**T.R.N.C**

**NEAR EAST UNIVERSITY**

**INSTITUTE OF HEALTH SCIENCES**

# COMPARISON AND UTILIZATION OF UNIVARIATE AND MULTIVARIATE STATISTICAL MODELS ON NON-SMALL CELL LUNG CANCER (NSCLC)

**OGUNJESA BABATOPE AYOKUNLE**

**Master of Science in Biostatistics**

**Advisor:**

**Asst. Prof. Dr. Özgür Tosun**

**NICOSIA, 2018**

**T.R.N.C**

**NEAR EAST UNIVERSITY**

**INSTITUTE OF HEALTH SCIENCES**

# COMPARISON AND UTILIZATION OF UNIVARIATE AND MULTIVARIATE STATISTICAL MODELS ON NON-SMALL CELL LUNG CANCER (NSCLC)

**OGUNJESA BABATOPE AYOKUNLE**

**Master of Science in Biostatistics**

**Advisor:**

**Asst. Prof. Dr. Özgür Tosun**

**NICOSIA, 2018**

**APPROVAL**

Thesis submitted to the Institute of Health Sciences of Near East University in partial fulfillment of the requirement for the degree of Master of Science in Biostatistics.

**Thesis Committee;**

| | |
|---|---|
| Chair of the committee: | Prof. Dr. S. Yavuz Sanisoğlu<br>Yıldırım Beyazıt Üniversitesi<br>Sig: ..... |
| Advisor: | Asst. Prof. Dr. Özgür Tosun<br>Near East University<br>Sig: ..... |
| Member: | Assoc. Prof. Dr. İlker Etikan<br>Near East University<br>Sig: ..... |
| Approved by: | Prof. Dr. K. Hüsnü Can Başer<br>Director of Health Science Institute<br>Near East University<br>Sig: ..... |

## DEDICATION

This research work is dedicated to my Beloved Parents, Mr & Mrs G.A Ogunjesa and my entire family members whose immeasurable supports and encouragement made it possible for me to gain the access to this higher education learning .

## ACKNOWLEDGMENT

I give thanks to the Lord Jesus Christ for the grace, ability, provisions, good health, guidance and guardian, for the successful completion of this thesis and the education program at large.To Him be all the glory, honour and adoration forever (Amen).

I equally extend my appreciation to the members of my family members; Mr and Chief (Mrs). G.A Ogunjesa, Mr and Mrs Tunde Abanishe , Mr and Mrs Gbenga Ogunjesa, Mr and Mrs Muyiwa Ogunjesa and Mr and Mrs Femi Ogunjesa for their great financial and moral support towards my education program as well as my well being in this country.

I especially want to appreciate the support, the time and the valuable contributions of my mentor, my lecturer and my advisor; Asst. Prof. Dr. Özgür Tosun who played a very critical role in imbibing the knowledge and understanding of Biostatistical concepts in me as well as meticulously guiding me in this research. My profound appreciation also goes to my Head of Department as well as the Vice Director of the Near East Health Institute; Prof. Dr. İlker Etikan who played an important role in guiding me in articles writings. I equally express my gratitude to Prof. Dr. S. Yavuz Sanisoğlu whose contributions in my thesis defense offered me more insightful understanding in Biostatistical research.

My sincere appreciation also goes to Pastor Sebestian Nlebedim and the Watchman Catholic Charistmatic Renewal Movement (WCCRM), North Cyprus family for the fellowship and togetherness we all shared together. I extend my immense gratitudes to the likes of Bro. Timilehin, Bro. Daniel, Bro. Jeremiah, Bro.Gabriel, Bro. Samuel, Bro. Dotun , Sis. Gift, Sis. Precious among others.

Likewise, I appreciate my coursemates such as Meliz, Devrim, and Kabiru for the moments we shared together.Also, to my library staff buddies ; Fatma, Eda, Aysur, Serap, Onur, Sahin, Pamela,Yakurp, Nasiru, Kubra, Imad, Muhammed, Abid, Mr Lisani and the host of others.

## ABSTRACT

### COMPARISON AND UTILIZATION OF UNIVARIATE AND MULTIVARIATE STATISTICAL MODELS 0N NON-SMALL CELL LUNG CANCER (NSCLC)

Ogunjesa, Babatope Ayokunle

Department of Biostatistics

Thesis Supervisor: Asst. Prof. Dr. Özgür Tosun

June, 2018

This study examined the application of Univariate, Bivariate, and Multivariate analysis for an insightful decision making process. The study makes use of a secondary data consisting of 548 patients suffering from a stage III Non- Small Cell Lung Cancer (NSCLC) from Cancer data repository. Fourteen (14) attributes made up of 6 quantitative and 8 qualitative variables ranging from clinical, laboratory and socio-demographic measures such as Age (yrs), Body Mass Index (BMI), N-Staging, World Health Oganisation (WHO) performance status and so on were considered in the study. The Univariate analysis was conducted on the obtained data using statistic such as mean, median, percentages and so on to describe the pattern and distribution of the variables. The Bivariate analysis involved the use of t-test, Mann Whitnney test as well as the Chi-Square to test for significance as regards to the patients' status of being dead or alive.

The Simple Logistic Regression Model (SLRM) was used to examine the patients' risk of death for each of the variables. It was found that the respective SLRM of the Age (yrs), Equivalent Radiation dose in 2-Gy fraction ($Eqd_2$) and the WHO performance status and the Treatment Method variables were respectively significant at a significance level of 0.05. However, all the

SLRM with a p-value of < 0.200 were then used to compute a final Multiple Logistic Regression Model (MLRM). The MLRM was significant, $\chi^2(15)$ = 54.00, $p < 0.001$.The model explained the 18.50% ( Nagelkerke $R^2$) of the variance in deaths of patients and 80.70% cases were correctly classified.

Patients with no chemotherapy treatment are 10.989 times at risk of dying compared to the patients subjected to a concurrent treatment plan. The Area under the Curve (AUC) of the Receiver Operating Characteristic curve for the MLRM of 75.30% provides a better analysis outcome than the ROC of the SLRM of the individual quantitative variables whose highest AUC value is 65.20% indicating that MLRM provides a better analysis result than Univariate analysis.

**Keywords:** Univariate, Bivariate, Multivariate analysis, Multiple Logistic Regression Model, Simple Logistic Regression Model, Non- Small Cell Lung , Cancer.

## TABLE OF CONTENTS

**LIST OF TABLES**

## LIST OF FIGURES

## LIST OF ABBREVIATIONS

| S/No: | ABBREVIATIONS | EXPLANATION |
|---|---|---|
| 1 | BMI | Body Mass Index |
| 2 | WHO | World Health Organization |
| 3 | $FEV_1$ | Forced Expiratory Volume |
| 4 | $EDQ_2$ | Equivalent Radiation dose in 2-Gy fraction |
| 5 | GTV | Gross Tumor Volume |
| 6 | NSCLC | Non- Small Cell Lung Cancer |
| 7 | SCC | Squamous Cell Carcinoma |
| 8 | ROC | Receiver Operating Characteristics |

# CHAPTER ONE

## 1.1 Introduction

In a complex world as ours, Statistics have been considered a precursor to an effective decision making process following the insight it thus provides (Pullinger, 2013). Its ability to model most scientific and non-scientific problems into solvable and actionable ways have been generally accepted in almost all fields of human endeavors. Equipped with various mathematical concepts, iterative processes and statistical based computations, solutions have been provided to various society oriented problems.

There are two broad divisions of statistics, namely descriptive and inferential statistics (Zheng et al., 2016). While the former gives the distribution and various patterns of data under observation, the latter gives a statistical evidence based solutions to research questions and hypothesis testings by making use of various Univariate, Bivariate and Multivariate Analysis approach (Khademi, 2016). The Univariate analysis considers a single variable examination mostly in descriptive format; the bivariate analysis evaluates two single dependent or independent variables with statistical tools such as t-Test, Wilcoxon signed rank test and so on while the multivariate analysis is about evaluating relationships as well as making inferential decisions among more than two variables (Canova et al., 2017; Kenkel, 2006).

The multivariate analytical method emphasizes on the prediction of a single outcome from a variety of two or more independent variables. This approach results into a model built-up which makes use of signs and the magnitude of the co-efficient to define relational effect on the dependent or response variable. Therefore, new values of the dependent variable can be

predicted and also the type of effect each of the independent variables holding other measures fixed can be measured (Bagleya et al.,2001).

Common examples of multivariate analysis models are linear regression, proportional hazard regression and the logistic regression. The linear regression model is based on a response outcome which is numerical in nature while proportional hazard regression is centered on a time to the occurrence of an event of interest. The logistic regression has an outcome variable with two possible events (Hosmer & Lemeshow, 2000; Park, 2013). Such dichotomous events could be high or low, dead or alive, diseased not diseased etc.

Also called the logit model, the logistic regression method is increasingly popular in use today (Oommen et al., 2011). In the medical field, its usage is equally highly pronounced, especially when a decision is geared towards understanding the probable treatment effect as regards improvements as well as efficacy (Tetrault al., 2008). Therefore, the focus of this research work is to demonstrate the usage of logistic regression in analyzing clinical research data and how analysis outputs can be used for clinical advising. The research will be making use of a cancer data for exploration purposes.

One of the most ravaging non-communicable diseases that have heightened the level of the universe disease problem with a high fatality is the Cancer disease (Awodele et al., 2011; Binu et al., 2007). Cancer can be described as the untamed growth of abnormal cells and its consequent spread to other parts of the body system. It has continued to pose a great challenge to the various national and local health systems of numerous nations in the world, both developed, less developed and least developed nations inclusive. On records according to Stewart & Kleihues (2003), the annual diagnosis of incidence of cancer is about 10 million people

comprising of more than 100 types with their different corresponding prognosis. Cancer is often named relative to their site of occurrence in the body. Examples of site occurrence of cancer in the human body include the digestive system (e.g., colon cancer), respiratory system (e.g. Larynx cancer), bones and joints, breast (breast cancer), soft tissue part (e.g., heart cancer), skin (e.g.,melanoma cancer), genital system (e.g. Colon cancer), urinary system (e.g., urinary bladder) and host of other parts (Siegel et al., 2017). The Lung cancer, which is the focus of this research work occurs in the respiratory system of the body and its occurrence is due to the growth of abnormal cells, which are also out of control in nature in the lung area of the body.The Lung cancer has the leading mortality rate compared to all other cancer types and in the year 2012 alone, about 1.8 million people were diagnosed with lung cancer while about 1.6 million death due to lung cancer was recorded (Brambilla & Travis, 2014; Silverstri & Jett, 2010). The lung cancer cells are generally grouped into two types, namely Non-small cell lung cancer (NSCLC) and Small cell lung cancer (SCLC) (Oser et al., 2015; Zappa, C., & Mousa, 2016). The group classification of these cells is important as they aid in treatment decision techniques as well as prognosis monitoring and evaluation. In this study, emphasis will be placed on the non-small cell lung cancer (NSCLC) relative to various risk factors associated with patients.

## 1.2 Objectives of the Research

The objective of this research is to show how univariate, bivariate and multivariate statistical tools can be adopted in the analysis so as to be able to draw inferences from research data. These methods will then be explored in a cancer data for risk factors evaluations.

## 1.3 Significance of the Study

The research will help us to show how various statistical modeling tools can be utilized to generate insights from clinical data, which often serves as a baseline for decision making. Several data analysis approaches will be compared and a systematic review to construct better multivariate logistic regression models based on the evidence collected from bivariate statistical tests will be utilized. Health and other health allied researchers, health and wellness oriented bodies will be able to understand how various risk factors affect the prognosis of non- small lung cancer cell (NSLCC) and how this can influence treatment plans for patients.

## 1.4 Thesis Structure

The first chapter of the thesis starts with the background information about the study. Here, basic concepts and methodology as regards to the application of statistical tools to problem solving mechanism were explained. Also, an introductory note on cancer diseases from the viewpoint of clinical and epidemiological perspectives will be discussed.

The second chapter explains the Regression model and as well as the concepts of Logistic regression which is a binary outcome modeling statistical tool. The oncology overview of the lung cancer will be discussed likewise the reviews of previous researches conducted using Multivariate analysis.

The third chapter will seek to explain the research methodology behind the study, the statistic terms used in logistic regression model will be described and the definition of the variables under

the non- small lung cancer (NSLCC) analysis will be enumerated. The results of the analysis of the data used in the study will be presented in the fourth chapter.

While the conclusion and summary of the research findings with necessary recommendations for future research will be presented in the fifth chapter of the study.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1: Introduction

This section of the research study presents the underlying principles of the statistical methods employed, associated literature overviews previously done in terms of application of bivariate and multivariate analysis models. Also, discussion of the occurrence of NSLCC as wells as the risk factors under consideration will be articulated.

## 2.2: Statistics and Inference

Basically, Statistics make use of samples drawn from a population of study for inferential decision purposes. Population in Statistics can be defined as an all- inclusive aggregate of objects, units or items from which certain information is needed to be ascertained (Banerjee & Chaudhury, 2010). Population can be in terms of total number of women that had child delivery in a hospital in a year, or total number of cells deformity in the kidney and so on. It is often impossible to get hands on all units present in a population for analysis; hence a selection of some units which are representative of all items in the population is done. The selected units are called samples, and the process of selection is called sampling.

In a bid to make inference concerning a population of the selected samples, the hypothesis testing and the confidence interval approach are the two methods employed. The interval method specifies a range of value with a certain percentage of confidence from which result has a probability to be obtained from while the hypothesis test evaluates the extent to which a result

can be attributed to chance. The hypothesis testing makes use of a *p-value* measure. The p-value which takes values that lie between 0 and 1 statistic is a probability stemming from the condition that signifies no form of difference in getting an anticipated value result or more extreme than what was actually observed (Dahiru, 2008). The nearer the assumed value is to 0, the higher the conclusion that the observed difference is not due to chance while the closer the p-value is to 1, entails the observed difference can be attributed to chance.

**2.3: Classification of Hypothesis Testing Methods**

Hypothesis testing can be categorized into two segments namely the parametric and non-parametric tests. The parametric test refers to methods with the assumptions that the population from which samples are drawn from follow a particular distribution pattern while the non-parametric test are methods that do not follow any form of distribution pattern, but often based on ranking typical of ordinal scaled observations (Mircioiu & Atkinson, 2017 ; Sedgwick, 2012). In parametric tests, the mean and standard deviation are used as a symmetric measure for the shape distribution. The population of quantitative variables in parametric tests is considered to be normally distributed. In order to validate this assumption since it is often impossible to have opportunity to the population, conclusion on normality is made upon the drawn samples through a normality test. Kolmogorov Smirnov, Shappiro Wilk and Anderson-Darling tests for assessing the test of normality of observations. In a case of repeated measures from the same set of sample, the normality test is sufficient to conclude the use of parametric analytical tests. However, in case samples are drawn from separate populations, the assumptions that the variances are equal needed to be fulfilled before parametric methods can be put to use. Most statistical software are embedded with algorithms to test for the homogeneity of variances. However, if otherwise, the

non-parametric tests are used. Examples of parametric methods include Independent samples-t test, Analysis of Variance (ANOVA), Regression models and so on while non-parametric tests include tests such as Kruska-Wallis test, Wilcoxon Signed Rank tests and so on. Due to the outcomes in medical data, the usage of regression models are well pronounced in this field.

## 2.4 Regression Models

The analyzing of the relationships and interactions between a dependent variable with other independent variable(s) is done usually with regression modeling (Al-Ghamdi, 2001). The adoption of these methods is well pronounced in the medical, scientific researches due to their ability to measure the relationship among variables, make cases for effects of variables that are confounded and also make predictions for the outcomes under study. This makes it one of the most versatile multivariate methods used in studying the relationship and dependency among variables of observation. The dependent variable also called the outcome variable is often expressed as the product or the addition of the coefficient of the independent variables under consideration. This model basically makes it possible to estimate the value of the dependent variable as well as understand the type and extent of contribution impact the independent variables have upon the dependent variable.

This form of relationship could be linear, cubic, exponential or logistic in nature. The linear regression model centered on the least squares methodology is quite common in use. The regression is called a simple regression model if only a dependent and an independent variable is considered while a multiple regression model consist of a single dependent variable and more than one independent variable (Uyanık & Güler, 2013).

A simple linear regression model is given as:

$$Y(X) = \beta_0 + \beta_1 X + \varepsilon \qquad \textit{equ (1)}$$

Where Y(X) is the dependent variable

$\beta_0$= Intercept (Constant) of the regression equation

$\beta_1$= Coefficient of the independent Variable *X*

$\varepsilon$ = Error Term

A multiple linear regression model is given as:

$$Y(X) = \beta_0 + \beta_1 X + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n + \varepsilon \qquad \textit{equ (2)}$$

$\beta_0$ = Constant of the regression equation

$\beta_1, \beta_2, \beta_2 X_2 \ldots \beta_n$ are the coefficients of the individual independent variables.

In order to make use of the ordinary linear regression model, there are several assumptions needed to be fulfilled (Alexopoulos, 2010; Schmidt & Finan, 2017). Such assumptions are as follows:

(a). There should be a linear relationship between the dependent and the independent variables.

(b). The variance of the error term should be constant

(c). The error terms should be normally distributed.

(d). For every pair of the dependent and independent variables, the error terms should not be highly correlated.

However, the drawback for this ordinary linear regression model is that the dependent variable can only take on a numerical scale measurement (i.e.Continuous). In the medical field, there are cases whereby dependent variable of choice gives a binary or multiple outcomes.(e.g. The success or failure of a surgical operation procedure, the status of a patient in terms of recovered or not recovered from an ailment after a drug administration), hence, the use of the ordinary

linear regression model cannot be employed in this scenario. In the event that the dependent variable consists of two possible outcomes, it is considered as a dichotomous dependent variable and those more than two level outcomes are referred to as multinomial dependent variable.

Therefore, the assumptions that govern the use of the ordinary regression model cannot hold for this form of variables since these assumptions cannot hold for dependent variables that are categorical. Hence, a regression model called logistic regression or logit regression is used.

### 2.4.1 Logistic Regression Model

Logistic regression is an iteratively measure based methodology which maximizes the strong combination of variables resulting in a higher chance of predicting the outcome of interest (Stoltzfus, 2011). Unlike the previously discussed ordinary least square regression model, none of the assumptions are needed to be fulfilled in the usage of a logistic regression model.

The modeling of the relationship between a dependent variable with two or more possible outcomes and an independent variable or group of independent variables is done by a Logistic regression method. According to Chatterjee and Hadi (2006), in case of a dependent variable with more than two possible outcomes, and the interest lies in determining the chance of the occurrence of one of the possible ordered outcomes, this is regarded as an ordinal logistic regression (Chatterjee & Hadi, 2006). In logistic regression, the analysis of interest is to predict the probability of the event outcome of the response variable rather than the actual value of the

response variable Y. The event occurrence of the dependent variable is denoted as 1 and the its corresponding probability of occurrence can be stated as P (Y=1).

Crammer (2002) stated that the first users of this method dated back to the 19th century. And in our contemporary time, many researches that have been published have frequently adopted the usage of this method, especially when interest lies in outcomes predictions that are dichotomous in nature (King & Zeng, 2001). There are two major approaches to solving this problem. One is the least squares estimations using a transformation technique while the other involved maximum likelihood estimation using complex algorithm (Mendehall & Sincich, 2003).

### 2.4.2 Least Squares Estimations Using Transformation Technique

The clear cut concept of a logistic regression is the natural logarithm that is attributed to the odds of the dependent variable. The odds of the occurrence of an event, simply typifies the chance or probability of the interest of an outcome occurring or not.

Suppose the chance or success of the event of interest occurring is *"P",* therefore, the non-occurrence of the event is given as "*1-P*". This can be written as

$$\text{Odds} = \frac{P}{1-P} = \frac{probability\ of\ success(event\ \ occurence)}{Probability\ of\ failure(Non-event\ occurence)}$$

Recall that the simple linear regression is stated as

$$\boldsymbol{Y(X) = \beta_0 + \beta_1 X + \varepsilon}$$

Given that the observed value of Y is expressed as the mean of a sub-population of Y values ($\mu_{y|x}$) for a given value of X, the error term $\boldsymbol{\varepsilon}$ is the difference between the observed Y and the regression line is zero. And this can be written as:

$\mu_{y|x} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 X} + \boldsymbol{\beta_1 X_2} + \ldots + \boldsymbol{\beta_1 Xn}$

It can further be written as:

$E_{(y|x)} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 X} + \ldots + \boldsymbol{\beta_1 Xn}$

From the three stated equations above, the right sides of the equality sign can take any values from negative to positive infinity ($-\infty$ and $+\infty$).

Thus, the ordinary regression model is not fitting when the dependent variable Y is binary because the expected value of Y, E(Y) is the probability that Y=1 and, therefore, is limited to take values between 0 to 1 (Wayne, 2010). Other assumptions whereby the ordinary linear regression model is not fitting is the problem of the non-normal errors and unequal variances.

Concerning the non-normal errors assumption, this has been violated since the dependent variable y and, hence the random error $\boldsymbol{\varepsilon}$ can take on only two values.

$\sigma^2 = Y - (\boldsymbol{\beta_0} + \boldsymbol{\beta_1} X)$

Therefore, when Y =1, $\varepsilon = Y - (\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X)$ and when Y= 0, $\varepsilon = Y - \boldsymbol{\beta}_0 - \boldsymbol{\beta}_1 X$.

Given that the sample size n is large, any conclusion drawn from the least square equation is considered valid even the error term is not normally distributed.

Concerning the unequal variance assumption violation, it can be deduced that the variance $\sigma^2$ of the error term $\varepsilon$ is a function of the expected value of Y i.e. E (Y) which is the probability that the response Y equals 1.

Concisely,

$\sigma^2 = V(\varepsilon) = E(Y)[1 - E(Y)]$

Since for the ordinary least square regression,

$E(Y) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_1 + \boldsymbol{\beta}_2 X_2 + \ldots \boldsymbol{\beta}_n X_n$

This implies that $\sigma^2$ is not constant and it also even depends on the values of the explanatory variables; therefore, the standard least squares of homoscedasticity or equal variances is violated. Mathematically, the application of the logarithm function of the linear regression model equation is given as:

$$\text{Logit}(y) = \text{In}(\text{odds}) = \text{In}(\frac{p}{1-p}) = \alpha + \beta x \qquad \textit{equ (3)}$$

This is referred to as the logistic regression model because the transformation of $\mu_{y|x}$ to $\text{In}(\frac{p}{1-p})$ is called the logit transformation.

The ratio $\frac{p}{1-p} = \frac{P(y=1)}{P(y=0)}$ is known as the odds of the event , y=1 occurring and is usually called the log-odds model.

The logarithm function makes it possible for the $\alpha+\boldsymbol{\beta}x$ to take values between 0 and 1 which was linearly impossible with the ordinary regression model of the least squares. The probability of success is denoted as "*p*" which takes the value of 1 and the probability of failure denoted as "1-p" which takes the value of 0. The outcome variable of the logistic regression follows the Bernoulli distribution due to the value of 1 and 0 it takes.

The "$\boldsymbol{\alpha}$ " and " $\boldsymbol{\beta}$" are called parameters which explains the intercept as well as the independent co-efficient respectively.

The equation (3) relationship can also be expressed as follows:

$$\text{P} = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{1}{1+e^{-(\alpha+\beta x)}} \qquad \textit{equ (4)}$$

Given that the exponential function is the inverse of the natural logarithm.

### 2.4.3 Multiple Logistic Regression Model

Given that $x_1$, $x_2$…….$X_n$ are a collection of independent variables and y is a binomial –outcome variable with probability of success=p, then the multiple regression equation can be stated as :

$$\text{Logit (y)} = \text{In(odds)} = \text{In}\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_2 + \beta_2 x_2 + \ldots\ldots + \beta_n x_n \qquad \textit{equ (5)}$$

y = 1 (if outcome is success) and 0 (if outcome is failure)

*Where α is the intercept and the βs' are the coefficients of the predictor variables*

From the above equation, the predicted probable outcome of the dependent variable with multiple independent variables can also be written as follows:

$$\mathbf{P} = \frac{e^{\alpha+\beta 1X1+\beta 2X2+\cdots+\beta nXn}}{1+e^{\alpha+\beta 1X1+\beta 2X2+\cdots+\beta nXn}} = \frac{1}{1+e^{-(\alpha+\beta 1X1+\beta 2X2+\cdots+\beta nXn)}} \qquad \textit{equ (6)}$$

### 2.4.4 Estimation of Odds ratios in Multiple Logistic Regression

Suppose there is a dichotomous independent variable (xj) which is coded as 1 if present and 0 if absent, thus, the odds ratio relating this independent variable to the dependent variable is estimated by :

$$\text{OR}=e^{\beta j}$$

This relationship expresses the odds in favour of success if x j=1 divided by the odds in favour of success if xj = 0 after controlling for all other variables in the logistic regression model.

### 2.4.5 The Logistic Regression Curve

As a result of the binary nature of the dependent variable, the ordinary least square method is not fitting to model the outcome hence; the mean of the dependent variable outcome is computed for

each category following the categorization of the independent variables. The plot of this relationship will form an "S" shaped plot otherwise called a sigmoidal curve, which appears a little bit linear in the middle, but gets flattened at the base and upper end of the curve (Wojciech , 2017). The logarithm function on the dependent outcome of the logistic regression model makes it possible for the model to overcome the problem of linearity as well as the normality and homoscedasticity of the error terms which are needed for the least square regression model. Also, it makes possible to make the prediction of the odds of the dependent outcome from the independent variable.

The simple logistic function can be stated as follows:

$$y = \frac{e^x}{1 + e^x}$$

Further expansion of the formula above leads to the derivative given below

$$y = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}} = \frac{1}{1 + e^{-(\alpha+\beta x)}}$$

The duality outcome of this model in relation to a single independent continuous result into a plot clearly different from the usual classical regression equation line. In a logistic model, such plot leads to two parallel line formations relative to the dependent variable outcomes. This is shown in the figure 1 diagram below:

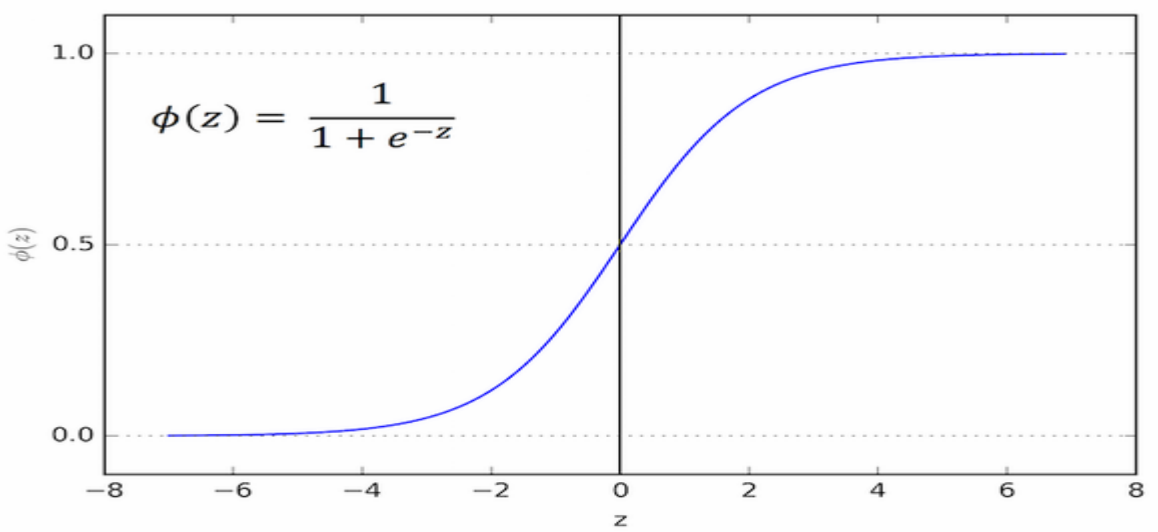


**Figure 1: Logistic Regression Curve**

### 2.4.6 Maximum Likelihood Estimation

The $\boldsymbol{\beta}$s' coefficients in the logistic regression can as well be obtained using the concept of maximum likelihood. Very similar property of the least squares estimators and the maximum likelihood estimation is that when the error terms of a linear regression model are normally distributed, both the estimates of the least squares as well as that of the maximum likelihood estimate are the same.

Given that x is a vector and Y is a Bernoulli distributed variable with 1 and 0 as the possible outcome, X is a linear function and the assumption that the probability Y = 1 is considered as a non-linear function of X.

The logistic regression model is stated as:

$$P(Y=1|x;\alpha,\boldsymbol{\beta}) = \sigma(\alpha+\sum_{j=1}^{d} Bjxj) = \frac{1}{1+exp-[\alpha+\sum_{j=1}^{d}\beta jXj]} \qquad equ\ (7)$$

### 2.4.7 Model Fit Statistics in Logistic Regression

**Log-likelihood**

Given that $L_o$ represents the log-likelihood of the logistic model with the constant term and $L_1$ denotes the log-likelihood of the model with the independent variables and the constant term. According to Menard (2001), the -2 log-likelihood (-2LL) for a dichotomous logistic regression which is the deviance for the is stated as:

$$L_o=\{(n_{y=1})In[P(Y=1)+(n_{y=0})In[P(Y=0)\}$$

-2LL can then be stated as -2 ($L_0$). The total number of events is noted as N while the total number of cases where Y=1 is denoted as $n_{y=1.}$ When the difference between $L_o$ and $L_1$ is multiplied by -2, the *-2LL* is chi-squared function and interpreted as:

$$\chi^2 = -2\ (L_o - L_1)$$

The $\chi^2$ function is used to test the hypothesis of the logistic regression model. When the model is significant, the alternative hypothesis is accepted and the null hypothesis that the coefficients of the model are equal to zero is rejected.

### 2.5 Literature of studies that Utilized Logistics Regression Model

From the early 70's, more disciplines began to recognize the use of logistic regression method in analyzing dichotomous variables. Though complex in manual computations, the proliferation of the use of logistic regression has become so popular, especially as a result of the availability of readily available statistical software packages and qualified analysts that made computations and as well as necessary interpretation. It is an essential multivariate tool constantly use in medical sciences and well pronounced in cancer studies (Zhou et al., 2004). From thoracic surgical researches, the chart below gave an analytical description of the increasing percentage of journal

publications that have made use of logistic regression model with their corresponding year of publication (Anderson et al., 2003).

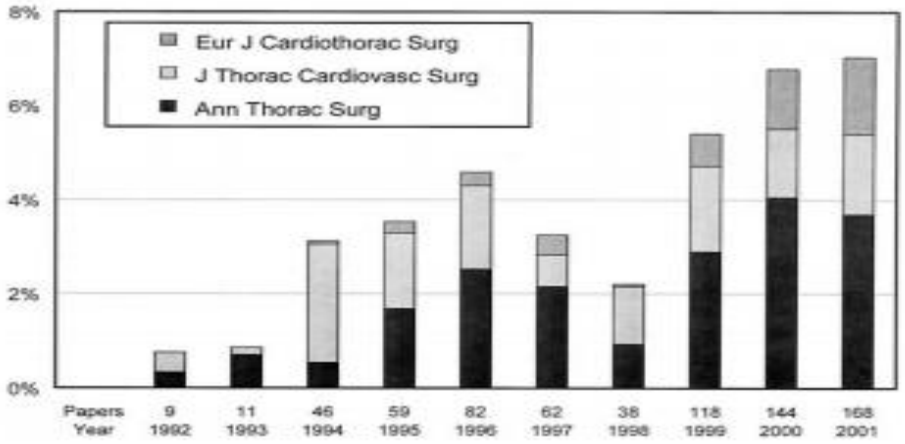


Figure 2: Thoracic Journal Publication

Excerpts of some of these studies in medical sciences are discussed as follows.

In a mammography research to determine the occurrence of breast cancer, 176 patients were selected for the study with patient menopause status, history of breast trauma, a related family member with cancer, presence of tissue mass as some of the risk factors considered in the study. Yusuf et al., (2013), made use of the multinomial logistic regression model and found that the risk of developing breast cancer is five times higher in patients with tissue mass than those not having.

Yoo et al., in 2012 conducted a study to identify the gene to gene interaction as well as the gene and environmental interaction using 1031 female patients found with non-small cell carcinoma. Four methods were considered in their study in order to compare and contrast best performing method in multigenic studies. The methods used included Logistic regression, Logic regression,

classification tree and random forest. In concluding their study, they found that Logistic regression is effective in explaining the effect of predictors on the outcome variable.

In order to examine the associated risk factors associated with post-operative anastomotic fistula with esophageal-cardiac cancer patients, Huang et al., (2017) adopted the usage of logistic regression to model these factors. Some of the factors considered in their study included age, gender, history of diabetes, smoking culture, surgery procedure. They concluded that female patients that underwent endoscopic surgery and also affected by renal dysfunction are at higher risk of post-operative anastomotic fistula.

Gupta et al., 2012 in their research on data mining, classification techniques applied to breast cancer concluded that Logistic regression technique is one of the veritable data mining technique to draw inference prognosis and risk factors associated with breast cancer occurrence.

By investigating the survival time of breast cancer patients using techniques like decision tree, logistic regression and Artificial Neural Network (ANN) methods, Delenin et al., 2005 used the Surveillance, Epidemiology and End Result (SEER) data research for their analysis. They concluded that logistic regression gave a good prediction; however, its predictive ability was lower compared to that of ANN and decision tree method.

Sathian in 2011 conducted a research with emphasis on how dependent dichotomous variable in medical research can be analyzed using the logistic regression method. The research question is to establish the effect of gender on blood clotting. The blood clothing was the dependent variable and was categorized as clotting time below or equal to 6 minutes (≤6 minutes) and a clotting time above 6 minutes (clotting time > 6 minutes). Nepalese students comprising of 64 male and

64 female students were used in the study. In conclusion, he found that female students are 3.46 times more likely to have a blood clotting occurrence in greater than 6 minutes duration compared to the male students.

Anderson et al., (2003) explaining the use of logistic regression in clinical studies applied this method on a dataset from a coronary artery bypass grafting study. The objective of the study was to investigate the effect of factors on death status of patients The independent variables considered in their study are patient age and the history of acute or chronic renal insufficiency (RENAL). The dependent variable is the death status of the patients under consideration. At the end of their study, they found that all the independent variables are statistically significant. The Odds value of the RENAL variable was 3.198 which indicate that the likelihood of a patient dying is tripled when there is history of acute or chronic renal insufficiency than when such history is absent.

Irfana et al., (2006) conducted a study on risk factors associated with ischemic heart disease. Examples of the risk variables considered in the study are cholesterol level, banaspati ghee, uric acid, residential location, age groups, protein level, phospholipids among other variables. Data used for the study consist of 585 patients from the Chandka medical college hospital in Pakistan. By using the backward stepwise elimination method of the logistic regression method, they found that variables such as ages of patients between 51 to 60 years, high cholesterol level, patient's residential area, usage of banaspati ghee heightened the risk of ischemic heart disease.
Pulmonary thromboembolism (PTE) is a condition that occurs as a result of blood cloth resulting in a blockage in the major arteries in the lung. Yoo et ., (2003) conducted a study to investigate

factors that tend to increase the risk of PTE. By using a retrospective autopsy record of 512 patients in a Brazil tertiary health institution, some risk factors considered in the study included age of patients, occurrence of trauma, cardiac thrombi, hypertension , sepsis and so on. By using a multiple logistic method, they found that the fatal prevalence of PTE is associated with age of the patients, patient trauma experience, pelvic vein thrombi, and right-sided cardiac thrombi.

## 2.6 An overview of Lung Cancer

The prevalence and incidence of cancer remain a heavy burden in both developing and the developed countries of the world. The risk of cancer is constantly on increment as a result of ageing population as well as behavioral lifestyles such as smoking, industrial pollution, obesity, and other associated factors linked to this disease. In a 2012 GLOBOCAN study, 8.2 million people died from cancer while 14.1% incidence was recorded worldwide. In this estimate, lung cancer is the most commonly diagnosed and a leading cause of fatality among the various types of cancer (Torre et al., 2012). Siegel et al., (2011) in their mortality trend of cancer diseases conclude that lung cancer has the highest record of death occurrences as depicted in the chart figure 3 below.

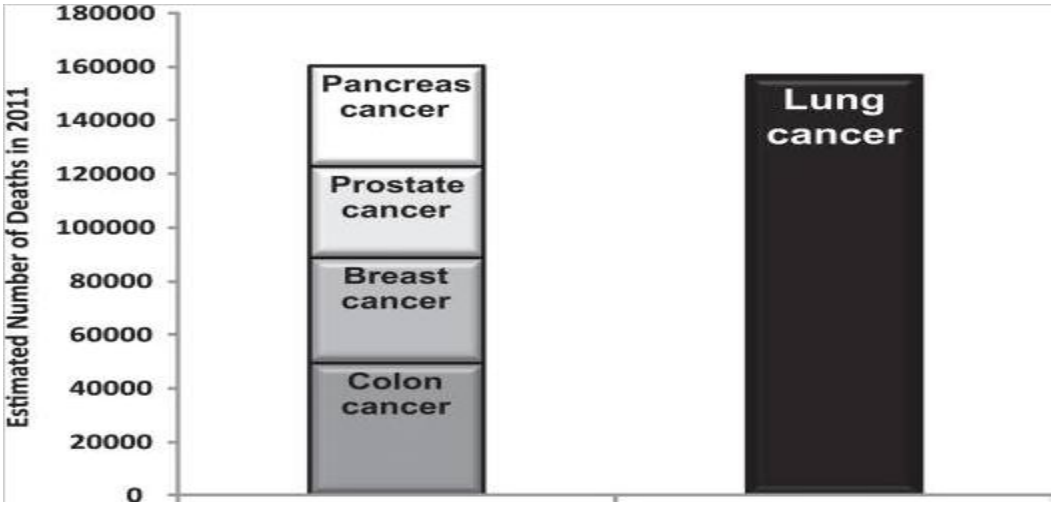


Figure 3: Death Incidence among major Cancer types ( Siegel et al.,2011)

Geographically, developed countries have higher cases of lung cancer occurrence as well as mortality from it while in most developing countries, these rates are lower, but showing a steady increasing tendency (Dela et al., 2011).

### 2.6.1 Lung Cancer Prognostic Risk Factors

Any attribute stemming from a patient that tends to influence disease diagnosis can be considered as a risk factor. In the oncology study of the cancer of the lungs, several risk factors have been identified from previous researches. These factors influence the survival outcomes of cancer patients as well as therapy procedures consider more appropriate. Cigarette smoking and for an example is strongly linked to the occurrence of about 80% to 90 % occurrence of lung cancer (Torre et al., 2015). By the year 2025, it has been projected that there will be about 1.9 billion smokers away from its current record of about 1.1 billion smokers.

And according to the World Health Organization, there will be a continuous increment in fatality from lung cancer as a result of the global rise in tobacco smoking and incidentally, cigarette smoking is quite attributed to the development of pulmonary carcinoma, which is also a major type of lung cancer (Guindon & Boisclair, 2009; Parkin et al.,1994). Aside cigarette smoking, the dietary intake of patients, age of a patient, gender category, treatment therapies, environmental factors as well as other socioeconomic factors tends to influence the prognosis and survival of lung cancer patients (Aldrich et al., 2013; Tu et al., 2016; Venuta et al., 2016; Suh et al., 2017; Zhu et al., 2017; Lin et al., 2018).

## 2.7 Overview of the Respiratory System

As the name implies, lung cancer occurs in lung organs found in the respiratory system of the body. The diagram below shows a typical respiratory system which equally houses the two lungs; organ which is elastic, spongy and cone-shaped in nature and structure located in the chest region of the body.

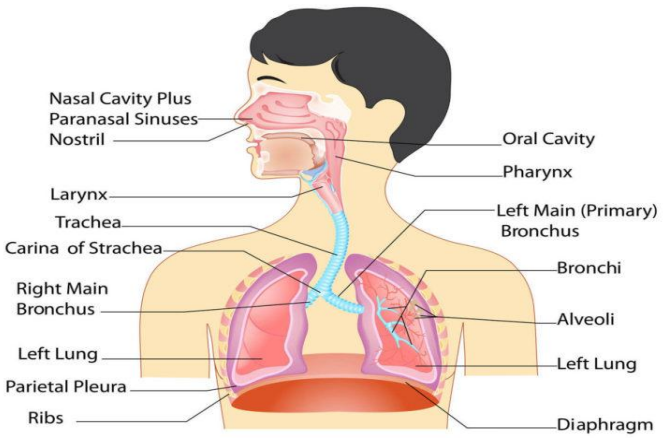


**Figure 4 : The Human Respiratory System**

The respiratory system is an advanced systemic structure in the body consisting of various organs responsible for the collective inhalation and exhalation of gases for the normal functioning of a biological body. It is made of components such as the mouth, pharynx, nose and nasal cavity, trachea (windpipes), larynx, bronchioles, bronchi, respiratory muscles and the lungs, which are classified into upper and lower tract respiratory system (Prince,1992). The lungs, which consist many air sacs called alveoli are a major component in this system and their primary role in the respiratory system involve prevention of harmful foreign objects through filtering, exchanging of breathe in and breathe out gases, conversion of inhaled airs and maintaining and adjusting the air to body temperature. The left lung comprising of two lobes and smaller in size primarily which made it successfully to house the heart and other critical structure while the right lung is bigger about 625 gm with three lobes. At each minute, there is a passage of about five litres of blood via the lungs necessary for gas conversion for inhalation as well as exhalation (Kowski, 2011).

Inhalation of air consists of about 21% of oxygen and other essential gases, but absence of carbon dioxide and is done through the nose and the mouth. The air passes through the bronchi and the windpipe then making an entry into the lungs via the right or left bronchi. The trapped air is further passed through the bronchioles, which are a smaller, divided, and complicated tubes. A secretion mechanism occurs in the bronchioles producing mucus which acts as screening agents to filter out foreign dirt objects contained in the trapped air. The bronchioles end with the alveoli in which the exchange of gases takes place. The alveoli make a rhythm of dilation and compression each moment there is an exchange of gases within it. The alveolus is compassed by a mass blood vessels known as the capillaries. The passage of the inhaled oxygen into the blood system through the blood plasma occurs when the level of concentration of oxygen is lower in

the capillaries but higher in the alveoli . The expelling of the carbon (IV) oxide therefore occurs when its concentration becomes higher in the capillaries as it diffuses into the alveoli, windpipes and finally into the atmosphere (Kowski, 2011).

## 2.8 Occurrence of the Lung Cancer and Histologic classification

The various air pathways down to the heart's capillaries are naturally designed to be free from any form of obstruction that can impede the flow cycle of gases. Any presence of impediment is capable of causing fatal damage to the whole respiratory system at large. Therefore, the uncontrolled growth of tissue that are malignant in nature commonly called cancer that grows in the lung region causes such obstruction which damages the natural sequence of the respiratory system**.** This uncontrolled growth of tissue forms a mass of lumps that can be referred to as malignant tissues and these tissues are cancerous**.**

These abnormal growths of tissue do occur more often when the lung is exposed to harmful substances that are carcinogenic in nature, especially as found in cigarette smoke and other poisonous substances such as asbestos. These substances damage the interior structure of the lung tissue as well as the bronchi which result into abnormal tissue growth that can metamorphosize into cancerous tumors (Rivera et al., 2003).

The lung cancer is majorly of two types, namely the non-small cell (NSCLC) and the small cell lung cancer (SCLC). The NSCLC is the most common malignant tissue growth and majorly consists of major types such as large cell carcinoma, squamous cell carcinoma, adenocarcinoma with other subdivisions (Koch, 2011). It accounts for about 85% of lung cancer type while the SCLC is said to account for the remaining 15 percent (Kowski, 2011; Keith, 2009). The table below shows an extract of the year 2015 World Health Organization (WHO) histological classification and sub- division of lung tumors.

Table 2. 1: An extract of the 2015 WHO Lung Tumor Classification (William et al., 2015)

| **Histologic Type and Subtypes** | **ICDO Code** |
|---|---|
| **Adenocarcinoma** | 8140/3 |
| Lepidic adenocarcinomae | 8250/3d |
| Acinar adenocarcinoma | 8551/3d |
| Papillary adenocarcinoma | 8260/3 |
| Micropapillary adenocarcinomae | 8265/3 |
| **Large cell carcinoma** | 8012/3 |
| Adenosquamous carcinoma | 8560/3 |
| **Squamous cell carcinoma** | 8070/3 |
| Keratinizing squamous cell carcinomae | 8071/3 |
| Nonkeratinizing squamous cell carcinomae | 8072/3 |
| Basaloid squamous cell carcinomae | 8083/3 |

The malignant epithelial tumor type of the adenocarcinoma is the most prevalent lung cancer cell type and it accounts for the major type of cancer found in non-smokers.

### 2.8.1 Lung Cancer Staging

In cancer research, staging which is a classification methodology plays a crucial decision making tool in terms of the therapy procedure, prognosis cycle, and information sharing considered fit for the studying and elimination of cancerous tumor growth. The staging classification is reviewed per time and developed by the International Association for the Study of Lung Cancer (IASLC). It is based on three major components, namely the extent of primary tumor (T), the involvement of regional lymph nodes (N) and the occurrence or non -occurrence of metastases

(M) and these are symbolically denoted as TNM descriptor (Uybico et al.,2010). This classification is broadly done from two major perspectives, namely the clinical staging description and the pathological description. The assessment of the lung cancer cell before undergoing any treatment procedure is referred to as the clinical staging while the prediction of the lung cancer growth following the pathological analysis of the lymph nodes, tumor and metastases through biospsy procedure is called the pathologic description (Rice, 2013).

**Table 2.2 : The Noninvasive Lung Cancer Staging TNM Description (Edge et al., 2010)**

| | |
|---|---|
| *Primary tumor* (*T*) | |
| TX | Primary tumor cannot be assessed, or tumor proven by the presence of malignant cells in sputum or bronchial washings but not visualized by imaging or bronchoscopy |
| T0 | No evidence of primary tumor |
| Tis | Carcinoma in situ |
| T1 | Tumor 3 cm or less in greatest dimension, surrounded by lung or visceral pleura, without bronchoscopic evidence of invasion more proximal than the lobar bronchus (i.e., not in the main bronchus)[a] |
| T1a | Tumor 2 cm or less in greatest dimension |
| T1b | Tumor more than 2 cm but 3 cm or less in greatest dimension |
| T2 | Tumor more than 3 cm but 7 cm or less or tumor with any of the following features (T2 tumors with these features are classified T2a if 5 cm or less): involves main bronchus, 2 cm or more distal to the carina; invades visceral pleura (PL1 or PL2); associated with atelectasis or obstructive pneumonitis that extends to the hilar region but does not involve the entire lung |
| T2a | Tumor more than 3 cm but 5 cm or less in greatest dimension |
| T2b | Tumor more than 5 cm but 7 cm or less in greatest dimension |
| T3 | Tumor more than 7 cm or one that directly invades any of the following: parietal pleural (PL3), chest wall (including superior sulcus tumors), diaphragm, phrenic nerve, mediastinal pleura, parietal pericardium; or tumor in the main bronchus (less than 2 cm distal to the carina[a] but without involvement of the carina); or associated atelectasis or obstructive pneumonitis of the entire lung or separate tumor nodule(s) in the same lobe |
| T4 | Tumor of any size that invades any of the following: mediastinum, heart, great vessels, trachea, recurrent laryngeal nerve, esophagus, vertebral body, carina, separate tumor nodule(s) in a different ipsilateral lobe |
| *Regional lymph nodes* (*N*) | |
| NX | Regional lymph nodes cannot be assessed |
| N0 | No regional node metastases |
| N1 | Metastases to ipsilateral peribronchial and/or ipsilateral hilar node(s) and intrapulmonary nodes, including involvement by direct extension |
| N2 | Metastases to ipsilateral mediastinal and/or subcarinal lymph node(s) |
| N3 | Metastases in contralateral mediastinal, contralateral hilar, ipsilateral or contralateral scalene, or supraclavicular node(s) |
| *Distant metastasis* (*M*) | |
| M0 | No distant metastasis |
| M1 | Distant metastasis |
| M1a | Separate tumor nodule(s) in a contralateral lobe tumor with pleural nodules or malignant pleural (or pericardial) effusion[b] |
| M1b | Distant metastasis (in extrathoracic organs) |

Based on more refined classification, the staging is concisely categorized into seven groups as shown in the table below. The advanced form of the NSCLC is categorized under the stage IIIB and IV groups.

**Table 2.3 : Stage Grouping in the 6th and 7th Editions of the TNM Staging (Uybico et al., 2010)**

| Stage | 6th Edition | 7th Edition |
|---|---|---|
| IA | T1, N0, M0 | T1a–T1b, N0, M0 |
| IB | T2, N0, M0 | T2a, N0, M0 |
| IIA | T1, N1, M0 | T1a–T1b, N1, M0 |
| | | **T2a, N1**, M0 |
| | | **T2b, N0**, M0 |
| IIB | T2, N1, M0 | **T2b, N1**, M0 |
| | T3, N0, M0 | T3, N0, M0 |
| IIIA | T3, N1, M0 | T1–T2, N2, M0 |
| | T1–T3, N2, M0 | T3, N1–N2, M0 |
| | | **T4, N0–N1**, M0 |
| IIIB | T4, N0–N2, M0 | **T4, N2**, M0 |
| | T1–T4, N3, M0 | T1–T4, N3, M0 |
| IV | T1–T4, N0–N3, M1 | T1–T4, N0–N3, M1a–M1b |

### 2.8.2 Performance Status

Originally developed by David A Karnofsky and his team of researchers in the year 1948, this is an assessment score used to rank the physical performance or activeness of a cancer patient by clinicians. It is considered a helpful guide to evaluate cancer patient survival period, life quality as well as a decisive tool for enlisting patients suitable for clinical trials of drugs or therapies (Firat et al., 2002). A newer performance metric score based on a 5 point scale called Eastern Co-operative Group (ECOG) performance scale was, however introduced to cater for necessary adjustments for simple denotation. It is also known as ECOG/WHO performance status score (Blagden et al., 2003).

**Table 2.4 : ECOG/ WHO Performance Score((Oken et al., 1982; Blagden et al., 2003)**

| ECOG PERFORMANCE STATUS* | |
|---|---|
| **Grade** | **ECOG** |
| 0 | Fully active, able to carry on all pre-disease performance without restriction |
| 1 | Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature, e.g., light house work, office work |
| 2 | Ambulatory and capable of all selfcare but unable to carry out any work activities. Up and about more than 50% of waking hours |
| 3 | Capable of only limited selfcare, confined to bed or chair more than 50% of waking hours |
| 4 | Completely disabled. Cannot carry on any selfcare. Totally confined to bed or chair |
| 5 | Dead |

## 2.9 Binary Logistic Model

As previously stated, the binary regression is a type of non-linear regression model whose outcome is dichotomous in nature by taking a 0 or 1 value as an outcome of success or failure. Mathematically , the simple binary logistic model is stated as :

$$\text{logit}(y) = \ln\left(p/1-p\right) = \alpha+\beta x$$

The multiple binary logistic model is given as :

$$\text{logit}(y) = \ln\left(p/1-p\right) = \alpha+\beta_1 x_1+\cdots+\beta_k x_k$$

Or

$$\log\left(\frac{\pi i}{1-\pi i}\right) = \sum_{k=0}^{K} xik\beta k \qquad i = 1,2,\ldots\ldots\ldots N \qquad\qquad equ\,(8)$$

Where Y takes the value of the event of interest which could be success or failure and *X* value(s) is the covariates or dependent variables.

### 2.9.1 Binary Logistic Model Assumptions

The Logistic regression model is not bound by the assumption of linearity, normal distribution or equal variances as seen in ordinary least square regression models, however, the assumptions that the parallel lines of the categorical variables need to be fulfilled. The dependent variable also needs to be dichotomous and relevant variables alone should be included in the model fitting.The coding of the outcome variable is also required to be correctly entered, for example, since the convectional probability for an event of success is denoted as 1 while the event of failure is denoted as 0.

### 2.9.2 Estimation of Parameter

Estimation of the k+1 of the $\boldsymbol{\beta}$ cooefficient of the *equ (8)* is the objective of logistic regression model. The maximum likelihood method is used to find this parameter estimate with a view to attain greater predictive probability from the observed data.

The probability distribution of the dependent variable is used to derive the maximum likelihood estimate which has a binomial distribution property (Czepiel, n.d).

The Joint probability density function of the maximum likelihood equation is given as:

$$f(y \mid \boldsymbol{\beta}) = \prod_{i=1}^{N} \frac{ni!}{yi!(ni-yi)!}\boldsymbol{\pi i}^{yi}\,(\mathbf{1}-\boldsymbol{\pi i})^{\,ni-yi} \qquad \ldots\ldots\ldots\ldots\ldots equ\ (10)$$

For every *ni trials*, the probability of success of yi is $\pi i^{yi}$ and the probability of ni-yi failure is $(1-\pi i)^{ni-yi}$ .

Expressing the $\boldsymbol{\beta}$ parameter as a known value for a fixed Y, we have derived the equation below

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{N} \frac{n_i!}{y_i!(n_i - y_i)!} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}$$ ……*equ (11)*

Performing two derivative processes on the above equation with respect to $\boldsymbol{\beta}$, we can obtain below expression

$$\prod_{i=1}^{N} \left( \frac{\pi_i}{1 - \pi_i} \right)^{y_i} (1 - \pi_i)^{n_i}$$ ……*equ (12)*

Taking the exponential of the two sides of *equ (8)*, we have:

$$\left( \frac{\pi_i}{1 - \pi_i} \right) = e^{\sum_{k=0}^{K} x_{ik}\beta_k}$$ ……..equ(13)

Solving for $\pi i$,

$$\pi_i = \left( \frac{e^{\sum_{k=0}^{K} x_{ik}\beta_k}}{1 + e^{\sum_{k=0}^{K} x_{ik}\beta_k}} \right)$$ …….*equ (14)*

Substititing *equ (13*) into the first term of the *equ (12)* and *equ (14)* into the second term of the *equ (12*), we have :

$$\prod_{i=1}^{N} (e^{\sum_{k=0}^{K} x_{ik}\beta_k})^{y_i} \left( 1 - \frac{e^{\sum_{k=0}^{K} x_{ik}\beta_k}}{1 + e^{\sum_{k=0}^{K} x_{ik}\beta_k}} \right)^{n_i}$$ ……*equ (15)*

$$\prod_{i=1}^{N}(e^{y_i \sum_{k=0}^{K} x_{ik}\beta_k})(1+e^{\sum_{k=0}^{K} x_{ik}\beta_k})^{-n_i}$$

……..*equ (16)*

Applying natural logarithm to *equ (16)*, we derive :

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N} y_i \left( \sum_{k=0}^{K} x_{ik}\beta_k \right) - n_i \cdot \log(1+e^{\sum_{k=0}^{K} x_{ik}\beta_k})$$

……..*equ (17)*

Finding the first derivative of the *equ (17)* by setting $\boldsymbol{\beta}$ to zero in order to obtain the critical point of the likelihood function, we have:

$$\begin{aligned}
\frac{\partial l(\beta)}{\partial \beta_k} &= \sum_{i=1}^{N} y_i x_{ik} - n_i \cdot \frac{1}{1+e^{\sum_{k=0}^{K} x_{ik}\beta_k}} \cdot \frac{\partial}{\partial \beta_k}\left(1+e^{\sum_{k=0}^{K} x_{ik}\beta_k}\right) \\
&= \sum_{i=1}^{N} y_i x_{ik} - n_i \cdot \frac{1}{1+e^{\sum_{k=0}^{K} x_{ik}\beta_k}} \cdot e^{\sum_{k=0}^{K} x_{ik}\beta_k} \cdot \frac{\partial}{\partial \beta_k} \sum_{k=0}^{K} x_{ik}\beta_k
\end{aligned}$$

$$= \sum_{i=1}^{N} y_i x_{ik} - n_i \cdot \frac{1}{1+e^{\sum_{k=0}^{K} x_{ik}\beta_k}} \cdot e^{\sum_{k=0}^{K} x_{ik}\beta_k} \cdot x_{ik}$$

$$= \sum_{i=1}^{N} y_i x_{ik} - n_i \pi_i x_{ik}$$

…….*equ (18)*

When the k+1 expression in *equ (18*) is set to zero and individual $\boldsymbol{\beta}$ parameters are solved for,the MLE for the $\boldsymbol{\beta}$ can therefore be derives follows:

$$\begin{aligned}\frac{\partial^2 l(\beta)}{\partial\beta_k\partial\beta_{k'}} &= \frac{\partial}{\partial\beta_{k'}}\sum_{i=1}^{N} y_i x_{ik} - n_i x_{ik}\pi_i \\ &= \frac{\partial}{\partial\beta_{k'}}\sum_{i=1}^{N} -n_i x_{ik}\pi_i \\ &= -\sum_{i=1}^{N} n_i x_{ik}\frac{\partial}{\partial\beta_{k'}}\left(\frac{e^{\sum_{k=0}^{K} x_{ik}\beta_k}}{1+e^{\sum_{k=0}^{K} x_{ik}\beta_k}}\right)\end{aligned}$$ ……*equ (19)*

Solving *equ (19)* by making the following differentiations assumptions,

$$\frac{\mathrm{d}}{\mathrm{d}x}e^{u(x)} = e^{u(x)}\cdot\frac{\mathrm{d}}{\mathrm{d}x}u(x)$$

And setting

$u(x) = \sum_{k=0}^{K} x_{ik}\beta_k.$

$$\left(\frac{f}{g}\right)'(a) = \frac{g(a)\cdot f'(a) - f(a)\cdot g'(a)}{[g(a)]^2}$$ ……..*equ (20)*

Putting this into *equ (19)*, we have the following expressions.

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{e^{u(x)}}{1+e^{u(x)}} = \frac{(1+e^{u(x)})\cdot e^{u(x)}\frac{\mathrm{d}}{\mathrm{d}x}u(x) - e^{u(x)}\cdot e^{u(x)}\frac{\mathrm{d}}{\mathrm{d}x}u(x)}{(1+e^{u(x)})^2}$$

$$= \frac{e^{u(x)}\frac{\mathrm{d}}{\mathrm{d}x}u(x)}{(1+e^{u(x)})^2}$$

$$= \frac{e^{u(x)}}{1+e^{u(x)}}\cdot\frac{1}{1+e^{u(x)}}\cdot\frac{\mathrm{d}}{\mathrm{d}x}u(x)$$ ……*equ (21)*

Thus, *equ (19)* can now as follows:

$$-\sum_{i=1}^{N} n_i x_{ik} \pi_i (1 - \pi_i) x_{ik'}$$

### 2.9.3 Logistic Regression Model Evaluation

**Likelihood Ratio Test**: In order to assess the strength contribution of each independent variable to the logistic regression model, the likelihood ratio test is conducted. The null hypothesis of the model assumes that there is no difference in the performance of a model irrespective of the presence of *K* independent variables (Zhang, 2015). The likelihood ratio test, however test the effectiveness of a logistic regression model with k independent variables against the model with the absence of independent variable(s). The test thus evaluate the changes that the outcome variable experiences as a result of the presence or absence of independent variable(s). In order to generate the test statistic, the -2log likelihood for the logistic model with no added variable is contrasted against the 2log likelihood of the logistic regression model with the variable(s). The resultant out gives a chi-squared measure which indicates how fit the model is with a K degree of freedom.

If the p-value for this contrasted differences is lesser than the given alpha level (often 0.05), we reject the null hypothesis and accept the alternative hypothesis by drawing a conclusion that at least one of the variables in the model has a significant effect.

**Hosmer & Lemeshow Test:** This test is equally a metric to access the goodness of fit of a logistic regression model(Cucchiara, 2012). The frequencies of observed events are contrasted against the expected frequencies of occurrence in the subdivision of the logistic regression equation.It is also based on a 10 probabilistic grouping method. It is mathematically stated as:

$$\mathbf{X_{HL}} = \sum_{g=1}^{n} \frac{(Og - Eg)^2}{Eg\left(1 - \frac{Eg}{ng}\right)}$$

The observed events are *Og* and *ng* but the expected event is *Eg*.

The index measuring this event comparison is also based on a Chi-square distribution. If the p-value is greater than the alpha value (say, 0.05), it means that the logistic regression model fits while at Hosmer and Lemeshow value below 0.05 is not considered a good model.

**Cox & Snell $R^2$ and Nagelkerke Pseudo $R^2$** : Like the $R^2$ in a linear regression model which explains the variability of the model and how much it performs, the Cox & Snell $R^2$ and Nagelkerke Pseudo $R^2$ equally plays this role in the Logistic Regression model. However, the Nagelkerke Pseudo $R^2$ is considered a more suitable measure to explain this variability than the Cox & Snell $R^2$ since it can get to up to 1 point numerical value which the Cox & Snell $R^2$ cannot attain (Hosmer and Lemeshow, 2000).

Mathematically, the Cox & Snell $R^2$ is stated as:

$$R^2_{CS} = 1.0 - \exp\left(\frac{2\ln L_{Full} - 2\ln L_{Intercept}}{N}\right)$$

While Nagelkerke Pseudo $R^2$ is stated as :

$$R^2_{Nag} = \frac{R^2_{CS}}{1.0 - \exp\left(\frac{-2\ln L_{Intercept}}{N}\right)}$$

Where $InL_{Full}$ is the loglikelihood of the model with an independent variable (current model) while $InL_{intercept}$ is the loglikelihood of the null hypothesis model.

### 2.9.4 Model Predictive Capability

**Classification Table:** This helps us to evaluate how well the binary logistic regression model can accurately predict an event occurrence in a close proximity to the already observed cases in consideration(Torosyan, 2017). This iterative method is done by classifying an event occurrence according to a pre-defined probability. The event of interest is classify as a success when is from the probability value upward while other cases below the set probability are considered (failure) . The sensitivity is considered the percentage of successes classification while the specificity is considered the percentage of failures. The event cases that are rightly classified is called true positives while those that are wrongly classify is called false negatives.

Table 2.5: Example of Classification Table

| | | **Expected** | |
|---|---|---|---|
| | | **1** | **0** |
| **Observed** | **1** | **a** | **b** |
| | **0** | **c** | **d** |

The "**a**" and the " **d** " are the true positive and true negative, respectively, while the "**c**" and the "**b** " are the falsely predicted cases.

Thus : Sensitivity = a/(a +c)

Specificity = d/(b+d)

If there are more cases recorded in the "**a**" and "**d** " rather than in the "**c**" and "**b** " cells, it means the model is a good fit in case classification. It can also be stated that higher values of sensitivity with a corresponding higher value of specificity is an indicator of a good model.

**Receiver Operating Characteristics (ROC):** This is a diagnostic technique with graphical output that tends to show the performance of a classification table analytically (Kumar & Indrayan, 2011; Tilaki, 2012). It produces a full graphic and table of score or value comparison rather than the two dimensional result the classification table depicts. The plot of the ROC consists of the sensitivity values which are plotted on the vertical axis and the 1-Specificity values which are plotted on the horizontal axis. The ROC is considered more informative as all the cut-off points can be compared and contrasted against one another**.** In many statistical software, the Area Under the Curve (AUC) result value tells the extent to which the model performs. Generally, any AUC value below 0.50 (50%) is considered a worthless model while those above that cut-off point of 0.50 (50%) are deemed okay. The closer the AUC value is to 1.00 (100%) the better the model.

# CHAPTER 3

# RESEARCH METHODOLOGY

## 3.1 Description of the Research Data

Data used in this study consist of 548 non small cell lung cancer patients with their various adjoining clinical, laboratory and demographic measurements. The data were obtained from the cancer data repository; an open source Cancer Research Organisation (CancerData, 2015). The variables used in the data analysis are:

Gender, WHO performance status, Body Mass Index (BMI), Forced Expiratory Volume (Fev1) in $1^{st}$ seconds, Smoking Status, Histology (Hist), T-stage, N-stage, Treatment method, Stage, Equivalent Radiation dose in 2-Gy fraction (Eqd2), Tumor Load, the Gross Tumor Volume (GTV) and the Patient Status (dead or alive)

The event of interest in the study is the death which is the binary regression dependent variable while other variables will be considered as the covariates.

**The Forced Expiratory Volume (Fev1)** is defined as the amount of air an individual can forcibly exhale from the lungs during a breath in 1 seconds. It is measured in percentage(%)

**Equivalent Radiation dose in 2-Gy fraction (Eqd2)** is a measure that accounts for the amount of radiation absorbed together with the effects that the type of the radiation exerts.

**Gross Tumor Volume (GTV)** is the extent to which a tumor is seen or observed during a standard examination procedure**.** It is measured in *ml.*

**Tumor Load (d)** can be defined as the number of cancer cells as well as the size of the tumor.

## 3.2 Research Analysis Methods

A step by step statistical research tools were being applied in the study.

Firstly, descriptive statistical tools was used to describe the nature and distribution pattern of the qualitative variables in terms of frequencies and percentages, while the mean, standard deviation (sd), median, minimum and maximum statistic will be computed for the quantitative variables. The normality test was also conducted on the quantitative variables to determine their suitability for a Parametric or a Non-parametric test. Parametric tests were applied to variables that are normally distributed while the Non-parametric tests were applied to the variables that are not normally distributed.

The test of hypotheses utilized in the study included the t-test, Mann Whitney test, Chi –square test, and the Logistic Regression method.

The Logistic Regression to be adopted include the Bivariate (Simple Logistic Regression) and the Multivariate Logistic Regression method (Multiple Logistic Regression model). Firstly, the Simple Logistic Regression was conducted for all the independent variables included in the study. Therefore, the Multiple Logistic Regression Model was then computed which included all the Simple Logistic Regression Models with a p-value $< 0.200$.

A Simple Receiver Operating Characteristic (ROC) curve for each of the quantitative variables was shown by making use of the probabilities derived from the logit functions of each of the Simple Logistic Regression Models for the quantitative variables. The ROC for the Multiple Logistic Regression Model was also generated by using the probabilities derived from the logit function of the Multiple Logistic Regression Model.

The SPSS (Demo version 20) software developed by the IBM Incorportion, New York, USA was used for all the analysis as well as the graphical outputs throughout the research. Also, The significance level of 0.05 was used for all the hypothesis testing computations.

# CHAPTER 4

# RESULT

**Table 4.1 : Descriptive Statistics of Quantitative Variables of the 548 Patients**

| Variables | Mean ± SD | Median(Min-Max) |
|---|---|---|
| **Age (Yrs)** | 65.69 ± 9.46 | 65.48 ( 36.00 – 88.00) |
| **BMI ($kg/m^2$)** | 24.99 ± 4.28 | 24.54 (14.30 – 39.50) |
| **Fev1 (%)** | 76.00 ± 21.19 | 77.00 (21.00 – 139.00) |
| **$Eqd_2$ (Gray)** | 59.63 ± 7.09 | 60.00 (39.80 – 77.90) |
| **GTV (ml)** | 88.57 ± 105.61 | 52.21 (0.00 – 725.00) |
| **Tumor Load (d)** | 122.80 ± 115.09 | 86.37 (3.40 -770.00) |

Table 4.1 gave the summary descriptive statistics of the quantitative variables of the data. The mean age (yrs) of the patients in the study is 65.69 ± 9.46 yrs while the average BMI value of the patients is 24.99 ± 4.28 $kg/m^2$ . The average Fev is 76.00 ± 21.19 %, while the mean $Eqd_2$ value is 59.63 ± 7.09 Gray. The median age of the patients is 65.48 yrs with a lower age of 36.00 yrs and the highest being 88.00 yrs old. The median value of the BMI is 24.54 $kg/m^2$ with the lowest value of 14.30 $kg/m^2$ and the highest value of 39.50 $kg/m^2$ . The median value of the

tumor load is 86.37 d with the patient having a lowest value of 3.40 d and the highest value of 777.00 d.

**Table 4.2 : Descriptive Statistics of Qualitative Variables (n= 548 Patients)**

| Variables | | n (%) |
|---|---|---|
| **Gender** | Male | 379 (69.20%) |
| | Female | 169 (30.80%) |
| **WHO Performance Status** | Restricted | 192 (35.00%) |
| | Capable | 287 (52.40%) |
| | Limited | 63 (11.50%) |
| | Missing | 6 (1.10%) |
| **Smoking Status** | Never/Ex Smoker | 305 (55.66%) |
| | Current Smoker | 202 (36.86%) |
| | Missing | 41 (7.48%) |
| **Histology** | SCC | 164 (29.90%) |
| | Adenocarcinoma | 81 (14.80%) |
| | LargeCells Carcinoma | 190 (34.70%) |
| | Others | 93 (17.00%) |
| | Missing | 20 (3.60%) |
| **T-Stage** | T0-1 | 74 (13.50%) |
| | T2 | 72 (31.40%) |
| | T3 | 60 (10.90%) |
| | T4 | 216 (39.40) |
| | Missing | 26 (4.70%) |
| **N-Stage** | N0 | 95 (17.34%) |
| | N1 | 15 (2.741%) |
| | N2 | 267 (48.72%) |
| | N3 | 167 (30.47%) |
| | Missing | 4 (0.73%) |
| **Staging** | | |
| | IIIA | 199 (36.30%) |
| | IIIB | 349 (63.70%) |
| **Treatment** | | |
| | No Chemo | 66 (12.00%) |
| | Sequential | 280 (51.10%) |
| | Concurrent | 202 (36.90%) |

In table 4:2, it can be found that the total number of male patients was 379 (69.20%) and the total number of female patients was 169 (30.80%). Patients that were considered to be capable of rendering self-care activities were 287 (52.4%), while those with limitation to render such activity were 63 (11.5%). In terms of tobacco usage, 305 (55.66%) of the patients were classified as none or ex-smokers while 202 (36.86%) were current smokers.

According to the T-stage Lung cancer classification, patients with a T4 stage had the highest number of patients in that category with 216 patients (39.4%) followed closely by the T2 category with 72 patients (31.4%) while the patients within the T3 stage had the lowest number of 60 patients (10.9%). Patients having Large cell carcinoma (LCC) lung cancer were the highest with a record number of 190 patients (34.7%), followed by patients with Small cell carcinoma with a total number of 164 patients (29.9%). Patients with adenocarcinoma type of lung cancer were 81 patients (14.8%), while 93 patients (17.00%) were unclassified. 20 (3.6%) of the patients in terms of histology classification were not reported.

In the N-Stage classification, the N2 stage category had 267 patients (48.72%) which made the category the highest, while the N1 stage had 15 patients (2.74%) which made it the lowest in the classification group. In terms of Staging classification, 199 patients (36.30%) of the patients were classified to be in the IIIA staging category, while 349 patients (63.70%) were classified into the IIIB stage. About the treatment method, 280 patients (51.1%) were subjected to sequential cancer treatment method while 202 patients (36.90%) were subjected to a concurrent cancer treatment plan. 66 patients (12%) of the patients were not subjected to any form of chemotherapy treatment.

**Table 4.3 : Bivariate Statistical Test for the Quantitative Variables**

| Variables | Outcome | n | Mean ± SD | Median(Min-Max) | Test Statistic | p |
|---|---|---|---|---|---|---|
| **Age(Yrs)** | Alive | 92 | 61.56 ± 7.83 | 62.12 (47.00-77.00) | Z = 2.434 | **0.015** |
| | Dead | 456 | 66.35 ± 9.54 | 66.74 (36.00-88.00) | | |
| **BMI** | Alive | 92 | 24.92 ± 4.35 | 24.31 (17.60-35.20) | Z = 0.058 | 0.954 |
| | Dead | 456 | 25.01 ± 4.26 | 24.65 (14.30-39.50) | | |
| **Fev(%)** | Alive | 92 | 78.88 ± 18.83 | 82.00 (27.00-122.00) | t = 1.300 | 0.194 |
| | Dead | 456 | 75.54 ± 21.54 | 76.50 (21.00-139.00) | | |
| **Eqd(Gray)** | Alive | 92 | 60.02 ± 7.29 | 60.18 (40.30-70.80) | Z = 3.577 | <**0.001** |
| | Dead | 456 | 59.57 ± 7.06 | 60.00 (39.80-77.90) | | |
| **GTV(ml)** | Alive | 92 | 75.32 ± 97.77 | 39.84 (0.00 - 366.50) | Z = 2.933 | <**0.001** |
| | Dead | 456 | 90.70 ± 106.84 | 55.78 (0.00 - 725.00) | | |
| **Tumor** | Alive | 92 | 115.87 ± 137.57 | 61.88 (3.40 - 629.5) | Z = 3.469 | <**0.001** |
| **Load(d)** | Dead | 456 | 123.91 ± 111.32 | 93.44 (4.60 - 770.00) | | |

Table 4.3 gives the result summary of the individual quantitative variables in the study. Because the Age, BMI, Fev, Eqd(Gray), GTV(ml) and the Tumor Load variables were not normally distributed, the Mann – U Whitnney test was used to test the significance of the difference within the patient outcome while the independent t-test was used to test the significance of the difference of the Fev variable because the Fev variable was normally distributed.

There is a statistically significance difference of the Age variable between the dead and alive patients ($p = 0.015$) which infers that the median age of the dead patients 66.74yrs (36.00-88.00) is higher than the median age 62.12yrs (47.00-77.00) of the patients that are alive. The median BMI value 24.65kg/$m^2$ (14.30-39.50) of the dead patients seem higher than that of the patients that are alive 24.31kg/$m^2$ (17.60-35.20) but there is no statistically significance difference between the dead and alive patients ($p = 0.954$).

The independent t-test shows that the Fev1 variable between the dead and alive patients is not statistically significant ($p = 0.194$). The $Eqd_2$ (Gray) variable between the dead and the living patients is significant ($p < 0.001$) which infers that the $Edq_2$ (Gray) median value 60.18 (40.30-70.80) for the dead patients is higher than that of the patients 60.00 (39.80-77.90) that are alive. The GTV variable between the dead and alive patients is also significant ($p < 0.001$) and it can be concluded that the median value 39.84 (0.00 - 366.50) of the GTV (ml) for the dead patients is higher than the median value 55.78 (0.00 - 725.00) of the patients that are alive .

The Tumor Load (d) variable between the dead and alive patients is equally found statistically significant ($p < 0.001$) which indicates that the median 93.44 (4.60 - 770.00) tumor level (d) Tumor Load of the dead patients is higher than the median 61.88 (3.40 - 629.5) Tumor level (d) of the patients that are alive.

**Table 4.4 : Bivariate Statistical Test for the Qualitative Variables**

| Variables | | Alive n (%) | Dead n (%) | $\chi^2$ | p |
|---|---|---|---|---|---|
| **Gender** | | | | 2.690 | 0.100 |
| | Male | 57 (15.00%) | 322 (85.00%) | | |
| | Female | 35 (20.70%) | 134 (79.30%) | | |
| **WHO Performance Status** | | | | 10.057 | **0.007** |
| | Restricted | 44 ( 22.90%) | 148 ( 77.10%) | | |
| | Capable | 41 (14.30%) | 246 ( 85.70%) | | |
| | Limited | 5 (7.90%) | 58 ( 92.10) | | |
| **Smoking** | | | | 0.107 | 0.743 |
| | Never/Ex Smoker | 48 (15.70%) | 257(84.30%) | | |
| | Current Smoker | 34 (16.80%) | 168 (83.20%) | | |
| **Histology** | | | | 7.526 | 0.057 |
| | SCC | 27(16.50%) | 137 (85.00%) | | |
| | Adenocarcinoma | 21 (25.90%) | 60 (74.10%) | | |
| | Large cell carcinoma | 24 (12.60%) | 166 (87.40%) | | |
| | Others | 18 (19.40%) | 75 (80.60%) | | |
| **T-Stage** | | | | 6.784 | 0.079 |
| | T0-1 | 16 (21.60%) | 58 (78.40%) | | |
| | T2 | 19 (11.10%) | 153 (89.00%) | | |
| | T3 | 11 (18.30%) | 49 (81.70%) | | |
| | T4 | 43 (19.90%) | 173 (80.10%) | | |
| **N-Stage** | | | | 7.664 | 0.053 |
| | N0 | 24 (25.30%) | 71 (74.70%) | | |
| | N1 | 4 (26.70%) | 11 (73.30%) | | |
| | N2 | 40 (15.00%) | 227 (85.00%) | | |
| | N3 | 23 (13.80%) | 144 (86.20%) | | |
| **Stage** | | | | 0.112 | 0.738 |
| | IIIA | 32 (16.10%) | 167 (83.90%) | | |
| | IIIB | 60 (17.20%) | 289 (82.80%) | | |
| **Treatment** | | | | 41.672 | < **0.001** |
| | No chemo | 4 (6.10%) | 62 (93.90%) | | |
| | Sequential | 27 (9.60%) | 253 (90.40%) | | |
| | Concurrent | 61 (30.20%) | 141 (69.80%) | | |

Table 4.4 gives the summary of the Bivariate analysis of the qualitative variables included in the study. The chi-square statistic shows that the WHO performance status category is significantly different ($\chi^2$=10.057,p < 0.05) relative to the patients' dead or alive outcome and the treatment methods are also significantly different ($\chi^2$=41.672, p <0.05) as regards to the patients' dead or alive outcome. There is no statistically significance difference in all other variables in respect to the dead or alive outcome of the patients.

In the gender category, 322 (85.00%) of the male patients died while 57 (15%) of them lived. 35 (20.70%) of the female patients live while 134 (79.30%) were deceased. 246 (85.70%) of the patients who are described as capable of rendering some basic self-care activities died while 41 (14.30%) of this class of patient live. Patients described as a current smoker recorded 168 (83.20%) death while 34 (16.80%) of them live. In the Non-smoker/Ex- smoker, 257 (84.30%) of these patients died, but 48 (15.70%) of them survived.

Patients with the large cell carcinorcoma recorded the highest number of death of 166 (87.40%), while patients with the occurrence of adenocarcinoma lived more with a total number 21 patients (25.90%), followed by 27 patients (16.50%) with small cell cancer. 153 Patients (89.00%) under the T2 category in the T-stage classification recorded the highest death in that classification while16 patients (21.60%) with T0-1 category has the highest surviving patient record.

In the N-stage classification, patients under the N1 category has 4 (26.70%) living patients while 144 (86.20%) patients under the N3 category recorded the highest number death occurrence followed by 227 (85.00%) patients under the N2. Patients with no chemotherapy treatment recorded the highest number of death of 62 (93.90%) patient record in the treatment

classification group while patients subjected to the concurrent treatment have the highest living patients of 61 (30.20%) patient record.

**Table 4.5 : Summary of the Bivariate Logistic Regression for each of the Variables**

Event of interest : Dead Patients Where ® = Reference Group

| Variable | B | SE | Odd Ratio 95%CI | Nagelkerke R-Square | Classification (%) | p |
|---|---|---|---|---|---|---|
| **BMI** | -0.002 | 0.035 | 0.998 (0.932 – 1.070) | 0.00 | 86.70% | 0.964 |
| **Age(yrs)** | 0.027 | 0.012 | 1.027 (1.004 – 1.051) | 0.02 | 83.20% | **0.022** |
| **$Edq_2$(Gray)** | -0.042 | 0.016 | 0.959 (0.929 – 0.990) | 0.02 | 83.20% | **0.010** |
| **Fev1(%)** | -0.008 | 0.006 | 0.992 (0.981 - 1.004) | 0.01 | 83.6% | 0.194 |
| **GTV(ml)** | 0.001 | 0.001 | 1.001 (0.999 - 1.004) | 0.00 | 83.40% | 0.368 |
| **Tumor Load(d)** | 0.001 | 0.001 | 0.992 (0.999-1.003) | 0.01 | 82.60% | 0.194 |
| **Gender** | | | | 0.01 | 83.20% | 0.102 |
| Female® | - | - | | | | |
| Male | 0.389 | 0.238 | 1.479 (0.925 - 2.353) | | | |
| **Smoking Status** | | | | 0.00 | 83.80% | 0.743 |
| Non-Smoker® | - | - | - | | | |
| Smoker | 0.08 | 0.245 | 1.084 (0.67-1.752) | | | |
| **Histology** | | | | 0.02 | 83.00% | 0.062 |
| SCC® | | | | | | |
| Adenocarcicoma | -0.574 | 0.330 | 0.563 (0.295 -1.074) | | | 0.081 |
| Large Cell | 0.310 | 0.303 | 0.725 (0.752 - 2.470) | | | 0.307 |
| Others | -0.197 | 0.336 | 0.821 (0.425 -1.588) | | | 0.558 |
| **Stage** | | | | 0.00 | 83.00% | 0.738 |
| IIIB® | | | | | | |
| IIIA | 0.080 | 0.240 | 1.083 (0.678-1.733) | | | |
| **T-Stage** | | | | 0.02 | 83.00% | 0.086 |
| T0-1® | | | | | | |
| T2 | 0.798 | 0.373 | 2.221 (1.070 - 4.612) | | | 0.320 |
| T3 | 0.206 | 0.437 | 1.229 (0.522 - 2.894) | | | 0.637 |
| T4 | 0.104 | 0.330 | 1.110 (0.581 – 2.118) | | | 0.752 |
| **N-Stage** | | | | 0.02 | 83.30% | 0.059 |
| N0® | - | - | - | | | |
| N1 | -0.073 | 0.630 | 0.930 (0.271 – 3.194) | | | 0.908 |
| N2 | 0.651 | 0.292 | 1.918 (1.083 – 3.399 | | | **0.026** |
| N3 | 0.750 | 0.326 | 2.116 (1.117 – 4.008) | | | **0.021** |
| **WHO Performance Status** | | | | 0.03 | 83.40% | **0.008** |
| Restricted® | - | | | | | |
| Capable | 0.579 | 0.241 | 1.784 (1.113 – 2.859) | | | **0.016** |
| Limited | 1.238 | 0.497 | 3.449 (1.303 – 9.130) | | | **0.013** |
| **Treatment** | | | | 0.120 | 83.20% | **<0.001** |
| No Chemo® | | | | | | |
| Sequential | -0.503 | 0.554 | 0.605 (0.204 – 1.791) | | | 0.364 |
| Concurrent | -1.190 | 0.554 | 0.149 (0.052 – 0.428) | | | **0.000** |

Table 4.5 gives the summary of the Bivariate Logistic Regression for each of the variables used in the study. The effect of the Age, and the $EDQ_2$ are the only significant variables in the quantitative variables, while others quantitative variables (BMI,GTV, and Tumor load), variables are not significant. The Nagelkerke R-Square values of each of the simple logistic regression models ranges from 0.00% to 1.60 %.

From the perspective of the categorical variables, it can be seen that only the WHO performance status variable and the Treatment variable were found to be significant in their respective logistic model. All other categorical variables such as Histology, Stage, T-stage, N-stage, and Smoking status are not statistically significant.

However, in order to develop a Multivariate Binary Logistic Regression Model, the Bivariate Logistic Models whose P-value $\leq 0.200$ are all variables included to build this model. These variables included the Age, $Edq_2$, Tumor load, Gender, Histology, T-stage, N-stage and the Treatment variable.

**Table 4.6: Omnibus Tests of Model Coefficients for the Multivariate Logistic Regression**

| | Chi-square | df | p-Value |
|---|---|---|---|
| Step | 54.400 | 15 | < **0.001** |
| Block | 54.400 | 15 | < **0.001** |
| Model | 54.400 | 15 | < **0.001** |

The Table 4.6 shows that the Multivariate Logistic Model is statistically significant since the p-value of < 0.001 is below the significant level of 0.05.

**Table 4.7: Model Summary (Multivariate for the Multivariate Logistic Regression comprising of all Logistic Regression Modes with p-value ≤ 0.200 )**

| -2**Log likelihood** | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|
| 383.303 | 0.111 | 0.185 |

The model summary shows that the multivariate logistic regression model explains between 0.111 and 0.182 variations of the effects on the probability of death occurrence in the patients. To present this more concisely, it can be stated that the Nagelkerke R Square value shows that the model explain 0.182 (18.2%) of the effects of the variables on the probability of death in the patients.

**Table 4.8: Hosmer and Lemeshow Test**

| **Chi - Square** | **df** | **p-Value** |
|---|---|---|
| 13.788 | 8 | 0.088 |

Table 4.8 shows the test of the fitness of the model. Since the p-value of 0.088 is greater than the alpha level of 0.05, it shows that the model *p=0.088 (>0.05)* is considered fit to the

data and significant in explaining the effect of the covariates on the death outcome of the patients

**Table 4.9: Classification Table for the Multivariate Logistic Regression Model**

| | | Predicted Outcome | | |
|---|---|---|---|---|
| | | **Alive** | **Dead** | **Percentage Correct** |
| **Observed** | **Alive** | 2 | 82 | 2.40 |
| | **Dead** | 7 | 370 | 98.10 |
| | **Overall Percentage** | | | 80.70 |

**Cut Off Point** = 0.5

The classification table describes how well the model categorizes the dependent outcomes. It can be seen that 82 living patients were erroneously classified as dead by the model and 7 dead patients were classified as living. The model does better at predicting patients that died than those living. However, the model correctly classifies 80.70% of the cases in the study.

**Table 4. 10: The Multivariate Logistic Regression Equations Summary**

| Variable | B | SE | Odd Ratio & 95% CI | p Value |
|---|---|---|---|---|
| **Age** | 0.07 | 0.015 | 1.007 (0.978 – 1.037) | 0.621 |
| **$Edq_2$(Gray)** | 0.024 | 0.019 | 0.976 (0.940 – 1.013) | 0.202 |
| **Tumor Load(d)** | 0.001 | 0.001 | 0.530 (0.998 - 1.003) | 0.530 |
| **Gender** | | | | 0.176 |
| **Female®** | | | | |
| **Male** | 0.387 | 0.285 | 1.472 (0.841 – 2.575) | |
| **Histology** | | | | 0.308 |
| SCC® | | | | |
| Adenocarcicoma | 0.514 | 0.387 | 0.598 (0.280-1.276) | 0.184 |
| Large Cell | 0.155 | 0.343 | 1.167 (0.596-2.286) | 0.652 |
| Others | 0.119 | 0.389 | 1.127 (0.526-2.413) | 0.759 |
| **T-Stage** | | | | 0.570 |
| T0-1® | | | | |
| T2 | 0.500 | 0.405 | 1.646 (0.746 – 3.645) | 0.217 |
| T3 | 0.005 | 0.505 | 0.995 (0.370 – 2.678) | 0.992 |
| T4 | 0.277 | 0.404 | 1.320 (0.598 – 2.913) | 0.492 |
| **Stage** | | | | 0.060 |
| N0® | | | | |
| N1 | 0.375 | 0.713 | 1.455 (0.360 – 5.884) | 0.599 |
| N2 | 0.814 | 0.373 | 2.257 (1.086 – 4.692 | **0.029** |
| N3 | 1.145 | 0.432 | 3.142 (1.348 – 7.322) | **0.008** |
| **Treatment** | | | | **<0.001** |
| No Chemo® | | | | |
| Sequential | 1.147 | 0.774 | 0.138 (0.700 – 1.446) | 0.137 |
| Concurrent | 2.910 | 1.887 | 0.091 (0.020- 0.414) | **0.002** |

Where ® = Reference Group

Table 4.10 gave the Multivariate Logistic regression. From the result output, it can be found that the significant variable in the model is the treatment variable with a p-value of <0.001. The age, Edq, Gender, Histology, T-stage and the N-stage variables are not significant. However, male patients were 1.472 (0.841 – 2.575) times more likely to die than female patients, though it is not statistically significant. The patients under the T2 stage were 1.649 (0.746 – 3.645) times more at risk of dying compared to patients in the T0-1 classification. In the N-stage classification, patients in the N2 category are 2.257 (1.086 – 4.692) times more at risk of death than those in the N0 stage while those in the N3 patients are 3.142 (1.348 – 7.322) times more at risk of death than the patients in the N0 category.

As regards to the treatment, patients undergoing sequential treatment plan are 0.137 (0.700 – 1.446) times more at risk of death than those subjected to no chemotherapy treatment. Otherwise, stating this, it can be stated that patients under the no chemotherapy treatments are 7.29 times (i.e. 1/0.137) more likely to die than those patients undergoing sequential treatment. Also, patients with no chemotherapy treatments are 0.091 (0.020 – 0.414) times more likely to die than those patients undergoing concurrent treatment. Likewise, it can be stated that patients under the no chemotherapy treatments are 10.989 times (i.e. 1/0.091) more likely to die than those patients undergoing concurrent treatment plan.

**Table 4.11: Area under the Curve for the ROC for the Quantitative Variables**

| Variables | Area under the Curve | SE | p-Value | CI (95%) |
|---|---|---|---|---|
| **Age(yrs)** | 0.652 (65.20%) | 0.039 | 0.001 | 0.576 – 0.728 |
| **BMI** | 0.512 (51.20%) | 0.048 | 0.808 | 0.417 – 0.606 |
| **Fev1(%)** | 0.558 (55.80%) | 0.044 | 0.221 | 0.644 – 0.472 |
| **Eqd2(Gray)** | 0.542 (54.20%) | 0.051 | 0.377 | 0.641 – 0.443 |
| **GTV(ml)** | 0.587 (58.70%) | 0.050 | 0.067 | 0.490 – 0.684 |
| **Tumor Load(d)** | 0.603 (60.30%) | 0.052 | 0.052 | 0.501 – 0.705 |

Table 4.11 presents the AUC of the quantitative variables used in the study. The statistical software used in the study was applied to compute the various Logit function of the bivariate logistic regression of these variables. The individual probabilities generated for each of the patients using the Logit function of their bivariate logistic regression models were then used to construct the ROC curves for the quantitative variables. It can be seen that the ROC for the best performing variable is the Age variable with an AUC of 0.652 (65.20%) closely followed by the Tumor load variable with an AUC of 0.603 (60.30%) and they are statistically significant. The lowest performing variable is the BMI (52.20%).

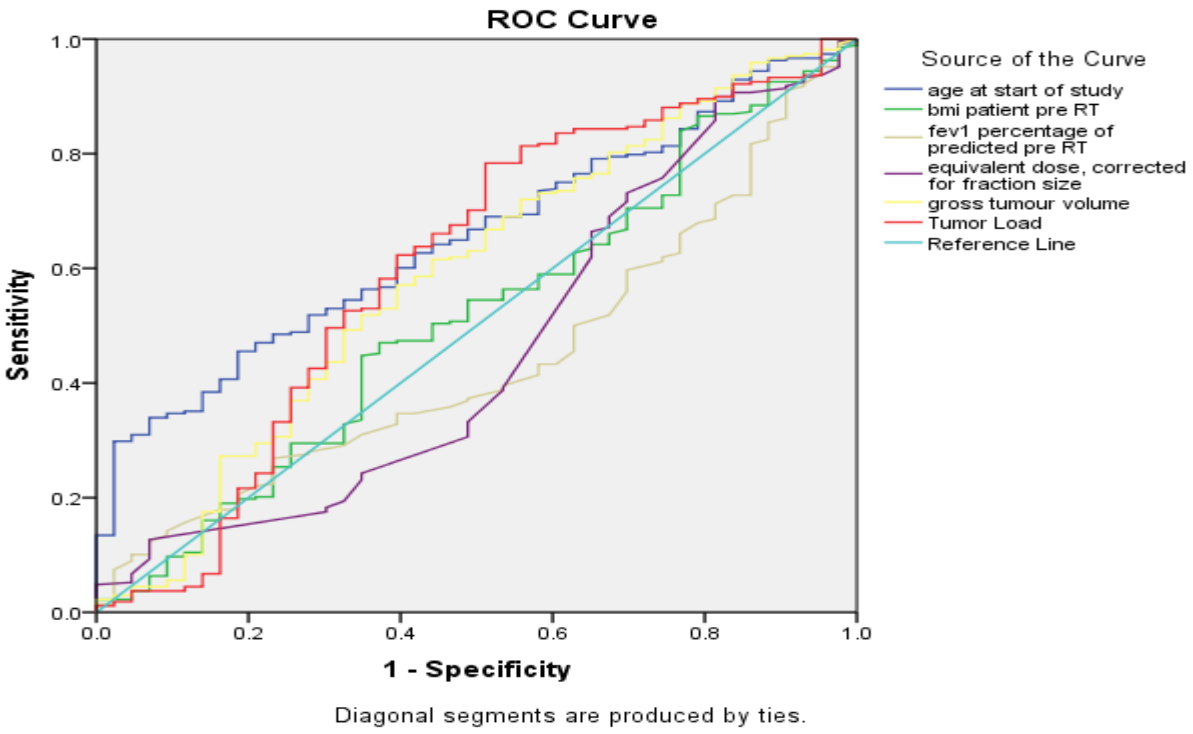


**Figure 5 : ROC Curve for the Quantitative Variables**

**Table 4.12: Area under the Curve for the ROC for the Multivariate Logistic Regression**

| Area under Curve | SE | p-Value | CI (95%) |
|---|---|---|---|
| **0.753** | 0.028 | **<0.001** | 0.696 – 0.089 |

Similarly to the ROC computation for the Bivariate Logistic Models for the quantitative variables, the Logit function of the Mulivariate Logistic Regression Model was used to compute the probability outcomes for each of the patients. These probability outcomes were subsequently used in the ROC in order to evaluate the model performance.

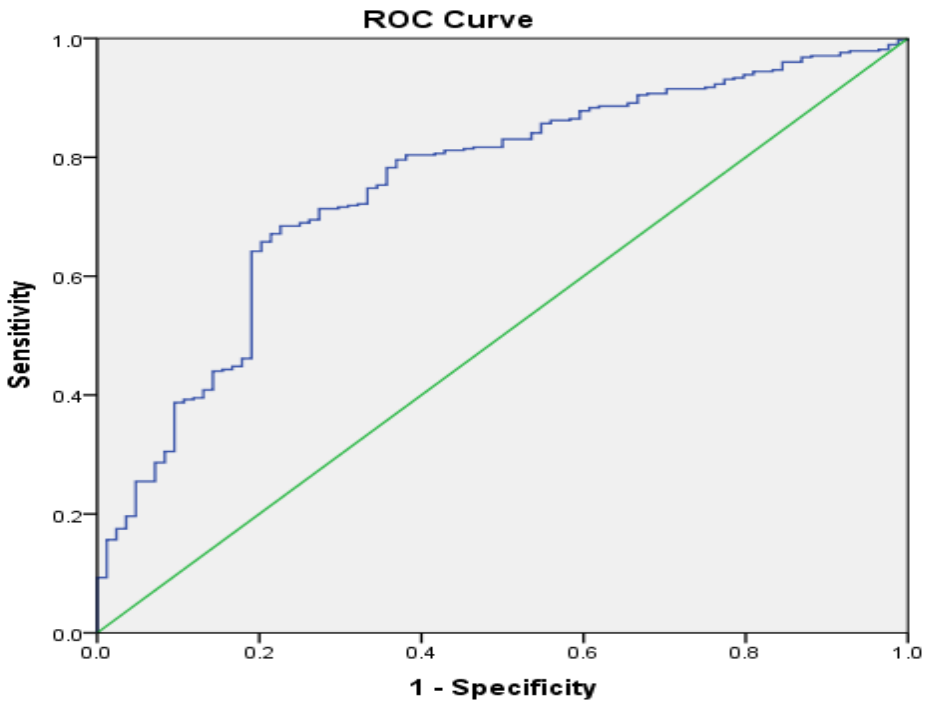


**Figure 6: ROC curve for the Final Multivariate Logistic Regression Model**

From the figure 6 and the diagram figure above, it can be seen that the Area under the Curve is 0.753 (75.3%) which indicates that the multivariate logistic regression is good enough to correctly classify patients that died or those that live in the study. The p-value of < 0.001 shows that the curve is statistically significant.

# CHAPTER FIVE

## 5.1 CONCLUSION AND RECOMMENDATIONS

As earlier indicated, statistics play an essential role in solving complex problems using analytic tools through which insights are being generated and proper decision guidance can be formulated especially in the medical field that is important for evidence based practice and intervention. The application of these various statistical tools is made to the thesis research data comprising of 548 Non-small cell lung cancer patients with a status of either been dead or alive. Using the status as the dependent variable,the dead patient, which is the event of interest in this thesis took the value of "1" and the living patients took the value of " 0 ". The some of the findings are reported as follows. The descriptive analysis result shows that death occurrence is higher in the female gender (85.00%) than the male gender (79.30%) though with no statistically significant difference .

The independent variables consist of thirteen (13) variables of which five (5) of them are quantitative while the remaining eight (8) variables are qualitative. Initially, the Bivariate Logistic Regression Model was applied to each of these variables to test for their individual significant effect on the event of interest (Death of Patient) at a significance level of 0.05. The the effect of the Age(yrs), and the $EDQ_2$ (Gray) are the only significant variables in the quantitative variables category and the WHO performance status variable and the Treatment variable are found significant in their respective logistic models.

However, variables with a p-value $\leq 0.200$ were all included to compute a Multivariate Logistic Regression Model. The variables that are finally included are Age(yrs), EDQ2(Gray), Gender,Histology, T-stage, N-stage and Treatment method.

The null hypothesis of the Multiple Logistic Regression that the covariates coefficients of the model are zero is rejected and that the model with the covariates is of good fit and found to be statistically significant $\chi 2\ (15)=54.00,\ (p<0.001)$ as given by the Omnibous test of the final model. The Nagelkerke $R^2$ value of 0.182 indicated that the model explained 18.20% of the variance in patients' death and 80.70 % of the cases are correctly classified as shown by the classification table. Increasing the age of the patients, $EDQ_2$ (Gray) and the Tumor load were associated with an increased likelihood of patients dying but these are not statistically significant. The treatment variable is the only significant variable in the Multivariate Logistic Model and patient under the no chemotherapy treatments is 7.25 times (i.e.1 /0.137) more likely to die than the patients undergoing sequential treatment.

The Receiver Operating Characteristics (ROC) of the Multivariate Logistic Model show that Area under the Curve (AUC) is 0.753 (75.3%) which indicates that the multivariate logistic regression is good enough to correctly classify patients that died or those that live in the study this is statistically significant. In contrast, the ROC for all other quantitative variables used in the study was analyzed and they were all found to have values below the ROC of the Multivariate Logistic Model with the highest being 0.652 (65.20%). This reflects the fact that Bivariate variables are not sufficient to predict an outcome of an event, but multivariate analysis with several variables provides a better predictive power for an outcome of interest.

The usage of logistic regression model is well pronounced in the medical research field. Prieto et al., (2017) conducted a study to investigate the biomarkers that could help identify the inception of Non-small cell lung cancer (NSCLC). The research subjects included 19 NSCLC patients and 19 control groups. Biomarkers such as Matrix metalloproteinases (TIMPs) and Tissue inhibitors of metalloproteinases (MPPs) were considered in the study. Their study made use of tests such as Mann-Whitnney test (which was used to test for the significance difference of the serum concentration), Fishers Exact test, ROC and the Logistic Regression model. They drew a conclusion that MMPs and TIMP-1 are found higher in the NSCLC patients' group serum than the control group and that MMP-9 is a good predictor of NSCLC in patients.

Similarly, a research was conducted by Hazra et al., (2017) on the prediction of lung cancer survivability using the logistic regression and the support vector machine (SVM) models. Variables included in their study consisting of 422 patients included age, gender, clinical M-stage, clinical N-stage, histology and overall stage. The SVM model was used to classify the event of death and being alive, of patients suffering from lung cancer, which was contrasted against the classification table of the logistic regression model for the best fit model with greater accuracy. Their study concluded that the logistic regression model gave a better classification accuracy of 77.40% than the SVM which gave a classification accuracy of 76.20%.

The increasing fatality associated with lung cancer could be attributable to late diagnosis and error in the treatment plan. However, with sophisticated analysis, such as logistic regression model formulation, early detection of lung cancer can be achieved and effective treatment plan can easily be formulated based on the various clinical, socioeconomic and laboratory analysis attributes.

It is however recommended that for further study, advanced machine learning algorithms can be used to learn more about the various interactions that exist within his variable and their respective effect on patient status. Also increment of sample size could lead to a greater insightful outcome. The prerequisite condition of choosing our variables into the multivariate logistic model which involves the selection of Bivariate Logistic Regression Models with a p-value $\leq 0.200$ might not be acceptable by divergent views, hence, it is recommended that subsequent research can be made with other significant levels to test if there are significant effect.

**REFERENCES**


Aldrich, M. C., Selvin, S., Wrensch, M. R., Sison, J. D., Hansen, H. M., Quesenberry, C. P., Wiencke, J. K. (2013). Socioeconomic Status and Lung Cancer: Unraveling the Contribution of Genetic Admixture. *American Journal of Public Health*. Vol.103, No 10, pp. e73–e80. http://doi.org/10.2105/AJPH.2013.301370

Alexopoulos, E. C. (2010). Introduction to Multivariate Regression Analysis. *Hippokratia*, Vol.14 , No 1, pp. 23–28.

Al-Ghamdi A.S., (2001). Using Logistic regression to Estimate the Influence of Accident Factors on Accident Severity. *Accident Analysis and Prevention,* Vol. 34, pp. 729–741.

Anderson Richard. P., Ruyun Jin , & Gary L. Grunkemeier ( 2003). Understanding Logistic Regression Analysis in Clinical Reports: An Introduction. *Annals of Thoracic Surgery*. Vol.75, pp. 753–757

Awodele Olufunsho, Adeyomoye A. A. , Awodele DF. , Fayankinnu Vincent B., Duro C. Dolapo (2011). Cancer distribution pattern in southwestern Nigeria**.** *Tanzania Journal of Health Research* , Vol. 13, Issue 2.

Bagleya Steven C., Halbert Whiteb, Beatrice A. Golomb (2001). Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. *Journal of Clinical Epidemiology*. Vol. 54, pp. 979–985

Banerjee, A., & Chaudhury, S. (2010). Statistics without tears: Populations and samples. *Industrial Psychiatry Journal*, Vol.*19* , No1, pp. 60–65.

Binu, V.S, Chandrashekhar, T.S, Subba, S.H. (2007) Cancer pattern in Western Nepal: A Hospital based retrospective study. *Asian Pacific Journal of Cancer Prevention*, Vol. 8, pp.183-186.

Blagden SP, SC Charman SC, Sharples LD , Magee LRA & Gilligan D (2003). Performance status score: do patients and their oncologists agree? *British Journal of Cancer*. Vol. 89, pp.1022 – 1027

Brambilla E, Travis WD (2014). Lung cancer. In World Cancer Report, Stewart BW, Wild CP (Eds), World Health Organization, Lyon.

Canova, S., Cortinovis, D. L., & Ambrogi, F. (2017). How to describe Univariate data. Journal of Thoracic Disease, Vol.9, No 6, pp. 1741–1743. http://doi.org/10.21037/jtd.2017.05.80

Cary Oberije, DirkDe Ruysscher , Ruud Hoube, Michelvan de Heuvel, Wilma Uyterlinde, Joseph O.Deasy, Jose Belderbos, Anne-Marie C.Dingemans, Andreas Rimner, Shaun Din, Philippe Lambin (2015). A Validated Prediction Model for Overall Survival From Stage III Non-Small Cell Lung Cancer: Toward Survival Prediction for Individual Patients. *International Journal of Radiation Oncology , Biology , Physics.* Vol., 92, Issue 4, pp. 935-944

Cramer, J.S., The Origins of Logistic Regression (December 2002). *Tinbergen Institute Working Paper* No. 2002-119/4. Available at SSRN: http://dx.doi.org/10.2139/ssrn.360300

Cucchiara Andrew (2012). Applied Logistic Regression. *Technometrics*, Vol. 34, No 3, pp. 358-359

Czpiel Scott (nd). Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation . Retrieved from https://czep.net/stat/mlelr.pdf on 1st of May,2018.

Dahiru, T. (2008). *P* – Value, a True Test of Statistical Significance? A Cautionary Note. *Annals of Ibadan Postgraduate Medicine*, Vol. *6, No* 1, pp. 21–26.

Dela Cruz, C. S., Tanoue, L. T., & Matthay, R. A. (2011). Lung Cancer: Epidemiology, Etiology, and Prevention. *Clinics in Chest Medicine*, Vol. *32*, No 4, pp. 605-644 http://doi.org/10.1016/j.ccm.2011.09.001

Delen Dursun, Walker Glenn & Kadam Amit (2015). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine*. Vol.34, pp. 113-127.

Edge SB, Byrd DR, Compton CC, Fritz AG, Greene FL, Trotti A(2010). *AJCC cancer staging manual* (7th ed). New York, NY: Springer.

Firat S, Byhardt RW, Gore E., (2002). Comorbidity and Karnofksy performance score are independent prognostic factors in stage III non-small-cell lung cancer: an institutional analysis of patients treated on four RTOG studies. Radiation Therapy Oncology Group. *International Journal Radiation Oncology Biology Physics*. Vol. 54, pp. 357–364.

Guindon GE, Boisclair D (2009). *Past, current and future trend in tobacco*. Washington, DC: International Bank for Reconstruction and Development. The World Bank; 2009. pp. 13–16.

Gupta, A., Shridhar, K., & Dhillon, P. K. (2015). A review of breast cancer awareness among women in India: Cancer literate or awareness deficit? *European Journal of Cancer*.Vol. 51, No 14, pp. 2058–2066. http://doi.org/10.1016/j.ejca.2015.07.008

Hatterjee, Samprit & Hadi, Ali. (2006). *Regression Analysis by Example*, Fourth Edition, Wiley & Sons, Inc,USA.

Hosmer David and Lemeshow Stanley (2000). Applied Logistic Regression.Second Edition. John Wiley & Sons. New York, N.Y, USA.

Huang Jinxi , Yi Zhou , Chenghu Wang , Weiwei Yuan , Zhandong Zhang , Beibei Chen & Xiefu Zhang (2017). Logistic regression analysis of the risk factors of anastomotic fistula after radical resection of esophageal-cardiac cancer. *Thoracic Cancer*. Vol. 8, No 6, pp. 666-671

Irfana P. Bhatti, Heman D. Lohano , Zafar A. Pirzado & Imran A. Jafri (2006). A Logistic Regression Analysis of the Ischemic Heart Disease Risk. *Journal of Applied Sciences*. Vol. 6 , Issue No 4 , pp. 785-788.

Keith, R. L. (2009). Chemoprevention of Lung Cancer. *Proceedings of the American Thoracic Society*, Vol. *6, No* 2, pp. 187–193. http://doi.org/10.1513/pats.200807-067LC

Kenkel N. C. (2006). On selecting an appropriate multivariate analysis. *Canadian Journal of Plant Science*, Vol.86, pp.663-676, https://doi.org/10.4141/P05-164.

King, G. & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, Vol. 9, pp. 137 -163.

Koch Andrea (2011). *Clinical Aspects of Inflammation in Non-small Cell Lung Cancer*.An unpublished thesis submitted to the Department of Medical and Health Sciences Linköping University, Sweden. Retrieved from http://www.diva-portal.org/smash/get/diva2:420479/FULLTEXT01.pdf on26th of May, 2018

Kowski, Margaret Anne (2011). *Gender Differences in Lung Cancer Treatment and Survival*. Graduate Theses and Dissertations submitted to the University of South Florida.Retrieved from .http://scholarcommons.usf.edu/etd/3191on 26th of May, 2018

Kumar Rajeev and Indrayan Abhaya (2011). Receiver Operating Characteristic (ROC) curve for medical researchers. *Indian Pediatrics*, Vol.48: 277. https://doi.org/10.1007/s13312-011-0055-4

Lin Xiaojing Lin , Liu Lingli , Fu Youyun , Gao Jing , He Yunyun , Wu Yang Wu and Lian Xuemei (2018). Dietary Cholesterol Intake and Risk of Lung Cancer: A Meta-Analysis. *Nutrients*. Vol 10, No 185; doi:10.3390/nu10020185.

Medenhall William and Sincich Terry (2003). *A second course in Statistics Regression Analysis*. 6 Edition.Pearson Education ,Inc, New Jersey,USA.

Menard, S. W. (2001). *Applied Logistic Regression Analysis* (quantitative applications in the social sciences) (2nd ed.). Thousand Oaks, CA: Sage Publications.

Mircioiu Constantin and Atkinson Jeffrey (2017). A Comparison of Parametric and Non-Parametric Methods Applied to a Likert Scale. *Pharmacy*, Vol. 5, No 2, pp.26 doi:10.3390/pharmacy5020026

Oken, M.M., Creech, R.H., Tormey, D.C., Horton, J., Davis, T.E., McFadden, E.T., Carbone, P.P(1982). Toxicity And Response Criteria Of The Eastern Cooperative Oncology Group. *American Journal of Clinical Oncolology*. Vol 5, pp. 649-655.

Oommen, T., Baise, L. G., & Vogel, R. M. (2011). Sampling bias and class imbalance in Maximum-likelihood Logistic Regression. *Mathematical Geoscience*.Vol. 43, 99–120

Oser, M. G., Niederst, M. J., Sequist, L. V., & Engelman, J. A. (2015). Transformation from non-small-cell lung cancer to small-cell lung cancer: molecular drivers and cells of origin. *The Lancet. Oncology*, Vol.*16, Issue* 4, e165–e172.

Park, Hyeoun-Ae (2013). An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain. *Journal Korean Academy Nursing.* Vol.43 No.2, pp.154-164, http://dx.doi.org/10.4040/jkan.2013.43.2.154

Parkin DM, Pisani P, Lopez AD(1994). At least one in seven cases of cancer is caused by smoking. Global estimates for 1985. International Journal of Cancer. Vol. 59, No 4, pp. 494–504

Price SA, McCarthy-Wilson L. Pathophysiology (1992): *Clinical Concepts of Disease Processes*. Fourth ed. St. Louis: Mosby

Pullinger John(2013). Statistics making an impact. *Journal of Royal Society of Statistics*. Vol. A 176, Part 4, pp. 819–839

Rice, T. W., & Blackstone, E. H. (2013). Esophageal Cancer Staging. Past, Present, and Future. *Thoracic Surgery Clinics*, Vol. 23, No 4, pp. 461-469. DOI: 10.1016/j.thorsurg.2013.07.004

Rivera Edgardo, Vicente Valero, Banu Arun, Melanie Royce, Rosni Adinin, Karen Hoelzer, Ronald Walters, James L. Wade III, Lajos Pusztai, Gabriel N. Hortobagyi (2003). Phase II Study of Pegylated Liposomal Doxorubicin in Combination With Gemcitabine in Patients With Metastatic Breast Cancer. *Journal of Clinical Oncology*. Vol. 21, No. 17, pp. 3249-3254.

Sathian Brijesh (2011). Reporting dichotomous data using Logistic Regression in Medical Research: The scenario in developing countries. Nepal Journal of Epidemiology. Vol.1, No 4, pp. 111-113

Schmidta Amand F., and Finana Chris (2017). Linear regression and the Normality Assumption. *Journal of Clinical Epidemiology*.Vol. 98, pp. 146–151

Sedgwick Philip (2012). Parametric v non-parametric statistical tests. *BMJ* ; 344:e1753 doi: 10.1136/bmj.e1753.

Siegel Rebecca L. , Miller Kimberly D., Jemal Ahmedin (2017). Cancer statistics, 2017. *A Cancer Journal for Clinicians*. Vol. 67, pp.7–30.

Silverstri GA, Jett JR (2010). *Clinical aspects of lung cancer*. In R Mason et al., Eds., Murray and Nadel's Textbook of Respiratory Medicine, 5th ed., Vol. 2, pp. 1116-1144. Philadelphia: Saunders

Sonia Blanco-Prieto , Leticia Barcia-Castro , María Páez de la Cadena , Francisco Javier Rodríguez-Berrocal , Lorena Vázquez-Iglesias , María Isabel Botana-Rial , Alberto Fernández-Villar and Loretta De Chiara (2017). Relevance of matrix metalloproteases in non-small cell lung cancer diagnosis. *BMC Cancer*. Vol. 17, No 823. DOI 10.1186/s12885-017-3842-z

Stewart, B. W. & Kleihues, P. (Eds.). (2003) *World Cancer Report*. Lyon, France: International Agency for Research on Cancer. www.scribd.com/.../World-Cancer-Report

Stoltzfus Jill C. (2011). Logistic Regression: A Brief Primer. *Academic Emergency Medicine* Vol. 18, pp. 1099–1104. doi: 10.1111/j.1553-2712.2011.01185.x

Suh, W. N., Kong, K. A., Han, Y., Kim, S. J., Lee, S. H., Ryu, Y. J., Chang, J. H. (2017). Risk factors associated with treatment refusal in lung cancer. *Thoracic Cancer*, Vol. 8, No 5, pp. 443–450. http://doi.org/10.1111/1759-7714.12461

Tetrault, J. M., Sauler, M., Wells, C. K., & Concato, J. (2008). Reporting of multivariable methods in the medical literature. *Journal of Investigative Medicine*. Vol.56, No7, pp.954-957. http://dx.doi.org/10.231/JIM.0b013e31818914f

Tilak Karimollah Hajian(2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian Journal of Internal medicine* .Vol. 4, No 2, pp. 627–635.

Torosyan Nare (2017). Application of binary logistic regression in credit scoring. An unpublished thesis submitted to the Institute of Mathematics and Statistics, University Of Tartu, Estonia. Retrieved from https://dspace.ut.ee/bitstream/handle/10062/57102/torosyan_nare_msc_2017.pdf?sequence=1&isAllowed=y on 1st of May, 2017.

Torre Lindsey A., ; Bray Freddie, Siegel Rebecca L., ; Ferlay Jacques, Joannie Lortet-Tieulent,; Jemal Ahmedin (2015). Global Cancer Statistics, 2012 . *International Journal of Cancer*. Vol.65, pp. 87–108

Torre Lindsey A, Rebecca L. Siegel, Elizabeth M. Ward, and Ahmedin Jemal (2015). Global Cancer Incidence and Mortality Rates and Trends—An Update. Cancer Epidemiology, Biomarkers &Prevention.Retrieved from https://commed.vcu.edu/Chronic_Disease/Cancers/2016/globalupdate.pdf on 21st of Feb , 2018.

Tu, H., Heymach, J. V., Wen, C.-P., Ye, Y., Pierzynski, J. A., Roth, J. A., & Wu, X. (2016). Different dietary patterns and reduction of lung cancer risk: A large case-control study in the U.S. *Scientific Reports*.Vol. *6*, No 26760. http://doi.org/10.1038/srep26760

Uyanık Gülden Kaya , Güler Neşe ( 2013). A Study of Multiple Linear Regression Analysis. *Procedia - Social and Behavioral Sciences*, Vol.106, No 10, pp. 234-240.

UyBico Stacy J., Wu Carol C., Suh Robert D., Le Nanette H., Brown Kathleen, Krishnam Mayil S (2010). Lung Cancer Staging Essentials: The New TNM Staging System and Potential Imaging Pitfalls. *RadioGraphics*. Vol. 30, pp. 1163–1181. https://pubs.rsna.org/doi/pdf/10.1148/rg.305095166.

Venuta, F., Diso, D., Onorati, I., Anile, M., Mantovani, S., & Rendina, E. A. (2016). Lung cancer in elderly patients. *Journal of Thoracic Disease*, *8* (Suppl 11), S908–S914. http://doi.org/10.21037/jtd.2016.05.20

Wayne W.Daniel (2010). Biostatistics:*Basic Concepts and Methodology for the Health Sciences*. 9 edition.John Wiley & Sons,INC.Hoboken,New Jersey,USA

William D. Travis, Elisabeth Brambilla, , Andrew G. Nicholson, , Yasushi Yatabe, , John H.M. Austin, , Mary Beth Beasley, Lucian. R. Chirieac, Sanja Dacic, Edwina Duhig, Douglas B. Flieder, Kim Geisinger, Fred R. Hirsch, Yuichi Ishikawa, Keith M. Kerr, Masayuki Noguchi, Giuseppe Pelosi, Charles A. Powell, Ming Sound Tsao, Ignacio Wistuba (2015). The 2015 World Health Organization Classification of Lung Tumors. *Journal of thoracic oncology*. Vol. 10, No 9, pp. 1243–1260

Wojciech Drozd (2017). Logistic Regression in the Identification of Hazards in Construction. *IOP Conf. Series: Materials Science and Engineering,* Vol. 245, No 062012. doi:10.1088/1757-899X/245/6/062012.

Yoo HH1, De Paiva SA, Silveira LV,Queluz TT., (2003). Logistic Regression Analysis Of Potential Prognostic Factors For Pulmonary Thromboembolism.*Chest*. Vol.123, No 3, pp. 813-821.

Zappa, C., & Mousa, S. A. (2016). Non-small cell lung cancer: current treatment and future advances. *Translational Lung Cancer Research*, Vol.*5*, *Issue* 3, pp.288–300. http://doi.org/10.21037/tlcr.2016.06.07

Zhang J, Wu Y (2005) Likelihood-ratio tests for normality.. *Journal of Statistical Computation and Simulation.* Vol. 49, pp. 709–721

Zheng Jie, Marcelline R. Harris , Anna Maria Masci, Yu Lin , Alfred Hero , Barry Smith and Yongqun He (2016). The Ontology of Biological and Clinical Statistics (OBCS) for standardized and reproducible statistical analysis. *Journal of Biomedical Semantics*, Vol. 7,No 53, pp. 1-13 DOI 10.1186/s13326-016-0100-2.

Zhou, X., K. Y. Liu, and S. T. C. Wong(2004). Cancer classification and Prediction using Logistic Regression with Bayesian Gene Selection. *Journal of Biomedical Informatics*.Vol. 37, No 4, pp. 249-259.

Zhu Yong-Jian , Bo Ya-Cong , Liu Xin-Xin , Qiu Chun-Guang(2017). Association of dietary Vitamin E intake with Risk of Lung Cancer: a Dose-Response Meta-Analysis. *Asia Pacific Journal of Clinical Nutrition.* Vol. 26, No 2, pp. 271-277

Zoubida Zaidi, Mokhtar Hamdi Cherif (2016). PS01.07: Geographical Distribution of Lung Cancer Mortality Worldwide. *Journal of Thoracic Oncology*. Vol. 11, No 11, pp. S273–S274