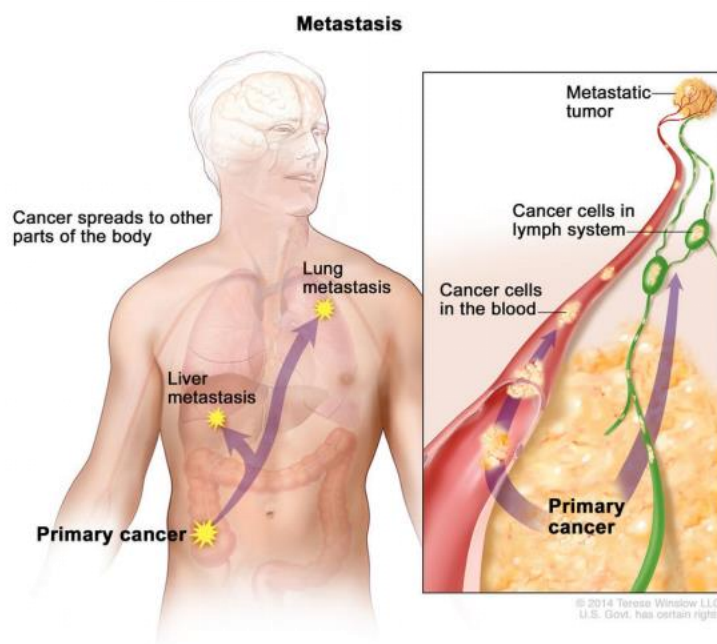
Cancer cells are unlike any other cell in the body and they act totally out of line with normal cell functions. As they develop, they begin to ignore the signals sent by

the body telling them when and when not to grow or even when to die. When they are allowed to accumulate, they begin to stray even further from normal acquiring unique traits affecting the ways they communicate, specialize, and locate themselves in tissues.

Normal Cells 👍		👎 Cancer Cells
Small, uniformly shaped nuclei Relatively large cytoplasmic volume		Large, variable shaped nuclei Relatively small cytoplasmic volume
Conformity in cell size and shape Cells arranged into discrete tissues		Variation in cell size and shape Disorganised arrangement of cells
May possess differentiated cell structures Normal presentation of cell surface markers		Loss of normal specialised features Elevated expression of certain cell markers
Lower levels of dividing cells Cell tissues clearly demarcated		Large number of dividing cells Poorly defined tumor boundaries

DRJOCKERS.com

When Cancer Spreads:

Metastasis

Cancer spreads to other parts of the body

Lung metastasis

Liver metastasis

Primary cancer

Metastatic tumor

Cancer cells in lymph system

Cancer cells in the blood

Primary cancer

© 2014 Terese Winslow LLC
U.S. Govt. has certain rights

In metastasis, cancer cells break away from where they first formed and form new tumors in other parts of the body.

A cancer that has spread from the place where it first formed to another place in the body is called metastatic cancer. The process by which cancer cells spread to other parts of the body is called metastasis.

Metastatic cancer has the same name and the same type of cancer cells as the original, or primary cancer. For example, breast cancer that forms a metastatic tumor in the lung is metastatic breast cancer, not lung cancer.

Under a microscope, metastatic cancer cells generally look the same as cells of the original cancer. Moreover, metastatic cancer cells and cells of the original cancer usually have some molecular features in common, such as the presence of specific chromosome changes.

In some cases, treatment may help prolong the lives of people with metastatic cancer. In other cases, the primary goal of treatment for metastatic cancer is to control the growth of the cancer or to relieve symptoms it is causing. Metastatic tumors can cause severe damage to how the body functions, and most people who die of cancer die of metastatic disease.

2.2.3 Types of Cancer Treatment

There are many types of cancer treatment. The types of treatment that you have will depend on the type of cancer you have and how advanced it is. Some people with cancer will have only one treatment. But most people have a combination of treatments, such as surgery with chemotherapy and/or radiation therapy. You may also have immunotherapy, targeted therapy, or hormone therapy.

2.2.4 Side Effects of Cancer Treatment

Cancer treatments and cancer can cause side effects. Side effects are problems that occur when treatment affects healthy tissues or organs.

Learn about steps you can take to prevent or manage the side effects listed below:

- Anemia
- Appetite Loss
- Bleeding and Bruising (Thrombocytopenia)
- Constipation
- Delirium

2.3 Prevention of cancer

Cancer that are closely linked to certain behaviors are the easiest to prevent. Some of the preventive measures are given as below,

Stop Smoking:

If you smoke, you should quit. Smoking is by far the leading risk factor for lung cancer, and it contributes to other cancers such as mouth, throat, cervical and bladder cancer. Your body begins recovering from smoking within minutes after quitting, and your risks for many cancers are cut in half five years after you quit. .

Be More Active:

Physical activity reduces your risk for several types of cancer, including breast and colon cancer. The American Cancer Society recommends that adults get at least 150 minutes of moderate-intensity activity each week. Children and teens should get at least 1 hour of moderate-intensity activity each week. Moderate activity includes things like walking, playing golf, doing yoga, and even mowing the lawn..

Eat More Plants:

Improving your diet is one of the most important things you can do to stay healthy, and certain foods can impact your cancer risk. The American Cancer Society recommends eating a healthy diet, with an emphasis on plant foods. Eat at least 2½ cups of vegetables and fruits each day, and minimize your intake of processed meats such as hot dogs and lunch meats. If you do eat meat, choose fish, poultry or lean cuts of beef.

Limit Alcohol Intake:

Drinking alcohol raises your risk for breast, throat, liver, colorectal and other cancers. The less you drink, the lower your risk. If you do drink, do so in moderation—no more than one drink a day for women, and no more than two drinks a day for men.

Watch Your Weight:

Too much weight around your midsection increases your risk for several types of cancer, including breast and colon cancer, as well as for other diseases like diabetes and heart disease.

Stand Up More:

Recent studies have found that people who spend most of their day sitting are more likely to develop colon and endometrial cancer. Be sure to stand up and walk around every two hours.

Protect Your Skin:

The sun, sunlamps and tanning beds all give off ultraviolet rays that cause skin damage and can lead to skin cancer. When you're in the sun, always apply sunscreen to your skin that has at least a sun protection factor (SPF) of 15. Some doctors

recommend at least SPF 30. Wear sun-protective clothing, hats that protect your skin and sunglasses to protect the skin around your eyes.

Get Vaccinated:

Scientists have developed some vaccines that protect your body from viruses like HPV that cause cancer. The U.S. Centers for Disease Control and Prevention recommends that all boys and girls get the HPV vaccine at age 11 or 12.

Get Screened:

Maintaining routine medical care is one of the best ways to lower your cancer risk. You should get a physical every year and talk to your primary care physician about which screening tests you need. If you don't have a primary care physician or Your doctor may recommend these screening tests. Everyone should get a colonoscopy to screen for colorectal cancer starting at age 45. Women should begin mammograms at age 50. Men who turn 50 should talk to their doctor about whether to get tested for prostate cancer. If you're a smoker or ex-smoker, you should talk to your doctor about a CT scan for lung cancer. You should see a dermatologist every year to check your skin from head to toe for unusual spots or moles that could lead to skin cancer.

2.4 Lung Cancer

When you breathe in, air enters through your mouth or nose and goes into your lungs through the trachea (wind pipe). The trachea divides into tubes called bronchi, which enter the lungs and divide into smaller bronchi. These divide to form smaller branches called bronchioles. At the end of the bronchioles are tiny air sacs known as alveoli.

Lung cancers typically start in the cells lining the bronchi and parts of the lung such as the bronchioles or alveoli.
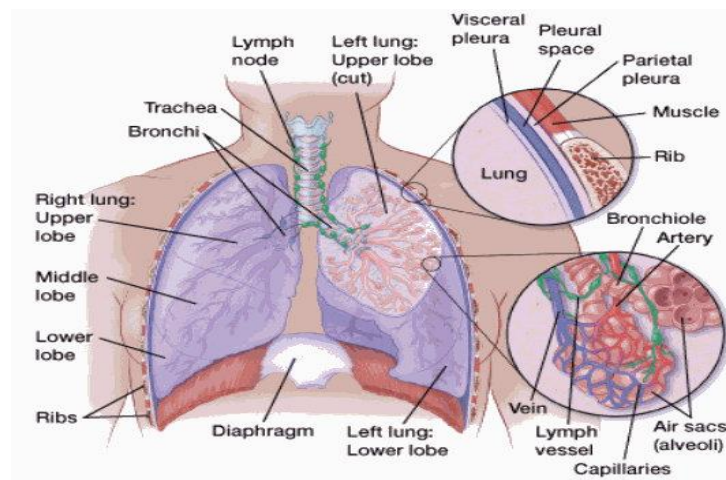
Normal structure and function of the lung

2.4.1 Types of lung cancer

There are 2 main types of lung cancer. They are Non-small cell lung cancer and Small cell lung cancer

Lung cancer

Lungs

© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.
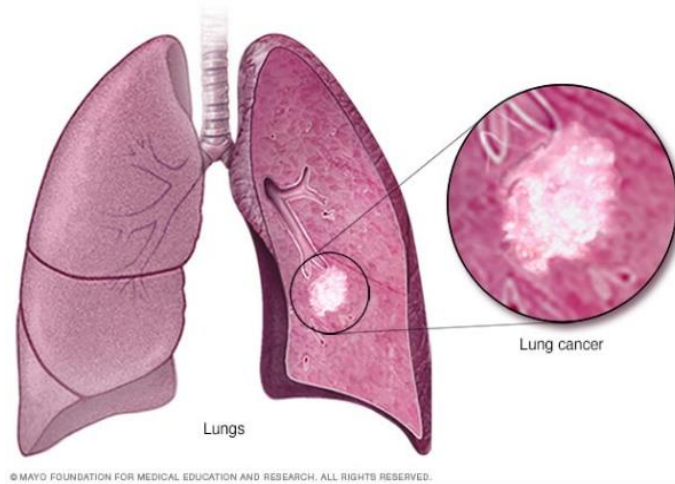
2.4.2 Non-small cell lung cancer (NSCLC)

About 80% adenocarcinoma, squamous cell carcinoma to 85% of lung cancers are NSCLC. The main subtypes of NSCLC are, and large cell carcinoma. These subtypes, which start from different types of lung cells, are grouped together as NSCLC because their treatment and prognoses (outlook) are often similar.

Adenocarcinoma

Adenocarcinomas start in the cells that would normally secrete substances such as mucus. This type of lung cancer occurs mainly in people who smoke or used to smoke, but it is also the most common type of lung cancer seen in people who don't smoke. It is more common in women than in men, and it is more likely to occur in younger people than other types of lung cancer. Adenocarcinoma is usually found in the outer parts of the lung and is more likely to be found before it has spread. People with a type of adenocarcinoma called **adenocarcinoma institute**(previously called **bronchioloalveolar carcinoma**) tend to have a better outlook than those with other types of lung cancer.

Squamous cell carcinoma:

Squamous cell carcinomas start in squamous cells, which are flat cells that line the inside of the airways in the lungs. They are often linked to a history of smoking and tend to be found in the central part of the lungs, near a main airway (bronchus).

Large cell (undifferentiated) carcinoma:

Large cell carcinoma can appear in any part of the lung. It tends to grow and spread quickly, which can make it harder to treat. A subtype of large cell carcinoma, known as **large cell neuroendocrine carcinoma (LCNEC),** is a fast-growing cancer that is very similar to small cell lung cancer.

Not otherwise specified (NOS) :

The histology subtype of non-small -cell lung cancer (NSCLC) is a significant factor when selecting treatment strategies. However cases are occasionally encountered that are diagnosed as not otherwise specified (NOS) prior to surgery , due to an uncertain histological subtype.

Diagnosis in non small cell lung cancer in survival analysis

Most lung carcinomas are diagnosed at an advanced stage and, as such, have a poor prognosis. The 5-year relative survival rates vary with the stage of disease at diagnosis, with reported rates being 49% for local disease, 16% for regional disease, and 2% for distant disease .

2.4.3 Small cell lung cancer (SCLC):

Small cell lung cancer is a rare fast-growing lung cancer. It can affect anyone but it typically affects people who have a long history of smoking tobacco. About 10% to 15% of all lung cancers are SCLC. It is sometimes called **oat cell cancer.**

How does small cell lung cancer affect the body?

Small cell lung cancer starts when healthy cells in your lungs mutate or change into cancerous cells. These cells then divide and multiply uncontrollably. Eventually, the cancerous cells clump together in masses (tumors) in your lungs.

These tumors may shed cancer cells that your blood or lymph pick up and carry throughout your body. (Lymph is fluid that travels through your body to your lymph nodes.)

Symptoms:

Lung cancer typically doesn't cause signs and symptoms in its earliest stages. Signs and symptoms of lung cancer typically occur when the disease is advanced.

Signs and symptoms of lung cancer may include:

- A new cough that doesn't go away

- Coughing up blood, even a small amount
- Shortness of breath
- Chest pain
- Hoarseness
- Losing weight without trying
- Bone pain
- Headache

Risk factors:

A number of factors may increase your risk of lung cancer. Some risk factors can be controlled, for instance, by quitting smoking. And other factors can't be controlled, such as your family history.

Risk factors for lung cancer include:

- **Smoking.** Your risk of lung cancer increases with the number of cigarettes you smoke each day and the number of years you have smoked. Quitting at any age can significantly lower your risk of developing lung cancer.

- **Exposure to secondhand smoke.** Even if you don't smoke, your risk of lung cancer increases if you're exposed to secondhand smoke.

- **Previous radiation therapy.** If you've undergone radiation therapy to the chest for another type of cancer, you may have an increased risk of developing lung cancer.

- **Exposure to radon gas.** Radon is produced by the natural breakdown of uranium in soil, rock and water that eventually becomes part of the air you breathe. Unsafe levels of radon can accumulate in any building, including homes.

- **Exposure to asbestos and other carcinogens.** Workplace exposure to asbestos and other substances known to cause cancer — such as arsenic, chromium and nickel — can increase your risk of developing lung cancer, especially if you're a smoker.

- **Family history of lung cancer.** People with a parent, sibling or child with lung cancer have an increased risk of the disease.

2.4.4 Prevention of Lung cancer:

- **Don't smoke.** If you've never smoked, don't start. Talk to your children about not smoking so that they can understand how to avoid this major risk factor for lung

cancer. Begin conversations about the dangers of smoking with your children early so that they know how to react to peer pressure.

- **Stop smoking.** Stop smoking now. Quitting reduces your risk of lung cancer, even if you've smoked for years. Options include nicotine replacement products, medications and support groups.

- **Avoid secondhand smoke.** If you live or work with a smoker, urge him or her to quit. At the very least, ask him or her to smoke outside. Avoid areas where people smoke, such as bars and restaurants, and seek out smoke-free options.

- **Test your home for radon.** Have the radon levels in your home checked, especially if you live in an area where radon is known to be a problem. High radon levels can be remedied to make your home safer. For information on radon testing, contact your local department of public health or a local chapter of the American Lung Association.

- **Avoid carcinogens at work.** Take precautions to protect yourself from exposure to toxic chemicals at work. Follow your employer's precautions. For instance, if you're given a face mask for protection, always wear it. Ask your doctor what more you can do to protect yourself at work. Your risk of lung damage from workplace carcinogens increases if you smoke.

- **Eat a diet full of fruits and vegetables.** Choose a healthy diet with a variety of fruits and vegetables. Food sources of vitamins and nutrients are best. Avoid taking large doses of vitamins in pill form, as they may be harmful. For instance, researchers hoping to reduce the risk of lung cancer in heavy smokers gave them beta carotene supplements. Results showed the supplements actually increased the risk of cancer in smokers.

CHAPTER 3

3.1 DATA DESCRIPTION

The data has been collected from "Confluence mobile -Cancer imaging archieve wiki " website. The above mention data is lung cancer data. It consists 11 variables that is Patient id, Survival time, Status, Age, Clinical t stage, Clinical s stage, Clinical m stage, Overall stage, Histology, Gender, Stage. We have considered "death due to cancer" as the event of interest and this data comes under stage.

3.2 EXPLANATION OF THE STUDY VARIABLES

The data consists of 366 observations with 11 variables. The 11variables have been described as follows:

Patient id: Institution code

Survival time : Survival time in days

Status: censoring status (0 = alive, 1 = death)

Age: This variable gives the age (completed years) of the patient when they were diagnosed with cancer . It ranges from 34 to 92.

3.2.1 Clinical t stage

In clinical staging, the T stage for cancer refers to the size and extent of the primary tumor. The values for clinical T stage vary depending on the specific cancer type being considered, but in general:

T1: The tumor is small and confined to the organ where it originated.

T2: The tumor has grown larger than T1 but is still confined to the organ where it originated.

T3: The tumor has grown beyond the boundaries of the organ where it originated and may have invaded nearby tissues or structures.

T4: The tumor has grown extensively and may have invaded nearby organs or tissues.

It's important to note that clinical staging is typically based on imaging tests and physical exams, and may not be as accurate as pathologic staging, which is based on examination of tissue samples obtained during biopsy or surgery.

3.2.2 Clinical s stage

The clinical stage of cancer is usually expressed using a combination of numbers and letters, with the numbers 0 through 4 indicating the extent of the cancer's spread. The stage of the cancer is based on various factors such as the size and location of the tumor, whether it has spread to nearby lymph nodes or other parts of the body, and other factors such as the tumor's grade (how abnormal the cells appear under a microscope) and the patient's overall health.

The stage numbers are typically assigned as follows:

Stage 0: This stage is used for non-invasive cancers, which have not spread beyond the layer of cells where they first developed.

 Stage 1: This stage is used for small, localized tumors that have not spread to nearby lymph nodes or other parts of the body.

 Stage 2: This stage is used for larger tumors that may have spread to nearby lymph nodes, but have not spread to other parts of the body.

Stage 3: This stage is used for tumors that have spread to nearby lymph nodes and/or other parts of the body, but are still considered to be treatable.

Stage 4: This stage is used for tumors that have spread to other organs or distant parts of the body, and are often considered to be more difficult to treat.

It's important to note that the staging system can vary depending on the type of cancer, and some types of cancer may have additional stages or different criteria for assigning stages. Additionally, the clinical stage may be updated based on additional testing or changes in the cancer's progression over time.

3.2.3 Clinical m stage

The clinical M stage can be assigned a value of 0, which indicates that there is no evidence of metastasis, or the spread of cancer to other parts of the body beyond the primary tumor site.

The clinical M stage is one component of the TNM staging system, which is used to classify the extent of cancer based on the characteristics of the tumor (T stage), whether it has spread to nearby lymph nodes (N stage), and whether it has spread to other parts of the body (M stage).

3.2.4 Overall stage

The overall stage of cancer refers to a summary of the cancer's extent based on the TNM system, as well as other factors such as the cancer's grade and the patient's overall health. The overall stage is typically expressed using Roman numerals and/or letters, with the exact stage grouping depending on the type of cancer.

However, the stage groupings you have mentioned - I, II, IIIA, and IIIB - are commonly used for several types of cancer, particularly solid tumors such as breast, lung, and colorectal cancer. These groupings are based on the following criteria:

Stage I: The cancer is small and has not spread to nearby lymph nodes or other parts of the body.

Stage II: The cancer may be larger and/or have spread to nearby lymph nodes, but has not spread to distant parts of the body.

Stage IIIA: The cancer has spread to nearby lymph nodes, and may have grown into nearby tissues, but has not spread to distant parts of the body.

Stage IIIB: The cancer has spread to nearby tissues and lymph nodes, and may have also spread to distant parts of the body.

It's important to note that the exact criteria for each stage grouping can vary depending on the type of cancer, and some types of cancer may have additional or different stage groupings. Additionally, the overall stage may be updated based on additional testing or changes in the cancer's progression over time.

3.2.5 Histology

This refers to the type of cancer cells that are present i the lung tissue. There are several different types of NSCLC, including adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. The type of cancer can affect treatment decisions and outcomes.

Gender

This refers to the biological sex of the patient. Some studies have suggested that gender may play a role in lung cancer risk and treatment outcomes.

Stage

Staging refers to the extent or severity of cancer in the body, based on the size of the tumor and whether it has spread to nearby lymph nodes or other organs. Staging is usually done using the TNM system, which stands for Tumor size and extent, lymph Node involvement, and Metastasis (spread to other parts of the body).

Cleaning of Data

Out of 422 observation,56 observations have missing values. Estimating the missing values is tedious and inclusion of them leads to wrong conclusion. so, we decided to omit the missing observations. After the deleting the missing observation we are left with 366 values.

CHAPTER 4

4.1 Kaplan-Meier Plot

The KM plot is described as follows:

- The x-axis is time, from 0 to the last observed time point.
- The y-axis is the proportion of subjects surviving. At time 0 all the subjects are alive without an event.
- A vertical drop indicates an event.

Table 1: KM Estimate

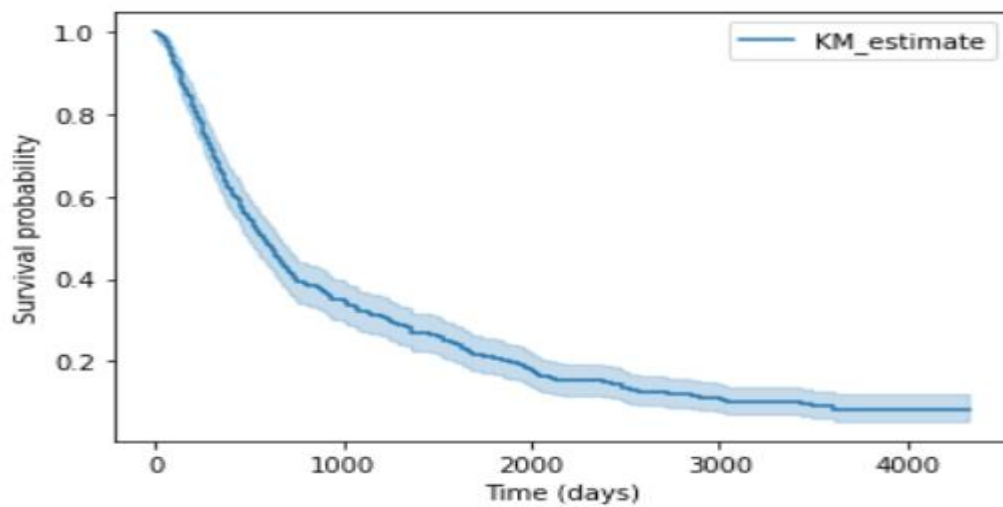| TIME LINE	
0.0	1.000000
10.0	0.997268
18.0	0.994536
25.0	0.991803
33.0	0.989071
...	....
4019.0	0.083247
4118.0	0.083247
4202.0	0.083247
4208.0	0.083247
4328.0	0.083247

Figure 1: Kaplan-Meier estimation

Finding the number of days a person was alive before the died

<lifelines.KaplanMeierFitter: "KM_estimate", fitted with 366 total observations, **44 right-censored observations**>

Median Survival Period (in days)

This is the time beyond which 50% of the individuals in the population under study are expected two survive, and is given by $S\{t(50)\} = 0.5$.

The median survival time: 632.0 days.

From the value of the median survival days, we can say, that after the diagnosis of lung cancer, a patient, on an average is alive for **632 days**.

Table 2:KM Estimate for confidence interval

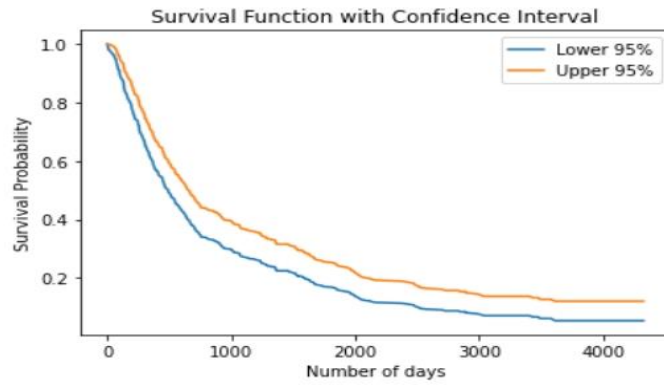|	KM_estimate_lower_0.95	KM_estimate_upper_0.95
0.0	1.000000	1.000000
10.0	0.980764	0.999615
18.0	0.978329	0.998631
25.0	0.974804	0.997349
33.0	0.971145	0.995884
...	...	...
4019.0	0.054009	0.120343
4118.0	0.054009	0.120343
4202.0	0.054009	0.120343
4208.0	0.054009	0.120343
4328.0	0.054009	0.120343

Figure 2: Survival function with confidence intervals.

SURVIVAL PROBABILITY WITH CUMULATIVE DENSITY

Table 3: KM estimate

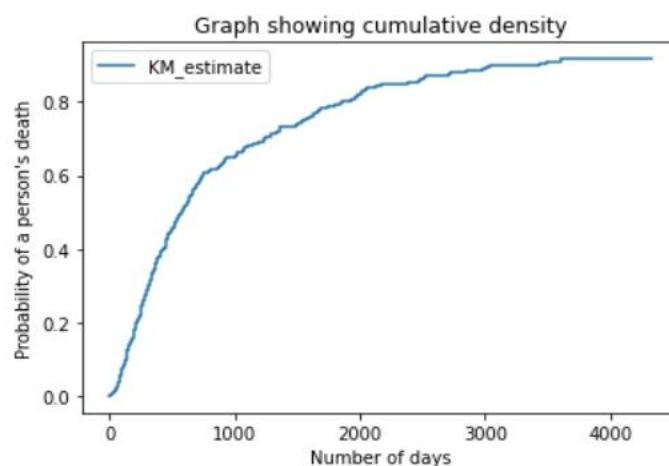| Timeline	Estimated S(t)
0.0	0.000000
10.0	0.002732
18.0	0.005464
25.0	0.008197
33.0	0.010929
...	...
4019.0	0.916753
4118.0	0.916753
4202.0	0.916753
4208.0	0.916753
4328.0	0.916753

Figure 3: Cumulative density

As the number of survival days increase, the probability of a person dying decreases.

The median time to an event:

The median survival time can help us get the amount of time remaining to an event.

Table 4: KM_estimate - Conditional median duration remaining to event

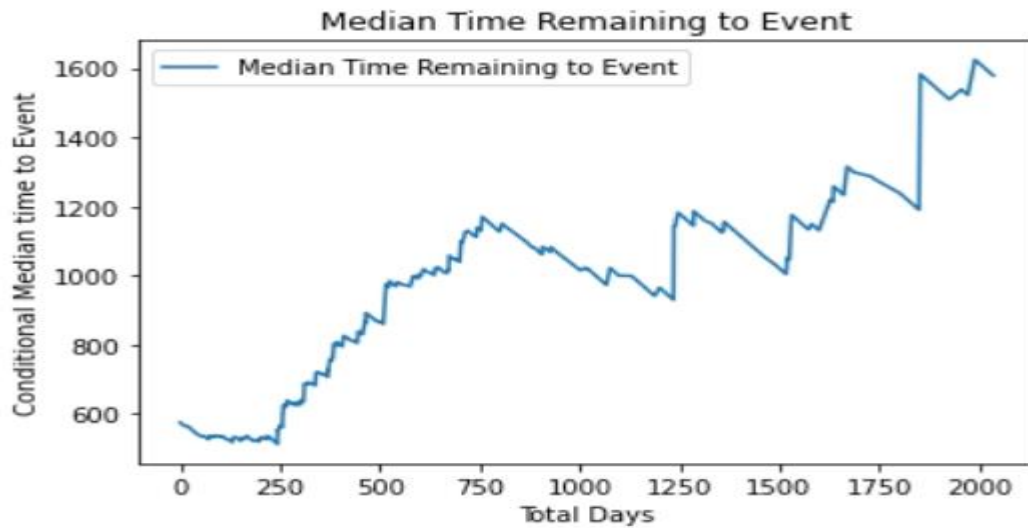| Timeline	Median Survival Time
0.0	575.0
10.0	565.0
18.0	565.0
25.0	558.0
33.0	550.0
...	...
4019.0	Inf
4118.0	Inf
4202.0	Inf
4208.0	Inf
4328.0	Inf

Figure 4: Median time remaining for an survival.

To check if sex of a subject has any affect on the survival
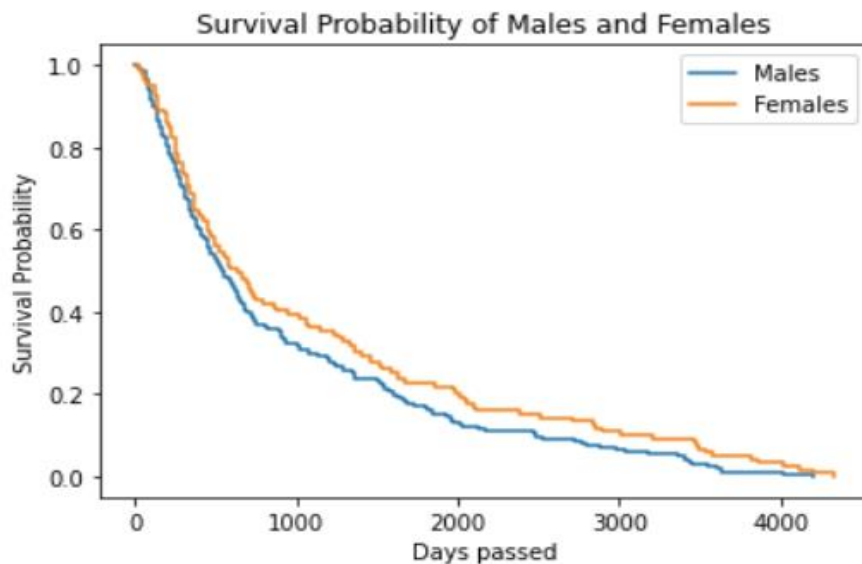

Figure 5: Survival probability of sex.

From the graph it can be inferred that females have comparatively higher chances of survival compare to males.

Conclusion

From this analysis we inferred that females are highly survived compared to males. We estimated survival probability of each observation and also median survival period is 613 days with 44 right censored.

4.2 COX REGRESSION

MULTIVARIATE COX REGRESSION

A Cox regression of time to death on the time-constant covariates is specified as follow:

Table 5: Time to death on the time-constant covariates.

	Coef	exp(coef)	se(coef)	z	Pr(>|z|)
Age	0.016148	1.016280	0.006693	2.413	0.0158 *
Gender	-0.146205	0.863981	0.127503	-1.147	0.2515
Clinical.t.stage	0.096973	1.101831	0.0638	1.519	0.1289
clincical.s.stage	0.154500	1.167075	0.066197	2.334	0.0196 *
clinical.m.stage	0.086931	1.090821	0.169908	0.512	0.6089
Stage	-0.209222	0.811216	0.108934	-1.921	0.0548

	exp(coef)	exp(-coef)	lower .95	upper .95
Age	1.0163	0.9840	1.0030	1.030
Gender	0.8640	1.1574	0.6729	1.109
Clinical.t.stage	1.1018	0.907	0.9722	1.249
clincical.s.stage	1.1671	0.8568	1.0251	1.329
clinical.m.stage	1.0908	0.9167	0.7819	1.522
Stage	0.8112	1.2327	0.6553	1.004

Concordance= 0.562 (se = 0.017)

Likelihood ratio test   =14.88 on 6 df, p=0.02

Wald   = 14.96 on 6 df, p=0.02

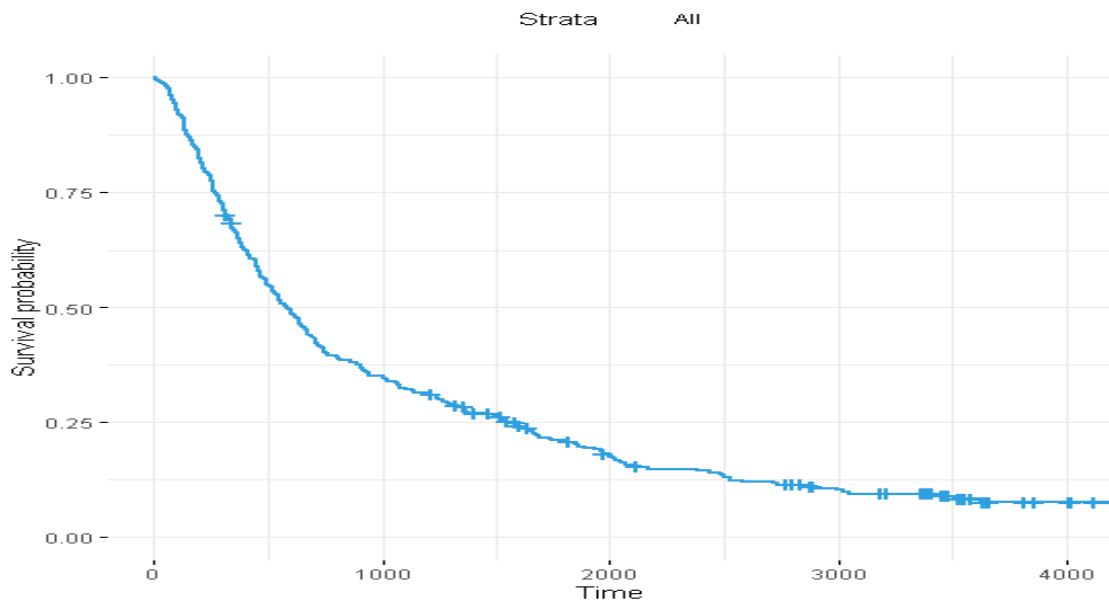**Score (log rank) test** = 15.01 on 6 df, p=0.02

Figure 6: Predicted survival proportion at any given point in time for a particular group.

Likelihood ratio test=8.63 on 3 df,

 p=0.03469

n= 366**,**

 number of events= 322

Conclusion

The p-value for all three overall tests (likelihood, Wald, and score) are significant, indicating that the model is significant. These tests evaluate the omnibus null hypothesis that all of the betas (β) are 0. In the above example, the test statistics are in close agreement, and the omnibus null hypothesis is soundly rejected.In the multivariate Cox analysis, the covariates age, clinical.s.stage and stage remain significant ($p < 0.05$). However, the covariate gender fails to be significant ($p = 0.25$, which is greater than 0.05).

The p-value for age is 0.0158, with a hazard ratio HR = exp(coef) = 1.0162, indicating a strong relationship between the patients' age and increased risk of death. The hazard ratios of covariates are interpretable as multiplicative effects on the hazard.

Similarly, the p-value for clincical.s.stage is 0.0196, with a hazard ratio HR = 1.167, indicating a strong relationship between the clinical.s.stage value and increased risk of death. Holding the other covariates constant, a higher value of clinical.s.stage is associated with a good survival.

The p-value for stage is 0.0548, with a hazard ratio HR = 0.811216 indicating a strong relationship between the stage value and decreased risk of death. Holding the other covariates constant, a higher value of stage is associated with a poor survival.

By contrast, the p-value for gender is now p=0.2515. The hazard ratio HR = exp(coef) = 0.8639, with a 95% confidence interval of 0.67 to 1.109. Because the confidence interval for HR includes 1, these results indicate that age makes a smaller contribution to the difference in the HR after adjusting for the clinical.s.stage values and patient's age, and only trend toward significance.

4.2.1 Residual:

It important to check the goodness-of-fit of a model, and potential outliers and influential observations to a fitted model. We are going check the normality of proportional hazard assumption.

Schoenfeld residuals

We are going to analyse the PH assumption using Schoenfeld residuals

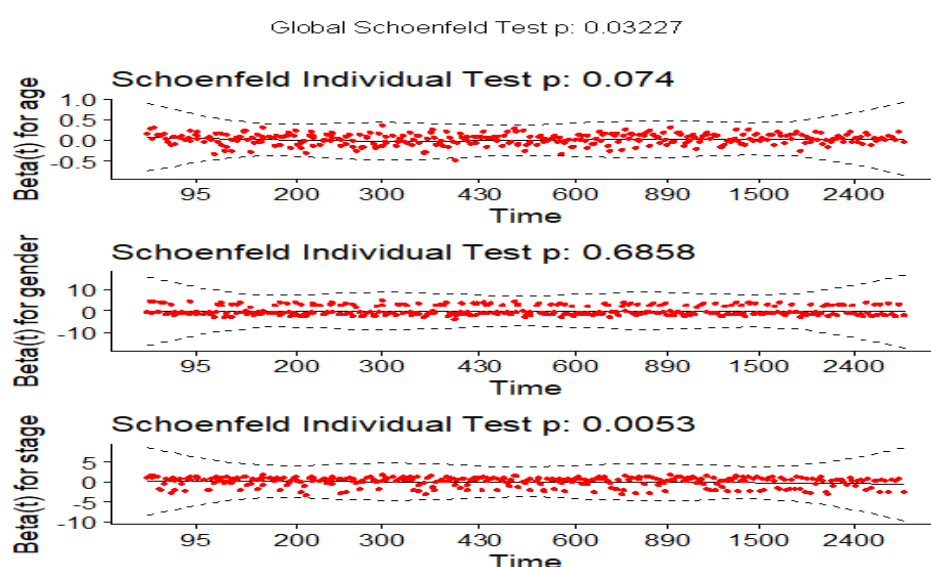|	Chisq	Df	P
Age	3.192	1	0.0740
Gender	0.164	1	0.6858
Stage	7.758	1	0.0053
GLOBAL	8.786	3	0.0323

Figure 7: Proportional hazard assumption Model

Deviance Residuals

It is possible to check outliers by visualizing the deviance residuals.
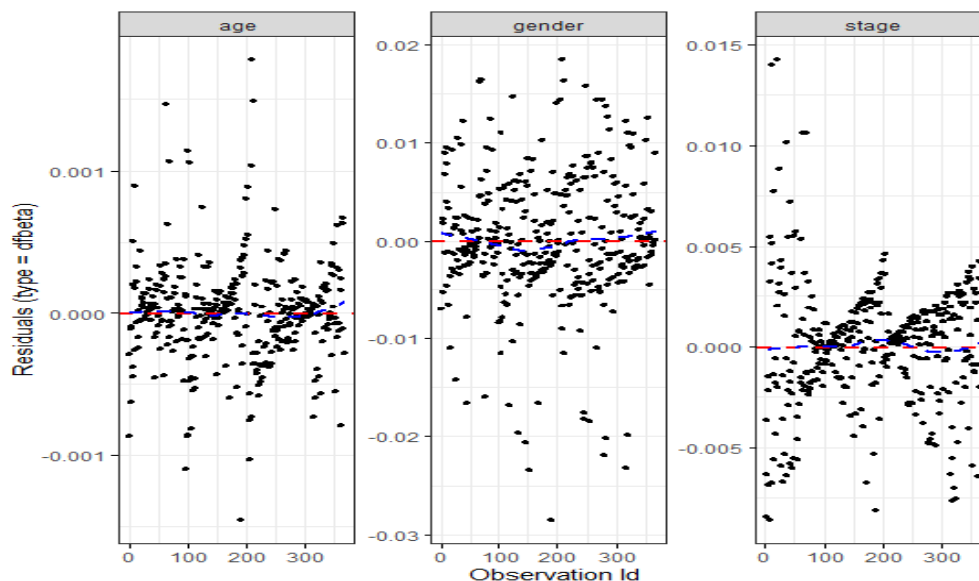

Figure 8: Deviance Residuals

Martingale residuals

Plotting the martingale residuals against continuous covariates is a common approach used to detect non-linearity.
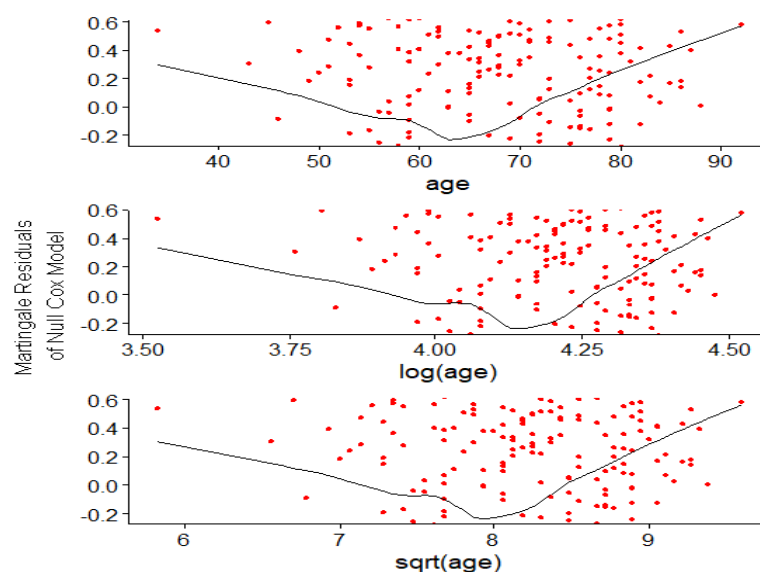

Figure 9: Martingale Residuals

Conclusion

From the output above, the test is statistically significant for each of the covariates, and the global test is also statistically significant. Therefore, we can assume there is no proportional hazards. From these comparing of three residuals, we interpreted that the Deviance residual is a good fit for this model.

4.3 Analysis of Logistic Regression

We are using the binary logistic regression to find the apparent error rate of misclassification with the help of SPSS.

Table 6: Case Processing Summary

Unweighted Cases		N	Percent
Selected Cases	**Included in Analysis**	365	99.7
	Missing Cases	1	.3
	Total	366	100.0
Unselected Cases		0	.0
Total		366	100.0

Table 7: Dependent Variable Encoding

Original Value	Internal Value
Alive	0
Death	1

Table 8: Categorical Variables Codings

		Frequency	Parameter coding		
			(1)	**(2)**	**(3)**
overall stage	**I**	62	1.000	.000	.000
	II	35	.000	1.000	.000
	IIIa	107	.000	.000	1.000
	IIIb	161	.000	.000	.000

Block 0: Beginning Block

Table 9: Classification Table

			Predicted		
			status		Percentage
	Observed		alive	death	Correct
Step 0	status	alive	0	44	.0
		death	0	321	100.0
	Overall Percentage				87.9

Variables in the Equation							
		B	S.E.	Wald	Df	Sig.	Exp(B)
Step 0	Constan	1.987	.161	152.817	1	.000	7.295

Variables not in the Equation[a]					
			Score	Df	Sig.
Step 0	Variables	survival time	152.407	1	.000
		Age	.420	1	.517
		clinical t stage	.820	1	.365
		clincical s stage	.374	1	.541
		clinical m stage	.554	1	.457
		overall stage	7.189	3	.066
		overall stage(1)	5.492	1	.019
		overall stage(2)	.014	1	.905
		overall stage(3)	.101	1	.751
		Gender	2.694	1	.101
		Stage	5.715	1	.017

Block 1: Method = Enter

Table 10: Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	135.598	9	.000
	Block	135.598	9	.000
	Model	135.598	9	.000

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	133.053[a]	.310	.596
Step	Chi-square	Df	Sig.
1	14.200	8	.077

Table 11: Classification table

			Predicted		
			status		
	Observed		alive	death	Percentage Correct
Step 1	status	alive	29	15	65.9
		death	6	315	98.1
	Overall Percentage				94.2

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	survival time	-.002	.000	62.202	1	.000	.998
	Age	.037	.457	.007	1	.935	1.038
	clinical t stage	.575	.314	3.349	1	.067	1.776
	clincical s stage	.218	.289	.572	1	.449	1.244
	clinical m stage	6.705	5379.533	.000	1	.999	816.614
	overall stage			6.648	3	.084	
	overall stage(1)	3.181	1.290	6.082	1	.014	24.061
	overall stage(2)	1.014	.954	1.128	1	.288	2.756
	overall stage(3)	1.213	.687	3.118	1	.077	3.364
	Gender	-.304	.499	.371	1	.543	.738
	Constant	2.712	2.505	1.172	1	.279	15.052

Conclusion

From these results you can see that stage ($p = .0014$) and stage1(p=0.014) added significantly to the model/prediction, but gender ($p = .543$) ,clinical t stage (p=.067) clinical s stage(p=.449),clinical m stage(p=.999) and age(p=.935) did not add significantly to the model. The logistic regression model was statistically significant, $\chi2(4) = 135.598$, p < 0.05. The explained variation in the dependent variable based on our model ranges from 31.0% to 59.0%, depending on whether you reference the Cox

& Snell R^2 or Nagelkerke R^2 methods, respectively. Nagelkerke R^2 is a modification of Cox & Snell R^2, the latter of which cannot achieve a value of 1. We obtain the confusion Matrix as:

	Alive	Death
Alive	29	15
Death	6	315

Apparent error rate $= \dfrac{n_1 M + n_2 M}{n_1 + n_2} = = \dfrac{6+15}{(29+15)+(6+315)} = 0.0301$

3% of our data is misclassified.

4.4 CHART REPRESENTATION OF STAGE
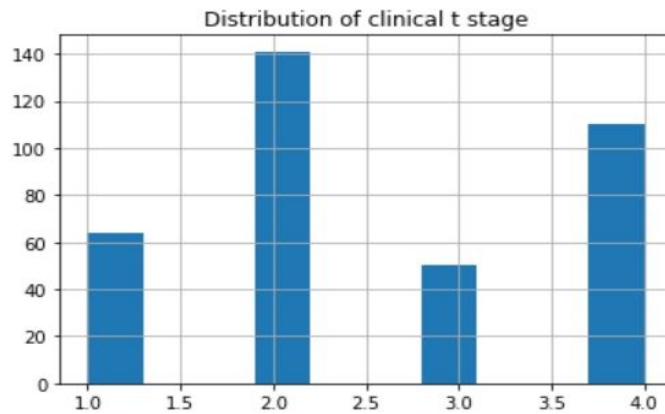
We are represented the stages in bar chart using Python.
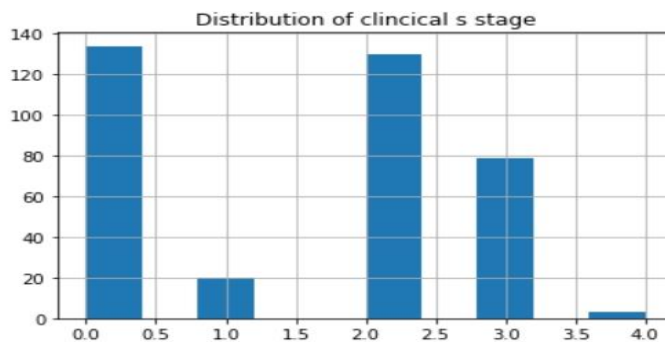

Figure 10: Clinical t stage

Figure 11: Clinical s stage

Conclusion

From the distribution of clinical.s.stage and clinical.t.stage We concluded that the t2 stage is higher compared to other stage whereas the s4 stage is less compared to s1 stage.

CHAPTER 5

CONCLUSION

5.1 SUMMARY OF THE STUDY

There are totally 443 cancer patients whose patient information is known and 44 patients are right censored due to many reasons. On an average, 50% of the patients are survived. we make the following inferences from the study.

Based on the Kaplan-Meier estimate, we observe that the survival probability is high for female Compared to male. In the clinical event of study, the survival probability is more for S stage and less for metastasis.

From the cox regression analysis, we interpretated that the variables reveals that the significant which are age, clinical S stage and Overall stage whereas the Gender has fails to be significant (<0.05). In the testing of hazard ratio has been higher for the variables Age and overall stage and also increased risk of death. In these analysis, clinical S stage has lesser risk of death compared to other variables.

In residual concept, we make the interpretation for the schoenfeld residual, with respect to the overall stage, each stage has significant difference in it. There is no proportional hazard assumptions.

From the logistic regression approach, we see that the variable stage and stage I has significant difference in this model and also 87.9% of the variance in lung cancer has been correctly classified .

REFERENCE

1. Allison PD (1995) Survival Analysis Using SAS: A Practical Guide. SAS Institute, Cary.

2. Bakr, (2017). Data for NSCLC Radiogenomics Collection.(Napel, Sandy).

3. Cox, D.R. (1972) Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B (Methodological), 34, 187-202. doi: 10.1111/j.2517-6161.1972.tb00899.

4. Grambsch, P.M. and Therneau, T.M. (1994) "Proportional hazards tests and diagnostics based on weighted residuals." Biometrika 81: 515-526. doi: 10.1093/biomet/81.3.515.

5. Kleinbaum, D.G. and Klein, M. (2012) "Logistic Regression for Grouped Survival Data: Linking the Past to the Present." Journal of Epidemiology and Biostatistics, 17:1, 1-10. doi: 10.1080/13595220310001619355

6. Korn EL, Graubard BI (1999) Analysis of Health Surveys. John Wiley and Sons, New York.

7. Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481. doi:10.2307/2281868

8. Lin, D.Y. and Wei, L.J. (1989) "The robust inference for the proportional hazards model." Journal of the American Statistical Association 84: 1074-1078. doi: 10.1080/01621459.1989.10478874.

9. Machin D, Campbell MJ, Tan SB, Tan SH (2018) Sample Size Tables for Clinical Studies, 3rd edn. Wiley-Blackwell, Hoboken.

10. Pocock SJ, Clayton TC, Altman DG (2002) Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. Lancet 359:1686–1689. doi:10.1016/S0140-6736(02)08594-X

ssss11. Schoenfeld, David (1982) "Partial residuals for the proportional hazards regression model." Biometrika 69: 239-241. doi: 10.1093/biomet/69.1.239.