MODELING BRONCHO-PNEUMONIA STATUS IN INFANTS USING DISCRIMINANT AND LOGISIC REGRESSION ANALYSES

BY

Sule Ahmed SHEHU
M.Sc/SCI/20789/2012 – 2013

A DISSERTATION SUBMITTED TO THE SCHOOL OF POSTGRADUATE STUDIES,
AHMADU BELLO UNIVERSITY, ZARIA
IN PARTIAL FULFILLMENT OF THE REQUIREENTS FOR THE AWARD OF DEGREE OF MASTER OF SCIENCE IN STATISTICS

DEPARTMENT OF STATISTICS,
FACULTY OF PHYSICAL SCIENCES
AHMADU BELLO UNIVERSITY,
ZARIA, NIGERIA

APRIL, 2017

DECLARATION

I, **Sule Ahmed Shehu** declare that the work in this dissertation titled "**MODELING BRONCHO-PNEUMONIA STATUS IN INFANTS USING DISCRIMINANT AND LOGISTIC REGRESSION ANALYSES** has been carried out by me in the Department of Statistics under the supervision of Dr. A. Isah and Dr. H. G Dikko.The information derived from the literature has been duly acknowledged in the text and a list of references provided. No part of this dissertation has been published.

Sule Ahmed SHEHU	_____	_____
Name of Student	Signature	Date

CERTIFICATION

This dissertation titled "**MODELING BRONCHO-PNEUMONIA STATUS IN INFANTS USING DISCRIMINANT AND LOGISTIC REGRESSION ANALYSES** by Sule Ahmed Shehu (MSC/SCI/20789/2012-2013) meets the rules and regulations governing the award of the Degree of Masters of Science of Ahmadu Bello University, Zaria and is approved for its contribution to knowledge and literary presentation.

Dr. A. Isah Chairman, Supervisory Committee	Date
Dr. H. G Dikko Member, Supervisory Committee	Date
Prof. B. A Oyejola External Examiner	Date
Dr. H. G Dikko Head of Department	Date
Prof. S. Z. Abubakar Dean, School of Postgraduate Studies	Date

DEDICATION

This work is dedicated to my late parents and my unforgettable uncle, for their unflinching support for my upbringing.

ACKNOWLEDGEMENT

All praises, Exaltation, Adulations, Idolization and devotions are due to ALMIGHTY ALLAH (SWT) alone for the successful completion of this level of achievement. May HE continue to shower on us His unflinching support, protection and guidance (Amen).

I must immediately show my immeasurable appreciation to my supervisors, Dr. A. Isah who apart from sparing his precious time amidst busy schedules to guide and supervised my work, played a vital role of mentorship, encouragement and above all imparting his wealth of knowledge on me as a lecturer. In the same line, I must also appreciate my minor supervisor, Dr. H. G Dikko whose inspiration, encouragement and supervisory roles cannot be quantify. The efforts and mentorship roles of Head of Department, Dr. H. G Dikko, Postgraduate Coordinator Dr. A. Yahaya, and other lecturers of the Department remain commendable.

I owe a great deal of appreciation and acknowledgement for the co-operation, support and prayers of my brothers; Moh'd, Usman, Shehu and Yaya; Sisters; Amina, Fatima, Hawau and a lot others that space will not allowed.

My special gratitude goes to my ever co-operative and supportive wife Mallama Khadijat and my Children Khadijat, Hauwa, Fatima, Ali-imran, Ahmed, Maimunat, Suleiman(Ama'ali) and Aishat for their support and patients during the program.

I am also indebted to all my course mates, staff and members of Niger State Polytechnic for their co-operation and support. I also appreciate the cooperation shown by Management and staff of University Teaching Hospital, Gwagwalada and Federal Medical Centre, Keffi Nassarawa State during data collection.

ABSTRACT

This work applies Discriminant Analysis and Logistic Regression models to predict the prevalence of Broncho-Pneumonia status (BPn) in infants. The data used in this study were collected from two tertiary health institutions in North Central Zone; University Teaching Hospital (UTH), Abuja and Federal Medical Centre (FMC), Keffi, Nassarawa State. Five predictors which are well-recognized for characterizing broncho-pneumonia in infants (baby's weight at birth, baby's weight 4week after, sex, mother's age and mother's occupation) were considered. One hundred and eighty (180) and two hundred and fifty three (253) infants with Low Birth Weight (LBW) were randomly sampled using simple random sampling technique from UTH, Abuja and FMC, Keffi respectively to build up the models. Both Linear Discriminant and Logistic Regression Models were fitted to the data for the two groups, and the best model was identified. Ten different samples of size 10 each were randomly taken from the dataset using SPSS package. The new datasets were used to validate the two models. It was observed that Discriminant Model is better used in the zone than Logistic Regression Model. We also find out that baby's weight at birth is best at discriminating between the two groups, since it has the least value of Wilk's Lambda compare to other predictor variables.

Table of Contents

TITLE PAGE…….. i
DECLARATION……ii
CERTIFICATION……..iii
DEDICATION……iv
ACKNOWLEDGEMENT……...v
ABSTRACT……...vi
TABLE OF CONTENTS……..vii
List of Tables……..ix
Acronyms……x
CHAPTER ONE……..1
INTRODUCTION……...1
1.1 Background of the study……..……1
1.2 Broncho-Pneumonia……5
1.3 Low Birth Weight……6
1.4 Statement of the Problem……8
1.5 Aim and Objectives of the Study……9
1.6 Significance of the Study……10
1.7 Scope of the Study……10
1.8 Definition of Terms……10
CHAPTER TWO……..….11
LITERATURE REVIEW……..11
2.1 Introduction……11
2.2 General Review……..11
2.3 Gap identified in literature……22
CHAPTER THREE……...23
RESEARCH METHODOLOGY…….……..23
3.1 Introduction…….…..23
3.2 Assumptions of Discriminant and Logistic regression analyses……..…….….23
3.3 Study Design…….…24
3.4 Population and Sample of Study……25
3.5 Data Collection……..25
3.6 Concept of Discriminant Analysis……26
3.7 Method used in this study……..27

3.8 Test For Significance of Canonical Correlation (Wilk's Lambda)…..……..……...30
3.9 Box's M - Test for the Equality of Covariance Matrices…… ……..……..32
3.10 Logistic Regression Model……..33
3.11 Chi-square test……35
3.12 Omnibus Chi-Square Test……...36
3.13 Multicollinearity……..37
CHAPTER FOUR……39
ANALYSIS, RESULTS AND DISCUSSIONS……39
4.1 Introduction……39
4.2 Descriptive Statistics……..39
4.3 Checking the Assumptions of Discriminant Analysis……...…39
4.4 Linear Discriminant Function 42
4.5 The Logistic Regression Model……45
4.6 Goodness of fit and classification power……...47
4.7 Discussion of Results……48
4.8 Major Findings……...49
CHAPTER FIVE……...51
SUMMARY, CONCLUSION AND RECOMMENDATIONS……51
5.1 Summary……51
5.2 Conclusion……52
5.3 Recommendations……....52
5.4 Contribution to knowledge……53
References……..54
Appendix I……57

List of Tables

Table 4.1: Descriptive Statistics on Continuos Variable……..……...……....38**Error! Bookmark not defined.**

Table 4.2: Categorical Variables Coding……...……..38

Table 4.3: Descriptive Statistics on the Two Groups……...…….....…..39**Error! Bookmark not defined.**

Table 4.4: Pooled Correlation Matrices……...39

Table 4.5: Pooled Covariance Matrices…….…..40

Table 4.6: Test of Equality of Group Means……..…….40

Table 4.7: Multicollinearity Results……40

Table 4.8: Test Results of Box's M……41

Table 4.9: Fisher's Classification Function Coefficients……41

Table 4.10: Canonical Discriminant Function Coefficients……..……..…….42

Table 4.11: Functions at Group Centroids……...……..43

Table 4.12: Classification Results……..43

Table 4.13: Variables in the Equation for the Sample Data……...….....44

Table 4.14: Omnibus Tests of Model Coefficients…….……..……45

Table 4.15: Hosmer and Lemeshow Test……..……45

Table 4.16: Samples Selected Using Discriminant Model…….……...46

Table 4.17: Samples Selected Using Logistic Regression Model…….……46

Acronyms

LRA	Logistic Regression Analysis
BPn	Broncho – pneumonia
BPD	Broncho – pulmonary Dysplasia
LBW	Low Birth Weight
VLBW	Very Low Birth Weight
E L B W	Extremely Low Birth Weight
LDA	Linear Discriminant Analysis
OR	Odds Ratio
NBI	Narrow Band Imagine
ASD	Angiogenic Squamous Dysplasia
SVM	Support Vector Machine
TLC	Total Lymphocyte Count
PCA	Principal Component Analysis
VAT	Visceral Adipose Tissue
NLR	Neutrophil – to – Lymphocyte Ratio
UTH	University Teaching Hospital
FMC	Federal Medical Centre

CHAPTER ONE

INTRODUCTION

1.1 Background of the study

Discriminant analysis is a procedure that can be used to build Discriminant functions which are linear functions of p-variables that can be used to describe or elucidate the differences among two or more groups. The goals of discriminant analysis include identifying the relative contribution of the p variables to separation of the groups and finding the optimal plane on which the points can be projected to best illustrate the configuration of the groups. Another use of discriminant analysis is the prediction or allocation of observations to groups, in which linear functions of the variables are employed to assign an individual sampling unit to one of the groups. The measured values in the observation vector for an individual or object are evaluated by the classification function to find the particular group to which the individual most likely belongs.

Interest in human development before birth is widely spread because of the interest in knowing more about our beginning and the desire to improve the quality of life. The intricate process by which a baby develops from a single cell is miraculous and few events are more exciting than a mother's viewing of her embryo during an Ultrasound examination. Human development is a continuous process that begins when an Oocyt (ovum) from a female is fertilized by a sperm (spermatozoa) from the male. By accepting the shelter of uterus, the foetus also takes the risk of disease or malnutrition and of biochemical immunological and hormonal adjustment.

Until the beginning of the Nineteenth Century, far more attention was paid to the collection and presentation of data than to their interpretation. Large volume of data were usually collected and frequently misinterpreted if indeed interpretation was attempted. However, since that time, the importance of scientific approach in the interpretation of data has been realized and great steps have been achieved in the development of appropriate methods.

In modern days, statistics has played a significant role in Biological, Pharmaceutical and Medical Sciences (Cornfield,1952). The application of multivariate statistical techniques to biological and medical data has dominated the areas of evidence-based medicine. Multivariate methods are relevant in virtually every branch of applied medicine, pharmacy and public health. They come into play either when we have a medical theory to test or when we have a relationship in mind that has some importance for medical decision or policy analysis in public health. Multivariate methods are also used in other disciplines.

Multivariate methods are prominently used on data to test a theory or to estimate a relationship in different disciplines. In some cases, especially those that involve the testing of medical theories, a formal multivariate model is constructed. The model consists of multivariate technique that describes various relationships. In most cases, the model is used to make predictions in either the testing of a medical theory or the study of a policy impact in pharmacy and public health.

Kirkwood and Stern (2008) defined discriminant analysis and classification as the multivariate techniques concerned with separating sets of objects or observations and

with allocating new objects or observations to previously defined groups. As a separation procedure, it is often employed on a one-time basis in order to investigate observed differences when causal relationships are not well understood. The immediate goals of discriminant analysis and classification are to describe the differential features of objects so as to find Discriminant function whose numerical values are such that the collections are separated as much as possible and to sort new objects or observations into two or more classes or groups.

In clinical situations, the status of a patient is assessed by the presence or absence of a disease. There are many factors to consider which may or may not correlate with the incidence of the disease. There has been numerous retrospective medical research studies published each year that review past medical records and charts of former patients to help determine some of the risk factors (or causing agents) of diseases that are of interest. Finding the risk factors and the potential risk factors can help to prevent the development of the disease. For all of the diseases, most of the risk factors considered are categorical variables i.e. variables taking on two or more possible values. (Hosmer and Lemeshow, 1989), two prominent statisticians, stated that 'the logistic regression model has become the standard method of analysis in this situation.'

Logistic regression analysis is also called "Binary Logistic Regression Analysis", "Multinomial Logistic Regression Analysis" and Ordinal Logistic Regression Analysis", depending on the scale type where the dependent variable is measured and the number of categories of the dependent variable. Logistic regression is divided into two; Univariate Logistic Regression and Multivariate Logistic Regression (Stephenson, 2006).

Like any other model building technique, the goal of the logistic regression analysis is “to find the best fitting and most parsimonious, yet biologically reasonable model to describe the relationship between an outcome (dependent or response) variable and a set of independent (predictor or explanatory) variables” (Hosmer and Lemeshow 1989). This statement motivates the purpose of this study to identify risk factors for low birth weight (LBW) in newborn infants using the statistical tools of logistic regression analysis.

The use of logistic regression dates back to 1845. It first appeared during the mathematical studies for the population growth at that time. The term logistic regression analysis comes from logit transformation, which is applied to the dependent variable. This case, at the same time, causes certain differences both in estimation and interpretation (Anderson, 2008)

In many application areas, such as epidemiological and biomedical studies, where outcomes may be occurrence or nonoccurrence, mortality (dead or alive), and so forth, logistic regression is the standard approach for the analysis of binary and categorical outcome data.

Logistic Regression Analysis (LRA) extends the techniques of multiple regression analysis to research situations in which the outcome variable is categorical. In practice, situation involving categorical outcomes are quite common. In the setting of evaluating an educational program, for example, predictions may be made for the dichotomous of success/failure or improved/not-improved. Similarly, in a medical setting, an outcome might be presence/absence of disease. The focus of this study is on the situations in

which the outcome variable is dichotomous, although extension of the techniques of LRA to outcomes with three or more categories (e.g improved, same, or worse) is possible.

In this Twenty First Century, statistics play an important role in many simulations, modeling and decision-making processes. This implies the need for statistical research in every facet of medicine; especially the evidence-based medicine. Anderson (2008) mentioned that the critical factor that separates statistical research from other ways of knowing the medical world is that statistical research is purely scientific in nature. In this sense, Science refers to both a system for producing medical knowledge and the knowledge produced. Also Science is a combination of an orientation towards a set of procedures, techniques, knowledge and instruments for gaining knowledge.

1.2 Broncho-Pneumonia

Pneumonia is an illness, usually caused by infection, in which the lungs become inflamed and congested, reducing oxygen exchange and leading to cough and breathlessness. It affects individuals of all ages but occurs most frequently in children and the elderly.. Historically, in developed countries, deaths from pneumonia have been reduced by improvements in living conditions, air quality, and nutrition. In developing world today, many deaths from pneumonia are also preventable by immunization or access to simple, effective treatments (Anthony, 2010).

Pneumonia can be caused by bacteria, viruses and fungi. *Streptococcus* pneumonia and *Haemophilus* influenza type b (Hib) are the most common causes of bacterial pneumonia while respiratory syncytial virus is the most common viral cause of pneumonia. A yeast-

like fungus- *Pneumocystis jiroveci* is often responsible for pneumonia deaths in HIV-infected infants.

Broncho-pneumonia or bronchial-pneumonial or bronchogenic pneumonia is a type of pneumonia characterized by multiple foci of isolated, acute consolidation, affecting one or more pulmonary lobules. It is one of two types of bacterial pneumonia as classified by gross anatomic distribution of consolidation (solidification). The other being lobar pneumonia. Broncho-Pneumonia is less likely than lobar pneumonia to be associated with streptococcus.

The broncho-pneumonia pattern has been associated with hospital acquired pneumonia, and with specific organisms' *staphylococcus aurous, klebsiella coli and pseudomonas.* In bacterial pneumonia, invasion of the lung parenchyma by bacteria produces an inflammatory immune response. This response leads to a filling of the alveolar sacs with exudates. The loss of air space and its replacement with fluid is called consolidation.

Broncho-Pulmonary Dysplasia (BPD) is a chronic type of lung disease prevalent among infants. This disease if present in a pregnant mother leads to low birth weight of infants at birth. It is a serious lung condition that affects infants. It mostly affects premature who need oxygen therapy (oxygen given through nasal prongs, a mask or a breathing tube). Most infants who develop BPD are born more than ten weeks before their due dates and weigh less than 2 pounds (about 1kg) at birth, and have breathing problems (Jobe, 2001).

1.3 Low Birth Weight

Low Birth Weight (LBW) is described as a birth weight of a live born infant of less than 2.5kg regardless of gestational age. Subcategories include; Very Low Birth Weight

(VLBW) which is less than 1.5kg and Extremely Low Birth Weight (ELBW) which is less than 1.0kg. Normal Weight at term of delivery is 2.5kg – 4.2kg. Most normal babies weigh above 2.5kg by 37 weeks of gestation. Intrauterine growth restriction refers to delayed growth within the uterus, which then leads to low birth weight. Some babies are just small and happen to weigh less than 2.5kg at birth, just like some adults are smaller than others. Though this is considered low birth weight, in these cases, it is not abnormal nor a cause for concern.

Using the discriminant and logistic regression is of interest to this study. We will use a sample of not less than 400 of infants drawn from an underlying population of children with low birth weight (kg). These children were confined to a neonatal intensive care unit, they required incubation during the first 12 hours of life, and they survived for at least 28 days and their weights measured four weeks later. Healthy infants are denoted by (0) while, Infected infants by (1). Factors that contribute to the risk of Broncho-Pulmonary Dysplasia (BPD) include high blood pressure in mothers, hypercholesterolemia in mothers and family history of tobacco smoking, among others (Eneh, 2011).

In strict terms, the application of statistical techniques to biological and medical data is called Biostatistics. Generally speaking, biostatistical methods are relevant in virtually every branch of applied medicine, pharmacy, nutrition and public health. They come into play either when we have a medical theory to test or when we have a relationship in mind that has some importance for medical decision or policy analysis in public health. Biostatistical methods in medicine are more or less empirical analysis using data to test a theory or to estimate a relationship in medicine, pharmacy, public health and other areas.

In some cases, especially those that involve the testing of medical theories, a formal statistical model is constructed. The model consists of statistical equations that describe various relationships. A biostatistical analysis begins by specifying a statistical model. Once a statistical model has been specified, various hypotheses of interest can be stated and empirically tested in terms of the unknown biological or medical parameters. An empirical analysis requires data which are used to estimate model parameters and to formally test hypotheses of interest. In most cases, the model is used to make predictions in either the testing of a medical theory or the study of a policy's impact in pharmacy and public health (Rencher, 2002).

Some statistical models in medical research may contain dichotomous factor; in form of a person is male or female; a person does or does not have a disease in question, to mention but a few. In all of these examples, the relevant information can be captured by defining a classification discriminant model.

1.4 Statement of the Problem

Birth weight less than 2.5kg is categorized as Low Birth Weight (LBW). It remains a significant public health problem in both developed and developing countries. These infants with LBW encounter greater neonatal morbidity and mortality and significantly higher rates of physical and mental handicaps later in life (Pope, 2010). Taking the infants population globally, the proportion of babies with a LBW is an indicator of a multifaceted public-health problem that includes the sex of an infant as well as the birth weight and weight four weeks after birth. Also, the mother's age and mother's occupation are important variables that could predict the LBW of the infant considered in the study. Therefore, the main problem which comes up in this particular study is how to construct

linear discriminant and logistic regression models that are capable of predicting the Broncho-Pneumonia(BPn) status of the infant using mother's age, mother's occupation, baby's sex, baby's weight at birth and baby's weight four weeks after birth as predictor variables. The core research issue is therefore to explore the predictive powers of both the Linear Discriminant Model and Logistic Regression Model as regards statistical modeling.

However, since the models comprise discrimination and classification, it is in the interest of the researcher to classify some infants as affected and unaffected patients of Broncho Pneumonia using Linear Discriminant and Logistic Regression Models. Hence, a suitable prediction model will be developed to satisfy the best methods of validation as well as diagnostics of statistical decisions. Moreover, the Linear Discriminant Model and Logistic Regression Model could be used to predict the BPn status of new cases of infants.

1.5 Aim and Objectives of the Study

The aim of this study is to investigate the broncho-pneumonia status in infants using linear discriminant and logistic regression models, and this will be achieved through the following objectives; by

i. constructing a linear discriminant and logistic regression models that are capable for predicting the Broncho-Pneumonia status in infants;
ii. predicting the Broncho-Pneumonia status of some infants (random selected cases) using the developed models;
iii. comparing the predictive powers of the two models for Broncho-Pneumonia;

iv. determining the predictor that has the most discriminating ability among the predictors.

1.6 Significance of the Study

The Linear Discriminant and Logistic Regression Models built in this study will give effective guide in evidence-based medicine. That is, to achieve useful projections of the BPn status of infants so as to isolate factors responsible for such. On the other hand, the study will assist medical researchers to ascertain the prevalence of BPn using the developed models.

1.7 Scope of the Study

The purpose of this study is to develop the models based on five predictors; Weight at birth, Weight four weeks after birth, Sex, Mother's age and Mother's occupation. The five independent variables are incorporated in both the linear discriminant and logistic regression models as the most relevant factors considered and captured by the study.

The study will also focus on North Central Zone, out of six geo-political zones of the Federation. The sample taken from two health tertiary institutions would be used for the analysis on prevalence of BPn among infants which lead to Low Birth Weight (LBW) in infants.

1.8 Definition of Terms

Broncho-Pneumonia (BPn): Is a type of pneumonia characterized by multiple foci of isolated, acute consolidation, affecting one or more pulmonary lobules.

Broncho-Pulmonary Dysplasia (BPD): Is a chronic type of lung disease prevalent among infants, this disease if present in a pregnant mother leads to low birth weight of infants at birth.

Low Birth Weight (LBW): Is described as a birth weight of a live born infant of less than 2,500g (5 pounds 8 ounces) regardless of gestational age.

Discriminant Function: Is a multivariate technique concerned with separating distinct sets of objects (or observations) and it gives the rule for allocating (observations) to previously defined groups.

Logistic regression or Logit deals with the cases where the response variable consists of two or more categorical values.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter is on the collection of relevant research works that provide a basis for the present study. It gives an overview of the prevailing theories and hypotheses and methodologies on the subject of study. A critical literature review shows how prevailing ideas fit into a particular study, and how the work agrees or differs from them.

2.2 General Review

Beki (2012) used discriminant analysis and binary logistic regression for tracking the incidence of Broncho-Pulmonary Dysplasia among infants. The researcher used three possible predictor variables i.e. weight at birth, weight four weeks later and gender and

built a discriminant model that is capable of tracking Broncho-Pulmonary Dysplasia (BPD) infants. The Discriminant models for the two locations were;

$Y=-2.860+0.035X_1-0.022X_2-0.658X_3$ and $Y=-3.539+0.001X_1-0.003X_2-0.795X_3$

(2.1)

The Logistic regression models for the two locations were given as;

$$p=\frac{e^{5.835-0.066x_1+0.042x_2-0.773x_3}}{1+e^{5.835-0.066x_1+0.042x_2-0.773x_3}} \text{ and } p=\frac{e^{9.113+0.001x_1-0.006x_2-1.260x_3}}{1+e^{9.113+0.001x_1-0.006x_2-1.260x_3}}$$

(2.2)

Were X_1 represent weight at birth (g), X_2 denote weight four weeks after birth (g) and X_3 represent sex.

The study predicted the BPD status of five new infants using the discriminant model in which all the five new cases were correctly predicted. The discriminant model built had a perfect classification of five new cases in Kaduna while it has misclassification of one of five new cases in Sokoto. Conversely, the study predicted the BPD status of five new infants using logistic model in which all the five new cases were correctly predicted or classified. Hence, the logistic model built has a classification of five new cases in Sokoto while it misclassified two of five new cases in Kaduna.

Danbaba *et al* (2013), carried out a research on low birth weight using logistic regression analysis to determine the prevalence of Low Birth Weight (LBW) and some of its risk factors in maternity hospitals in Wushishi Local Government of Niger State. Data from a sample of 200 live births were collected in the hospital from June – September 2011. The

data were collected by obtaining the mother's age at birth, mother's weight at birth, mother's education level, mother's occupation, gestational age, birth interval, twin or singleton birth and parity. The study fitted the logistic regression model to the data. The analysis of variance and chi-square tests were used to know the variables or factors that have statistically significant effect on birth weight at the 95.0% confidence level. The Odds Ratio (OR) of the risk factors of LBW was found using a multivariate logistic regression. They established the fundamental model for multiple regression analysis, with the assumption that the outcome variable was a linear combination of a set of predictors. For response variable (y), and a set of n-predictor variables, x_1, $x_2 \ldots x_n$, used the model as;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \qquad (2.3)$$

where β_o is the expected value of Y when the x's are set to 0, β_i is the regression coefficient for each corresponding predictor variables, Xi and ε is the error of the prediction.

The analysis showed that, there was no significant difference in prevalence between boys and girls (14.9% versus 13.9%) i.e. p=0.578.

Carlos *et al,* (2007) conducted a research which involves the construction of model for the prediction of Broncho-Pulmonary dysplasia for seven-day old infants. The objective was to develop a predictive model capable of identifying which premature infants have the greatest probability of presenting BPD based on assessment at the end of the first week of life. In this study, the data were collected retrospectively from January 1998 through December 2001 and prospectively from December 2001 through July 2003. The researcher's target population was infants born with less than three weeks after birth

weight of less than 1500g. The logistic regression was employed for the data analysis. Four variables maintained a significant relationship with the outcome and were used to construct the formula to calculate the probability of BPD [Probability (%) = $(e^{logit})/1 + (e^{logit})$], where **logit** is the β value for each variable multiplied by the corresponding variable minus the constant $(logit = \beta x\ variable\,(0/1) + \beta x\, variable\,(0/1) - constant)$ and **e** is the base for natural logarithms with a value of 2.71828.

In order to validate the model, it was applied to different populations. The researcher used a sample of 247 children, of whom 68 developed BPD which was classified to 35 (51.4%) mild BPD, 20 (29.4%) moderate BPD and 8 (11.9%) severe BPD. The results showed that the model had a 93.7% probability of correctly classifying the children. In summary, the researchers concluded that at the end of the first week of life, the predictive model developed was capable of identifying newborn infants at increased risk of developing BPD with high degree of sensitivity.

Cornfield (2010), reported the relationship between risk of coronary heart disease and two variables; serum cholesterol level and systolic blood pressure, in subjects in a long–term follow–up (prospective) study. The endpoint of interest in his study was whether or not the subject developed coronary heart disease (myocardial infarction or angina pectoris).

Uthman (2008) carried out a study on the effect of birth weight on infant mortality found that children born with low birth weight are more likely to die during the first year of life compared to children born with normal weight, independent of child's sex, birth order,

(pregnancy care and delivery care), maternal education and nutritional status, household access to clean water and sanitation.

Denan *el al, (*2002) observed that there are large variations in the incidence and severity of this disease. According to the National Institutes of Health of USA (NICHD) consensus, the most recent report of the incidence of BPD in Latin America comes from the Neonatal Group study, very-low-birth-weight (VLBW) /I,k,min some Asian part of the world.

Tapia *el al,* (2006), observed that the BPD is a chronic pulmonary disease which affects premature infants and contributes to their morbidity and mortality. Despite substantial changes in incidence, risk factors and severity after the introduction of new therapies and mechanical ventilation (MV) techniques, BPD remains common.

Vitmalkumar *et al,* (2011) conducted a study to determine the prevalent risk factors of Nephropathy in type-2 diabetic patients. A tertiary hospital was used for the study aimed to build a binary logistic model for predicting Nephropathy status among type-2 diabetic patients using age, sex, socio-economic status, and duration of Nephropathy history as covariates. Furthermore, the researcher used a random sample of 200 patients suffering from type-2 diabetes where data on some risk factors like age, sex, socio-economic status and duration of Nephropathy history were collected using questionnaire.

In their summary of findings, it was discovered that as the duration of type-2 diabetes increases, the incidence of Nephropathy also increases significantly. Hence, all the type-2 diabetic patients, especially those with increased duration should be screened for Nephropathy and be made aware of the complications.

Vishwa *et al* (2015) stated that discriminant analysis and classification are multivariate techniques concern with separating district sets of objects (or observations) and with allocating new objects (or observations) to previously defined groups. Discriminant analysis is rather exploratory in nature. As a classificatory procedure, it is often employed on a one-time basis in order to investigate observed differences when casual relationships are not well understood. Classification procedures are less exploratory in the sense that they lead to well-defined rules, which can be used for assigning new objects. It is possible to have classifications into two or more multivariate normal populations, but the study shall be restricted to classifications into two normal populations denoted by π_1 and π_2.

Adler and Dondlon (2010) in their study concluded that morphometric crown traits in the deciduous dentition can be used to classify sex of juvenile skeletons (11 months to 12 years) of European descent from linear discriminant functional analysis with accuracy between 70.2% and 74.8%.

Fernandez *et al*, (2010) used a discriminant analysis to investigate whether FT – Raman spectroscopy as spectroscopic fingerprint techniques combined with some chemometric tools can be used as a rapid and reliable method for the discrimination of honey according to their sources. In their study, they used developed models exploratory techniques as the fishers criterion, supervised methods as Partial Least Squares -Discount Analysis (PLS-DA) or Support Vector Machine (SVM) which all show a correct classification ratio between 85% and 90% of average showing Raman spectroscopy combined to chemometric treatment is a promising way for rapid and non expensive discrimination of honey according to their regions.

Onyiriuka (2010), discovered that maternal parity have a significant influence on the incidence of delivery of LBW infants in twin gestations. As in previous studies, the incidence was higher in primiparous compared with multiparous counterparts, suggesting that the uteri of multiparous women are more efficient in nurturing and promoting intrauterine growth of twins; accounting for the relatively lower incidence of LBW of twin infants.

Rees *et al*, (2005) found lifestyle behaviours such as inadequate nutrition, smoking, mothers themselves who were low birth weight, low pregnancy weight gain, increasing maternal stress and/or depression, domestic violence and maternal regret and/or rejection of pregnancy to be significant factors. Other socio-demographic factors according to Lasker *et al*., (2005) were low maternal age (under 18), high maternal age (over 35), low educational level, poverty, ethnicity and late or no antenatal care.

Yu-Shun *et al*, (2011) observed that breast cancer is the most common type of cancer in women, while the mortality rate of breast cancer of females over 40 years old is extremely high. If detected early, it can be treated early, and the mortality rate of breast cancer can be reduced. Therefore, the image processing technologies has been adapted to automatically breast images, select suspicious regions, and provide alerts to assist in doctors' diagnosis, reduce misdiagnosis rates due to fatigue of doctors, and improve diagnostic accuracy. In order to assist physicians in clinical diagnosis, a set of breast cancer detection algorithm was designed in this paper through the Fuzzy theory and Linear Discriminant Analysis (LDA). The experimental results show that the accuracy of the current detection methods can be improved to generate a breast cancer detection

system used in the auxiliary medical diagnosis system, effectively reduce physicians' determination time, and improve accuracy.

Erimafa *et al*, (2009) used discriminant analysis to predict the class of degree obtainable in a university system. In this study, it was clearly stated that, the conditions for predictive discriminant analysis were obtained, and the analysis yielded a linear discriminant function which successfully classified or predicted 87.5 percent of the graduating students' class of degrees. The function had a hit ratio of 88.2 percent when generalized.

Vajpayee *et al*, (2005) conducted a research that tried to classify HIV-seropositive Antiretroviral Treatment (ART)-naïve Indian individuals on Centers For Disease Control (CDC) criteria of clinical symptoms and Cluster Differentiation 4 (CD4)% and CD4 counts. The optimum cut-off values of CD4 counts and CD4% obtained were compared with the CDC recommended values. The study also aimed to investigate the CDC staging of HIV-1 patients, on the basis of CD4 counts and CD4%, and the clinical implications in terms of HIV treatment and prophylaxis of these two staging criteria in an Indian population.

Mobili *et al*, (2010) used both principal component Analysis (PCA) and Partial Least Square– Discriminant Analysis (PLS-DA) in the analysis and interpretation of the Raman spectra collected from microorganism of different species recorded in the spectral range of 2000 to 200 cm^{-1}. To develop a classification rule, the researcher used PLS-DA in a Leave-One-Out Cross Validation (LOOCV) method for the calibration and validation of a classification model. It was asserted that, results obtained showed an acceptable

classification among the strains under study; thereby, suggested it to be useful tools for the classification and discrimination of similar samples.

Hamid *et al,* (2011) used a supervised micro-calcifications based on Fisher's Linear discriminant analysis by methodological approach of breast density which allow them to identify micro-calcifications even in difficult cases (i.e when there is not high contrast between the micro- calcification and the sound.

Thomas *et al* (2007) studied Prevention and treatment of broncho-pulmonary dysplasia, current status and future prospects. The review summarizes the existing evidence for potential non-ventilator-dependent strategies to prevent or ameliorate broncho-pulmonary dysplasia (BPD). Oxygen plays an important pathogenetic and therapeutic role for BPD. Targeting infants with lower saturation of peripheral oxygen levels than traditionally used seems justified. Inhaled nitric oxide has not proven effective on pulmonary outcome in extremely low birth weight infants. Diuretics can ameliorate lung function transiently. High intramuscular doses of vitamin A reduce the risk of BPD. Prophylactic application of natural surfactant may also confer benefits. Recently, early administration of caffeine has been shown to decrease risk of BPD. However, assessment of long-term effects is needed before routine use can be recommended. Owing to potential short-term and long-term effects, postnatal corticosteroids should be restricted to the most severe manifestations of BPD. Superoxide dismutase and 1-proteinase inhibitor have not shown efficacy in preventing BPD. The potential role of anti-inflammatory therapies like Clara Cell 10 kDa protein has yet to be defined.

Pyng *et al,* (2009) carried out a research on Color Fluorescence Ratio for Detection of Bronchial Dysplasia and Carcinoma. The aim of the study was to determine if color fluorescence ratio (i.e. red-to-green ratio R/G ratio) added to autofluorescence bronchoscopy could provide an objective means to guide biopsy. Subjects at risk for lung cancer were recruited at two centers: VU University Medical Centre (Amsterdam) and BC Cancer Agency (Canada). R/G ratio for each site appearing normal or abnormal was measured before biopsy. R/G ratios were correlated with pathology, and a receiver operating characteristic curve of R/G ratio for high-grade and moderate dysplasia was done. Following analysis of the training data set obtained from two centers, a prospective validation study was done. In their study, 3362 adequate biopsies from 738 subjects with their corresponding R/G ratios were analyzed. In conclusion, the Color fluorescence ratio can objectively guide the bronchoscopist in selecting sites for biopsy with good pathologic correlation.

Shibuya *et al,* (2003) investigated the use of high magnification bronchovideoscopy combined with narrow band imaging (NBI) for the detailed examination of Angiogenic Squamous Dysplasia (ASD). This was carried out in relation to bronchial vascular patterns with abnormal mucosal fluorescence in heavy smokers at high risk for lung cancer. Forty eight patients with sputum cytology specimens suspicious or positive for malignancy were entered into the study. It was concluded that high magnification bronchovideoscopy combined with NBI was useful in the detection of capillary blood vessels in ASD lesions at sites of abnormal fluorescence. This may enable the discrimination between ASD and another pre-invasive bronchial lesion.

Emin *et al*, (2014) conducted a research to find out the relationship between diabetic nephropathy and Visceral Adipose Tissue (VAT). The Neutrophil-to-Lymphocyte Ratio (NLR) and Platelet-To-Lymphocyte Ratio (PLR) are simple, inexpensive, and useful markers to determine inflammation. The aim of research was to investigate the association between diabetic nephropathy, NLR, and PLR as inflammatory markers. The methods used were a cross-sectional study involving 200 diabetic patients. The patients were separated into three groups according to their albuminuria levels. A correlation analysis showed that albuminuria was significantly correlated with EAT, disease duration, creatinine, Estimated Glomeruler Filtration Rate (eGFR), PLR, and NLR levels. Additionally, in a binary logistic regression analysis, EAT, NLR, and PLR were found to be independently associated with albuminuria. And, concluded as determining various inflammatory cytokines and measuring abdominal, VAT in diabetic patients is complex and expensive. Simply measuring EAT and calculating NLR and PLR can predict inflammation and albuminuria in patients with diabetes.

Suzanne *et al*, (1997) carried out a study to investigate the mortality and short term neurologic outcome of Extremely Low Birth Weight (ELBW) infants requiring rescue High Frequency Ventilation (HFV). Methods used by the researchers entailed retrospective review of all infants < 1000g cared for at a single level III Neonatal Intensive Care Unit (NICU) between 7/93 to 7/96, n =152. Infants receiving rescue HFV were compared to those not receiving HFV by *Student's* t-test and chi-square. Forward stepwise discriminant analysis was performed to account for potential confounding variables. All data are expressed as mean ± SD. The result of the analysis was given as the overall 144/152 (95%) required mechanical ventilation. In conclusion, population of

ELBW infants, the need for rescue HFV is associated with 58% mortality. However, infants surviving after rescue HFV have the same low incidence of Periventricular Leukomalacia (PVL) and severe Intraventricular Hemorrhage (IVH) as infants surviving after conventional mechanical ventilation.

2.3 Gap identified in literature

The present work is different from other previous works such as Beki (2012) in which a combination of discriminant and binary logistic regression were used and considered only three variables (Sex, Weight at birth and Weight after four weeks) as against the present study in which five predictos were considered. In Danbaba *et al*, (2013) the study investigated some variables considered from mothers' aspect and use logistic regression. This study used both Mother and Baby's independent variables capable of having Broncho-pneumonia in infants. The variables were; Baby's weight at birth, Baby's weight four weeks after, Baby's gender, Mother's age and Mother's occupation.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

This chapter focuses on the methods and data collection from the study area. It is necessary to critically study our methods and procedures as a precondition for achieving the desired goals. The research explored the prediction powers of the discriminant function and the logistic model as regards to the proper applications of biomedical modeling and to compare same for classifying the Broncho-Pneumonia (BPn) status of infants. The variables considered in this research; baby's weight at birth, baby's weight 4 weeks after, baby's sex, mother's age and mother's occupation are medically adequate to elucidate the difference between a normal and Broncho-Pneumonia (BPn) patient.

In this study, the researcher shall particularly build discriminant models with prior information for predicting the Broncho-Pneumonia (BPn) status of infants using the five variables.

3.2 Assumptions of Discriminant and Logistic regression analyses

i. The predictors are not correlated with one another, i.e. there is no multicollinearity. Correlation between two predictors is constant across groups.
ii. The variance-covariance matrixes of all the independent variables are homogenic.
iii. Independent variable with the best discriminating ability between the two groups
iv. The independent variables do not need to be multivariate normal
v. The dependent variable is binary.
vi. The predicted values are probabilities and are therefore restricted to (0,1)

3.3 Study Design

To achieve the research objectives, this work is a combination, both in purpose and in design of discrimination and classification analysis. It is discrimination as it seeks to draw a line between the BPn status of infants using their weight at birth, weight four weeks after birth, sex, mother's age and mother's occupation. On the other hand, in the classification design, the researcher is not interested in a mere collection of facts but models would be used to classify the BPn status of an infant whose BPn status is not known earlier.

However, in discrimination and classification designs, the major statistical components form the basis of the research design which includes both the sampling plan and the modeling procedures. The sampling plan is the methodology used for selecting the sample from the population. The modeling procedure is the algorithms or formulae used for obtaining models of population values from the sample data and for estimating the reliability of these models.

The Sex and Mother's occupation were categorized as follows; in sex variable, male is coded as 0 and female as 1. In Mother's occupation, House wife is coded 0, Civil servant as 1 and 2 for Business mother.

The five selected predictors mentioned above are capable of characterizing Broncho-Pneumonia in infants. The experience and records from medical section (Danbaba, 2013), these variables are believed to vary significantly between unaffected and affected (BPn) infants. These variables were abbreviated as follows;

x_{bw_1}= $baby's\ weight\ at\ birth\ (kg)$

x_{bw_2}= $baby's\ weight\ 4\ weeks\ after\ (kg)$

x_s= $sex, where\ x_s\ is\ coded\ as\ 1\ for\ male\ and\ 2\ for\ female$

$x_{ma} = mother's\ age$
x_{mo}= $mother's\ occupation\ coded\ as\ 0\ for\ house\ wife, 1\ for\ civil\ servant\ and\ 2\ for\ business\ mother$

3.4 Population and Sample of Study

This study was conducted on the available dataset from the two tertiary health institutions. The population is infinite as infants are given birth to on daily basis. Any baby with low birth weight is a potential BPD which is mainly caused through pneumonia in infant; hence, the population size cannot be specified at any point in time.

A total of 433 infants' files were inspected from the two institutions. In UTH Abuja, 180 infants were treated; 80 out of the 180 infants were affected with pneumonia, and in FMC, Keffi, 96 of 253 available infants were affected.

3.5 Data Collection

Data were carefully and technically extracted directly from the individual patient's medical folder from the two randomly selected tertiary health institutions within the zone, namely; University Teaching Hospital, Abuja and Federal Medical Center, Keffi, Nasarawa State. The baby's weight at birth (kg), baby's weight four weeks after (kg), baby's gender, Mother's Age and Mother's Occupation data were collected and tabulated as independent variables.

It is necessary to outline the framework of the major components involved in the sampling design and data collection procedures adopted in this research. The North Central Zone consists of six states and FCT Abuja. The States are Niger, Kogi, Kwara,

Nasarawa, Benue and Plateau. Each of these has at least one tertiary health institution. The simple random sampling (SRS) scheme with size n=2 was used for data collection.

3.6 Concept of Discriminant Analysis

Assuming there are two multivariate normal populations with equal variance-covariance matrices, $N(\mu_{1,}\sum)$and $N(\mu_{2,}\sum)$; where μ_i is the vector of means of the ith population and $\sum$ is the variance-covariance matrices of the two populations. The probability density function of ith population is given as follow;

$$P_i(X) = \frac{1}{(2\pi)^{\frac{p}{2}}|\sum|^{\frac{1}{2}}} exp\left[-\frac{1}{2}(X-\mu_i)'\sum^{-1}(X-\mu_i)\right]$$

(3.1)

The ratio of the densities of two multivariate normal populations is given below; Usman, 2011:

$$\frac{P_1(X)}{P_2(X)} = \frac{\exp\left[-\frac{1}{2}(X-\mu_1)'\sum^{-1}(X-\mu_1)\right]}{\exp\left[-\frac{1}{2}(X-\mu_2)'\sum^{-1}(X-\mu_2)\right]} \geq k$$

$$= \exp\left[-\frac{1}{2}\left\{(X-\mu_1)'\sum^{-1}(X-\mu_1) - (X-\mu_2)'\sum^{-1}(X-\mu_2)\right\}\right] \geq k \qquad (3.2)$$

By taking the natural logarithms of equation (3.2) above; which is monotone increasing we have:

$$-\frac{1}{2}\left\{(X-\mu_1)'\sum^{-1}(X-\mu_1) - (X-\mu_2)'\sum^{-1}(X-\mu_2)\right\} \geq \log k \qquad (3.3)$$

The second term of (3.3) above is the Mahalonobis square distance between $N(\mu_1,\sum)$ and $N(\mu_2,\sum)$. For k suitably chosen (which of course can be one and then log k will be

zero), the left hand side of equation (3.3), can be expanded and reposition to get the following equation:

$$X' \sum^{-1} (\mu_1 - \mu_2) - \frac{1}{2} (\mu_1 + \mu_2)' \sum^{-1} (\mu_1 - \mu_2) \geq \log k \tag{3.4}$$

The first expression of equation (3.4) above is well known as Fisher's linear discriminant function which is linear in the component of the observation vector.

3.7 Method used in this study

In this work, five predictors that are well recognized for characterizing Broncho-Pneumonia status in infants were considered. These variables are baby's weight at birth, baby's weight 4week after, sex, mother's age and mother's occupation.

By the method of Euclidean distance, the mean vectors and the covariance matrices of a sample of unaffected infants (π_1) and affected infants (π_2) were stated below;

Let

$$\overline{X}_i = \begin{pmatrix} \bar{x}_{i1} \\ \bar{x}_{i2} \\ ,, \\ ,, \\ \bar{x}_{ip} \end{pmatrix} \tag{3.5}$$

where $\overline{X}_i$ represent the sample mean vector and i denote the two groups (unaffected and affected).

Let $\bar{x}_{i1}, \ldots, \bar{x}_{ip}$ represent the individual mean vectors for the five variables i.e. p=5

For instance;

$$\overline{X}_{i1} = \frac{1}{k}\sum_{i=1}^{n} X_{i1} \tag{3.6}$$

where $\overline{X}_{11}$ is the mean of the first variable in group one, while $\overline{X}_{21}$ represent the mean of the first variable in group two, k is the number of the cases and n is the sum of all observations in a Particular group.

The sample variance-covariance matrix is given as;

$$S_i = \begin{pmatrix} s_{ii} & \cdots & s_{ip} \\ \vdots & \ddots & \vdots \\ s_{pi} & \cdots & s_{pp} \end{pmatrix} \tag{3.7}$$

where S_i denote variance-covariance matrix, for $i = 1, 2,$ S_{ii} denotes an individual variance and

$S_{ip} = S_{pi}$ denotes an individual covariance for $p = 1, 2, \ldots, 5$.

The illustrations are given below,

$$S_{ij} = \frac{1}{k_i}\sum_{i=1}^{ni} \left(X_{ij} - \overline{X}_{.i}\right)^2 \qquad i = 1,2 \quad j = 1,\ldots,5 \qquad \text{(general variance)} \tag{3.8}$$

$$S_{12} = \frac{1}{k}\sum_{i=1}^{n} \left(X_{i1} - \overline{X}_1\right)\left(X_{i2} - \overline{X}_2\right) \qquad \text{(an individual covariance)} \tag{3.9}$$

Let π_1 denote group one (unaffected infant) and

π_2 denote group two (affected infant)

The Euclidean distance of the unaffected infants (π_1) is;

$$\hat{l}_1 = \overline{X}_1' S_p^{-1}(\overline{X}_1 - \overline{X}_2) \tag{3.10}$$

and Euclidean distance of the affected infants (π_2) is;

$$\hat{l}_2 = \overline{X}_2' S_p^{-1} (\overline{X}_1 - \overline{X}_2) \quad (3.11)$$

where S_p denotes the pooled variance matrix

The mean Euclidean distance used in this study for the two groups is given as;

$$\widehat{M} = \frac{1}{2}\left(\hat{l}_1 + \hat{l}_2\right) \quad (3.12)$$

and the discriminant function is calculated by

$$\hat{Y} = X' S_p^{-1} (\overline{X}_1 - \overline{X}_2) \quad (3.13)$$

Therefore, the classification rule is that

$if \quad \hat{Y} \geq \widehat{M} \quad classified\ as\ group\ one\ (\pi_1)$ and

$if \quad \hat{Y} < \widehat{M} \quad classified\ as\ group\ two\ (\pi_2)$

where $\hat{Y}$ denote the estimate of the discriminant function, and $\widehat{M}$ denote the mean Euclidean distance for unaffected and affected (BPn) groups

$$\mathbf{X}' = \left(X_1 \quad X_2 \right) \quad (3.14)$$

$$S_p = \frac{n_1 S_1 + n_2 S_2}{n_1 + n_2} \quad (3.15)$$

Since $n_1 \neq n_{2,,}$equation (3.15) will be used, but if $n_1 = n_2$ then the estimated pooled

variance S_p above becomes:

$$S_p = \frac{S_1 + S_2}{2} \tag{3.16}$$

where $S_1 and\ S_2$ are the respective sample variance covariance matrices for the two groups, and $n_1 and n_2$ are the sample size of the two groups respectively.

The Fisher's linear discriminant model used is:

$$Y_{HS} = \beta_0 + \beta_1 X_{BW_1} + \beta_2 X_{BW_2} + \beta_3 X_S + \beta_4 X_{MA} + \varepsilon \tag{3.17}$$

Where;

Y_{HS} $denotes\ response\ probability(Health\ Status)$

X_{BW_1} $denotes\ baby's\ weight\ at\ birth$

X_{BW_2} $denotes\ baby's\ weight\ 4\ weeks\ after$

X_S $denotes\ baby's\ Sex$

X_{MA} $denotes\ mother's\ age$

X_{MO} $denotes\ mother's\ occupation$

β_0 $is\ the\ expected\ value\ of\ Y\ when\ the\ X's\ are\ set\ as\ 0$

β_i $is\ the\ regression\ coefficient\ for\ each\ corresponding\ predictor\ variable$

ε $is\ the\ error\ of\ the\ predictor.$

In this study, '0' is used to represent "unaffected" and '1' represents "affected (BPn)". The mean of the dichotomous random variable Y_{HS}, designated by P_{HS}, is the proportion of times that the pneumonia takes the value '0'. Equivalently;

$$P_{HS} = P(HS = 0) = P(\text{Unaffected}) \tag{3.18}$$

3.8 Test For Significance of Canonical Correlation (Wilk's Lambda)

The degree of linear relationship existing between two set of variables can be measured by means of canonical correlation which takes on values between minus-one and plus-one inclusive ($-1 \leq r \leq +1$). The closer the value of canonical correlation is to one, the stronger the degree of linear relationship existing between the two set of variables. Also, the stronger the degree of linear relationship existing between the two set of variables, the better the linear discriminant function between the set of variables. The significance of canonical correlation is measured by the Wilk's' Lambda statistic as follows;

Hypothesis for Canonical Correlation:

H_0: There is no linear relationship between the two set of variables

H_1: There exists a linear relationship between the two set of variables

Test statistic:

$$\lambda = \frac{|W|}{|W + H|}$$

(3.19)

Where;

W is residual variance

H is variance due to linear relationship

W+H is the total variance.

Decision Rule:

Reject H_0 if $p < 0.05$ otherwise accept H_0 at the 5% level of significance.

Reason for the Use of Canonical Correlation

This research work would investigate the strength of the underlying relationship between several pairs of variables. It is therefore, justifiable to use the Canonical Correlation coefficient whenever such relationships are to be measured.

3.9 Box's M - Test for the Equality of Covariance Matrices

Box's M is used to determine whether two or more covariance matrices are equal. It is also used to test for homogeneity of covariance matrices. Alvin, (1934) stated that, the basic assumptions of the linear discriminant model is that the two covariance matrices must be equal. Hence, the Box M test is used to investigate this assumption; otherwise the discriminant model may be misleading. The Box M test procedures are given below:

Hypothesis for Box's M Test:

H_0: The two covariance matrices are not equal

H_1: The two covariance matrices are equal

Test statistic:

$$M = \frac{|S_L|}{|S_S|}, \tag{3.20}$$

where S_L and S_S are the larger and smaller variance respectively.

Decision Rule:

Reject H_0 if $p<0.05$ otherwise accept H_1 at the 5% level of significance.

$$P = \Pr\left(F_{\nu 1,,\nu 2} > M\right)$$

where; M ---- calculated value of Box's M

$F_{\nu 1,\nu 2}$ ---- f-distribution with v_1,v_2 df

Reason for the Use of Box's M

The reason for using Box's M is that, it is justifiable and even necessary to investigate the equality of the two covariance matrices. That is, if they are equal, then the linear discriminant model is appropriate otherwise the non-linear discriminant model is applied. Hence, the Omnibus test is applied.

3.10 Logistic Regression Model

Binary Logistic or Logit regression deals with the binary case, where the response variable consists of just two categorical values. Logistic regression model is mainly used to identify the relationship between two or more explanatory variables (X_i) and the dependent variable (Y). Logistic regression model has been used for prediction and determining the most influential explanatory variables on the dependent variable (Cox and Snell, 1994). The logistic regression model is the most frequently used regression model for the analysis of these data. It is important to understand that the goal of an analysis using this model is the same as that of any other regression model used in statistics; that is, to find the best fitting and most parsimonious interpretable model to describe the relationship between an outcome (response) variable and a set of independent (explanatory) variables.

The most important difference between a logistic regression model and the linear regression model is that the outcome variable in logistic regression is *binary* or *dichotomous*. In linear regression model it is assumed that an observation of the outcome variance may be expressed as $y = E(y/x) + \varepsilon$. This difference between logistic and linear regression is reflected both in the form of the model and its assumptions.

Unlike linear regression, which predicts the actual values of the response variable, logistic regression models the probability associated with each level of the response variable by finding a linear relationship between predictor variables and a link function of these probabilities. Different link functions offer different goodness-of-fit for the data. The following link functions are common, and during data analysis, the link function that offers the best goodness-of-fit for the data is chosen.

(i) logit (ii) normit/probit (iii) gompit (complementary log-log)

Binary and ordinal logistic regression offers all three link functions; nominal logistic regression offers only the logit link function. In order to simplify notation, we use the quantity $\pi(x) = E(Y|X)$ to represent the conditional mean of Y given X when the logistic distribution is used. The specific form of the logistic regression model used is:

$$\pi(\mathrm{x}) = \frac{e^{\beta_0+\beta_1\mathrm{x}+\beta_2\mathrm{x}+\beta_3\mathrm{x}+\beta_4\mathrm{x}+\beta_5\mathrm{x}}}{1+e^{\beta_0+\beta_1\mathrm{x}+\beta_2\mathrm{x}+\beta_3\mathrm{x}+\beta_4\mathrm{x}+\beta_5\mathrm{x}}} + \varepsilon$$

(3.21)

A transformation of $\pi(x)$ that is central to the study of logistic regression is the *logit transformation*. This transformation is defined, in terms of $\pi(x)$, as:

$$g(\mathrm{x}) = \ln\left[\frac{\pi(\mathrm{x})}{1-\pi(\mathrm{x})}\right] = \beta_0 + \beta_1\mathrm{x} \qquad (3.22)$$

the importance of this transformation is that $g(x)$ has many of the desirable properties of a linear regression model. The logit, $g(x)$, is linear in its parameters, may be continuous, and may range from $-\infty$ to $+\infty$, depending on the range of x.

Logistic regression models are adequate for those situations where the dependent variable of the regression problem is binary. That is, the dependent variable has only two possible

outcomes, e.g., "success/failure" or "normal/abnormal". We assumed that these binary outcomes are coded as 1 and 0 (Hosmer and Lemeshow, 2000). The application of linear regression models to such problems would not be satisfactory since the fitted predicted response would ignore the restriction of binary taking on values for the observed data. In this work, an attempt is made to estimate a population regression equation as;

$$Y_{HS} = \beta_0 + \beta_1 X_{BW_1} + \beta_2 X_{BW_2} + \beta_3 X_S + \beta_4 X_{MA} + \beta_5 X_{MO} + \varepsilon \qquad (3.23)$$

The response Y_{HS} is continuous, and is assumed to follow a normal distribution. The study will predict the mean value of the response corresponding to a given set of values for the explanatory variables.

However, there are many situations in which the response of interest is dichotomous rather than continuous. Examples of variables that assume only two possible values are disease status (the disease is either present or absent) and survival following surgery (a patient is either alive or dead). (Hosmer and Stanley, 1989). In general, the value '0' is used to represent a "success" or the outcome we are not interested in, and '1' represents a "failure". The mean of the dichotomous random variable Y, designated p, is the proportion of times that it takes the value '0'. Equivalently;

$$P_{HS} = p(HS = 0) = p(success) \qquad (3.24)$$

Our interest is to estimate the probability (p) associated with a dichotomous response for various values of an explanatory variable.

3.11 Chi-square test

In this research work, the sex and mother's occupation variables used are categorical variables. The sex has two categories, male and female, and mother's occupation has three which are house wife, civil servant and business mother. Hence, it is important to test, during the modeling process, whether BPn infection depends on sex and/or mother's occupation which is better analyzed by the chi-square statistic.

Hypothesis for Chi-square Test:

H_0**:** The two variables are not independent

H_1**:** The two variables are independent

Test statistic:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

(3.25)

where O_{ij} is observed value and e_{ij} is expected value.

$$P = \Pr(\chi_\upsilon^2 > \chi_0^2)$$

Decision Rule:

Reject H_0 if $p<0.05$ otherwise accept H_0 at the 5% level of significance.

Reason for the Use of Chi-square Test

One of the reasons for the use of chi-square is to investigate the inter-dependency among sex and mother's occupation against BPn. The Chi-square test of independence is the appropriate statistical tool for the investigations of such inter-dependency.

3.12 Omnibus Chi-Square Test

The Omnibus Chi-square test is a log-likelihood ratio test for investigating the model coefficients in logistic regression. The test procedures are as follows:

Hypothesis for Omnibus Chi-square test:

H_0**:** The model coefficients are not statistically significant

H_1**:** The model coefficients are statistically significant

Test statistic:

$$\chi^2 = 2\left[\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} \ln\left(\frac{O_{ij}}{e_{ij}}\right)\right] \tag{3.26}$$

where;

O_{ij} *denote observed values, and*

e_{ij} *denote expected values*

Decision Rule:

Reject H_0 if $p < 0.05$ otherwise accept H_1 at the 5% level of significance.

$$P = \Pr\left(\chi_{\upsilon}^2 > \chi_0^2\right)$$

where;

χ_0^2 - - - chi-square calculated

χ_{ν}^2 - - - chi-square value with vdf

Omnibus test is used to investigate the significance of the model coefficient in logistic model.

3.13 Multicollinearity

A critical condition for the application of least squares is that the explanatory variables are not perfectly linearly correlated (i.e.,$r_{xixj} \neq 1$). The term multicollinearity is used to denote the presence of linear relationships (or near linear relationships) among

explanatory variables. If the explanatory variables are perfectly linearly correlated, that is, if the correlation coefficient for these variables is equal to unity, the parameters become indeterminate: it is impossible to separately obtain numerical values for each parameter and the method of least squares breaks down (Draper and Smith, 1998)

Effects/Consequences of Multicollinearity

Since $\beta = (x^I x)^{-I} x^I y$

(3.27)

Where $(x^I x)^{-I} = \frac{Cof\,(x^I x)^T}{\det(x^I x)}$

(3.28)

If the $x's$ are highly correlated, then

i. $\det(x^I x) -> 0$

$=> (x^I x)^{-I} -> \infty$

$therefore\ \beta -> \infty$

ii. $\text{var}(\beta) = s^2(x^I x)^{-I} -> \infty$

(3.29)

Variance is infinite. This results in insignificant t-ratios

$$t^* = \frac{\widehat{\beta} - \beta}{SE(\widehat{\beta})} = \frac{\widehat{\beta} - \beta}{\sqrt{\widehat{\delta}^2}}$$

(3.30)

iii. The variables of the parameter estimate are unnecessarily high.

CHAPTER FOUR

ANALYSIS, RESULTS AND DISCUSSIONS

4.1 Introduction

In this chapter, the data are fitted to the linear discriminant and logistic regression models. The results of the analyses are presented and discussed. The data were analyzed using SPSS version 21.0

4.2 Descriptive Statistics

Table 4.1: Descriptive Statistics on Continuous Variable

Variables	N	Min	Max	Sum	Mean	Std.Dev	Variance
Weight at birth(kg)	433	0.70	2.80	831.00	1.9192	0.40804	0.166
Weight after 4weeks(kg)	433	1.00	3.70	1015.10	2.3443	0.49942	0.249
Mother's age	433	13	41	11985	27.68	5.776	33.367
Valid N (listwise)	433						

Table 4.1 shows the predictor variables with Mother's age indicating the highest value of mean and standard deviation. While, the weight at birth recorded the smallest values for the mean, standard deviation and variance

Table 4.2: Categorical Variables Coding

Categorical Variables		Frequency	Percent	Valid Percent	Cumulative %
Sex	1	200	46.2	46.2	46.2
	2	233	53.8	53.8	100.0
Mother's Occupation	0	199	46.0	46.0	46.0
	1	137	31.6	31.6	77.6
	2	97	22.4	22.4	100.0
Health Status	0	257	59.4	59.4	59.4
	1	176	40.6	40.6	100.0

The table 4.2 shows the values of which categories were coded as the reference for each of the explanatory variables

Table 4.3: Descriptive Statistics on the Two Groups

Health Status	Variables	Mean	Std Dev.	N
Unaffected	Weight at birth(kg)	2.0934	0.27820	257
	Weight after 4weeks(kg)	2.4932	0.45223	257
	Mother's age	27.5370	5.43207	257
Affected(BPn)	Weight at birth(kg)	1.6648	0.43373	176
	Weight after 4weeks(kg)	2.1270	0.48629	176
	Mother's age	27.8864	6.25539	176
Total	Weight at birth(kg)	1.9192	0.40804	433
	Weight after 4weeks(kg)	2.3443	0.49942	433
	Mother's age	27.6790	5.77638	433

In table 4.3, the predictor in the two groups (that is, Mother's age) has the largest mean and standard deviation respectively. However, the predictor in unaffected group is associated with the predictor in affected group with smallest mean and standard deviation (that is, Weight at birth)

4.3 Checking the Assumptions of Discriminant Analysis

In this section, we check for the assumptions of Discriminant analysis base on the outcomes of the Pooled Correlation Matrices, Pooled Covariance Matrices, Test of Equality of Group Means and Multicollinearity.

Table 4.4: Pooled Correlation Matrices

	BWABirth	BW4WAfter	MoAge
BWABirth	1.000	0.320	-0.031
BW4WAfter	0.320	1.000	0.027
MoAge	-0.031	0.027	1.000

Table 4.4 shows that the data set satisfy the assumptions of discriminant analysis which states that, the predictors are not highly correlated with each other (that is, baby's weight at birth and baby's weight four weeks after), which has the correlation of 0.320. The correlation between self predictor is constant across groups.

Table 4.5: Pooled Covariance Matrices

	BWABirth	BW4WAfter	MoAge
BWABirth	0.122	0.052	-0.062
BW4WAfter	0.052	0.217	-0.072
MoAge	-0.062	-0.072	33.414

Table 4.5 reveals that all within variables are insignificant, that is,. BWABirth to BWABirth is 0.122 > 0.05. It also shows that, one variable with others have at most one significance while the remaining are insignificant

Table 4.6: Test of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
BWABirth	0.733	156.842	1	431	0.000
BW4WAfter	0.870	64.410	1	431	0.000
SEX	0.996	1.731	1	431	0.189
MoAge	0.999	0.382	1	431	0.537
MoOccup	0.998	0.846	1	431	0.358

	BWABirth	BW4WAfter	Sex	MoA	MoOC

Table 4.6 is on the potentials of variables, as Wilks' Lambda value tend to 1, the lesser is the ability of the predictor to discriminate. It is observed that baby's weight at birth is best at discriminating between the two groups. This satisfies one of the assumptions of the study

Table 4.7: Multicollinearity Results

BWABirth	1				
BW4WAfter	0.441672	1			
BGender	0.049825	0.004814	1		
MoA	-0.04173	-0.03546	-0.03308	1	
MoOC	0.119901	0.075955	-0.01509	0.026302	1

These results show that there is no significant correlation among independent variables

Table 4.8: Test Results of Box's M

	Box's M	58.133
F	Approx.	3.825
	df1	15
	df2	566140.580
	Sig.	0.000

Table 4.8 investigates the equality of the two covariance matrices. The test statistic is clearly stated in equation (3.20). Since p-value of 0.000 □ α-value of 0.05, it confirmed the equality of the covariance matrices for the two groups.

Conclusion: if all the assumptions of Discriminant analysis are satisfied, the study will adopt Discriminant analysis

4.4 Linear Discriminant Function

The coefficients in table 4.9 are used to develop the model for the Unaffected and Affected groups

Table 4.9: Fisher's Classification Function Coefficients

HStatus

	Unaffected	Affected
BWABirth	13.867	10.633
BW4WAfter	8.646	7.663
SEX	6.565	6.341
MoAge	0.887	0.888
MoOccup	-0.472	-0.061
Constant	-43.166	-34.806

Table 4.9 provides information (coefficient) that will be used to construct the models for the unaffected and affected groups. It alsoshow that baby's weight at birth have the largest value (BWABirth = 13.867 and 10.633) in the two groups. Mother's occupation has the lowest coefficients from the status. In the two groups, the response to health status is negative when all independent variables stand as zero (0). The negative constant value in unaffected group indicate the increases in response (y) as predictors (x) increases, likewise the affected group.

The Fisher's linear discriminant model for each group is constructed as follows;

Unaffected $group\ (\pi_1)$

$$Y_1 = -43.166 + 13.867\mathrm{X}_{\mathrm{BW}_1} + 8.646\mathrm{X}_{\mathrm{BW}_2} + 6.565X_S + 0.887X_{MA} - 0.472X_{MO} \quad (4.1)$$

$Affected\ group\ (\pi_2)$

$$Y_2 = -34.806 + 10.633X_{BW_1} + 7.663X_{BW_2} + 6.341X_S + 0.888X_{MA} - 0.61X_{MO} \quad (4.2)$$

Equations (4.1) and (4.2) are the developed models for Unaffected and Affected group respectively.

Table 4.10: Canonical Discriminant Function Coefficients

BWABirth	2.417
BW4WAfter	0.735
SEX	0.167
MoAge	0.000
MoOccup	-0.308
Constant	-6.373

Unstandardized coefficient allow for comparing variables measured on different scales. A coefficient with larger absolute values corresponds to variable with greater discriminating ability. Table 4.10 shows that baby's weight at birth has a coefficient value of 2.417, so it has a greater discriminating ability among the predictors.

$$Y_{HS} =$$

$$-6.373 + 2.417X_{BW_1} + 0.735X_{BW_2} + 0.167\mathrm{X_S} + 0.000\mathrm{X_{MA}} - 0.308\mathrm{X_{MO}} \qquad (4.3)$$

In equation (4.3), Baby's Weight at Birth has the largest coefficient which confirms the earlier of wilks' lambda in table 4.6, as the variable with the highest discriminating power.

Table 4.11: Functions at Group Centroids

Health Status	Euclidean values
Unaffected	0.544
Affected(BPn)	-0.794

Table 4.11 is an unstandardized canonical discriminant functions evaluated at group means. The mean of the discriminant function of two Euclidean distance is shown in equation (4.4):

The Cutoff point ($\hat{\mathbf{M}}$) is computed as follows;

$$\hat{\mathbf{M}} = \tfrac{1}{2}\left(\overline{Y}_1 + \overline{Y}_2\right) = \tfrac{1}{2}(0.544 - 0.794) = -\mathbf{0}.125$$

(4.4)

Therefore, the classification rule is stated as;

Classify as group 1 (Unaffected) if $Y_{HS} \geq -0.125$

Classify as group 2 (Affected) if $Y_{HS} < -0.125$

Table 4.12: Classification Results

			Predicted Group Membership		
		Health status	Unaffected	Affected(BPn)	Total
Original	Count	Unaffected	207(80.5%)	50(19.5%)	257(100%)
		Affected(BPn)	60(34.1%)	116(65.9%)	176(100%)

The classification table shows the practical result of using discriminant model on all the cases. The 433 patients were used in model fitting. From the sample 207 of 257 (that is, 80.5%) of Unaffected infants were correctly classified and 50 of the Unaffected were misclassified. On the other hand, 116 of 176 (that is, 65.9%) of affected infants were correctly classified while 60 (that is, 34.1%) were misclassified.

4.5 The Logistic Regression Model

Using the given Logistic Regression Model in equation (3.21), SPSS Package was used to analyze the data and the following table of model coefficients was generated as in table 4.12

Table 4.13: Variables in the Equation for the Sample Data

	B	S.E.	Wald	df	Sig	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
BWABirth	-2.915	0.365	63.671	1	0.000	0.054	0.026	0.111
BW4WAfter	-0.939	0.268	12.269	1	0.000	0.391	0.231	0.661
Sex(1)	0.209	0.237	0.778	1	0.378	1.233	0.774	1.963
MoAge	-0.002	0.021	0.008	1	0.931	0.998	0.959	1.039
MoOccup			7.841	2	0.020			
MoOccup(1)	-0.812	0.302	7.217	1	0.007	0.444	0.245	0.803
MoOccup(2)	-0.748	0.330	5.128	1	0.024	0.473	0.248	0.904
Constant	7.916	1.095	52.271	1	0.000	2741.241		

From table 4.13 the p-values of Gender and Mother's Age were insignificant, this implies that the two variables have no effect on BPn. The odds ratio of 1.23 indicate that the likelihood of effect of female is 23% higher than that of male, since male stand as reference. In mother's occupation, the outputs of the p-values of the two categories are less than significant value of 0.05. We then conclude that there is significant difference in the effect of the three categories (i.e. House wife, Civil servant and Business mother). Since House wife serve as reference, the two odds ratios (0.44 and 0.47) implies that the likelihood of effect of Civil servant and Business mother is 44% and 47% respectively lower than that of House wife.

From the table above, we have fitted logistic regression model as;

$$P_{HS} = \frac{e^{7.916-2.915X_{BW_1}-0.939X_{BW_2}+0.209X_S-0.002X_{MA}-0.812X_{MO_{(1)}}-0.748X_{MO_{(2)}}}}{1+e^{7.916-2.915X_{BW_1}-0.939X_{BW_2}+0.209X_S-0.002X_{MA}-0.812X_{MO_{(1)}}-0.748X_{MO_{(2)}}}} \qquad (4.5)$$

Alternatively:

$$ln\left(\frac{\hat{p}_{HS}}{1-\hat{p}_{HS}}\right)$$

$$= e^{\hat{\beta}_0+\hat{\beta}_1X_{BW_1}+\hat{\beta}_2X_{BW_2}+\hat{\beta}_3X_S+\hat{\beta}_4X_{MA}+\hat{\beta}_5X_{MO}+\varepsilon_{ij}} \qquad (4.6)$$

$$ln\left(\frac{\hat{p}_{HS}}{1-\hat{p}_{HS}}\right) = e^{7.916-2.915X_{BW_1}-0.939X_{BW_2}+0.209X_S-0.002X_{MA}-0.780X_{MO}}$$

The logistic model 4.6 will be used to predict the affected status of infants using a cut value or threshold probability of 0.5.

Table 4.14: Omnibus Test of Model Coefficients

		Chi-square	Df	Sig.
Step 1	Step	148.089	6	0.000
	Block	148.089	6	0.000
	Model	148.089	6	0.000

The table 4.14 indicates the omnibus test for the parameter convergence at the final stag of the iteration. Here the chi-square is highly significant (chi-square=148.089 with df=6 and $p<0.05$)

Table 4.15: Hosmer and Lemeshow Test

Step	Chi-square	Df	Sig.
1	6.134	8	.632

This table suggests that the model is a good fit to the data since p=0,632>0.05. However, the chi-square statistic shows that the BPn depend on the categorical data

4.6 Goodness of fit and classification power

It is pertinent to use the discriminant model to classify BPn status of infants, thus, equation (4.3) is used to test the goodness of fit and classification power.

Table 4.16: The outcome of randomly selected samples using Discriminant model

Sample	Sample size	Correctly classified	Misclassified

1	10	7	3
2	10	6	4
3	10	8	2
4	10	8	2
5	10	9	1
6	10	7	3
7	10	8	2
8	10	9	1
9	10	7	3
10	10	7	3
Total	100	76	24

This table shows the Ten random samples containing ten cases each taken from dataset. It was discovered that the average of 7.6 is correctly classified, while 2.4 misclassified.

Table 4.17: The outcome of randomly selected samples using logistic regression model

Sample	Sample size	Correctly predicted	Mispredicted
1	10	4	6
2	10	7	3
3	10	6	4
4	10	4	6
5	10	7	3
6	10	4	6
7	10	5	5
8	10	7	3
9	10	5	5
10	10	5	5
Total	100	54	46

Using equation (4.5), it was observed that 5.4 is averagely predicted and 4.6 is unpredicted. The outcome of the two goodness of fit above shows that the Discriminant model is better than Logistic regression model for prediction of broncho-pneumonia status in infant.

4.7 Discussion of Results

From the analysis carried out it has been observed that the Disrciminant model fit the data set. In comparing the outcome of the analyses, it was discovered that 76% of the random selected cases were correctly classified using the discriminant model built from the dataset. In logistic regression model, we discovered that 54% cases were correctly classified. In comparing the outcomes using the two models on the cases, it is observed that the discriminant model is best used in North Central Zone than Logistic regression model. However, if normality assumption fails, the logistic model is more preferred to the discriminant models. But if the normality assumption holds, the principle of parsimony prevails which model fits better for a particular situation must be determined using the goodness of fit results. In this case, the prediction of BPn is better done with discriminant model than logistic model in North Central Zone.

4.8 Major Findings

The following are the summary of findings;

i. There was no correlation among the independent variables (i.e. no presence of multicollinearity).

ii. In equation (4.3) above, the model indicate that baby's weight at birth make the highest contribution to the dicriminant function. Baby's weight after four (4) weeks and baby's sex contributed positively to the function and the last two variables (mother's age and mother's occupation) made low contribution.

iii. The discriminant model classified 80.5% cases and misclassified only 19.5% of the

Unaffected group. In Affected group, 65.9% were classified and 34.1%

misclassified. The developed discriminant model classified 76% randomly selected

cases and misclassified only 24%, this result validates the original classification.

iv. In table 4.4, it is observed that baby's weight at birth is best at discriminating between the two groups since it has the smallest value of wilks' lambda (that is, 0.73.

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.1 Summary

In this study, discriminant model and binary logistic regression were used for predicting the occurrence of Broncho-pneumonia among infants using five variables as predictor variables i.e. baby's weight at birth, baby's weight four weeks after, baby's sex, mother's age and mother's occupation.

The objectives of this research are to construct the discriminant and logistic regression model that is capable of tracking BPn infants based on their variables used and also to compare and contrast the predictive power of the discriminant model and logistic regression for Broncho-Pneumonia.

The transcription and experimental method of data collection was used in this study and the research was carried out in Abuja with data obtained from University Teaching Hospital, Gwagwalada and Nasarawa with data obtained from Federal Medical Centre, Keffi. The data were gathered and tabulated for 180 and 253 low birth weight infants respectively.

Discriminant analysis and logistic regression were multivariate techniques employed for the analysis of the work. Box's M test and Wilk's Lambda were used to confirm the

equality of the Covariance matrices and also to confirm the significance of the canonical correlation respectively.

5.2 Conclusion

In this research, linear discriminant and logistic regression model were applied to data collected for Broncho-Pneumonia from North Central Zone taking a case study of UTH, Abuja and FMC, Keffi Nasasrawa State. The result shows that the prediction of BPn is better done with discriminant model than logistic regression model in the zone.

Ten random samples of size 10 each taken from dataset were used to test the goodness of fit of the two models developed. The models were used for the prediction of the BPn status of selected samples. In discriminant model the average of 7.6 is correctly classified, while it misclassified 2.4. The study has predicted the BPn status of selected samples using the logistic model built in which an average of 5.4 were correctly predicted.

Equations (4.3) and (4.5) are the developed linear discriminant and logistic regression models constructed in this study, while, the related reviewed linear dicriminant and logistic regression models were in equation (2.1)and 2.2) respectively.

In this research, it was observed that linear discriminant model has a perfect classification than Logistic regression model. We also discovered that ‘baby’s weight at birth’ is the predictor that is best discriminating between the two groups

5.3 Recommendations

i. The researcher recommend that the models developed in this study could assist the doctors and other health practitioners to detect and monitor the prevalence and control of BPn among infants

ii. It is recommended that the discriminant model built should be used for BPn cases in the zone particularly at University Teaching Hospital, Abuja and Federal Medical Centre, Keffi Nassarawa State.

iii. It is also recommended that larger sample size and health facilities be used in further study. And use of other statistical package especially those dedicated to multivariate analysis on this area in order o elucidate intensive information or results.

iv. Doctors and Clinics should adopt the use of the models built in this research to discover the prevalence of BPn among infants so that adequate measures for prevention and control of Broncho-Pneumonia can be taken early enough.

5.4 Contribution to knowledge

This study used both baby and mother's attributes to identify the best model used in the study while previous studies only used either the baby attributes or that of the mother.

Re-sampling technique which was achieved through SPSS package was also used to validate the models.

The study also identified the predictor variable with the highest discrimination between the two (2) groups.

References

Anderson, A. A. (2008). Tracking Broncho-pulmonary dysplasia (BPD) using the logistic regression model. *International Journal of Statistics* 65, 59-66.

Anthony, J., Brooks, W. A., Malik J. S., Douglas, H. and Kim, M. E. (2010): Pneumonia research to reduce childhood mortality in the developing world. *The Journal of Clinical Investigations*. 2, 120-128

Beki, O. D. (2012). The Use of Discriminant Model and Logistic Regression for Tracking the Incidence of Broncho-Pulmonary Dysplasia among Infants. A Dissertation Submitted to Usman Danfodio University, Sokoto Nigeria.

Carlos, A. B., Christieny, C. M., Moreira, E. L. M,, Jose, R. R and Guilherme, M. S. (2007). Binary logistic model for predicting Broncho- Pulmonary Dysplasia (BPD). *J. Pediar*, 83(2), 163-170.

Cornfield, A. (2010). Binary logistic model for predicting Myocardia infection. *American Journal of Public Health,* 81, 15-21.

Danan, C. (2002). Gelatinase activities in the airways of premature infants and development bronchopulimonary dysplasia. *American Journal of physiology,* 283, 86-93.

Danbaba, A., Audu, M., and Mahmud, A. M. (2013). Investigation of risk factors of low birth weight using multivariate logistic regression analysis: *Journal of Humanities, Science and Technology.* A publication of Niger State Polytechnic. 3, 59-77

Draper N. R. and Smith, H. (1998): *Applied Regression Analysis.* John Wiley & Sons Inc. USA

Emin, M. A., Levent, D., Adelet, O. and Aysu, T. (2014). Epicardial Adipose Tissue, Netrophil–To-Lymphocyte Ratio and Platelet Lymphocyte Ratio with Diabetic Nephropathy. *International Journal of clinical and experimental medicine.* 7(7), 1794-1801

Eneh, A.U. and Ugwu, R. O. (2011). The Proportion of Low Birth Weight Due to Small Gestational Age (SGA) and Prematurity. *The Internet Journal of Pediatrics and Neonatology. 13, 1-5.*

Erimafa, J. T., Iduseri, A. and Edokpa, I. W. (2009). Application of discriminant analysis to predict the class of degree for graduating students in a university system. *International Journal of Physical Sciences.* 4(1), 16 - 21.

Fernaudez, P., Abbas, O., Dardenne, P. and Baeten, V. (2010). Discrimination of Corsican Honey by FT- Raman spectroscopy and Chenometrics. *Journal on Biotechnology Agron Soc Environ.* 15(1), 75 - 84.

Hosmer, D. W., and Lemeshow, S. (1989). A review of goodness of fit statistics for use in the development of Logistic models. *American Journal Epidemiology.* 115, 92-106.

Hosmer, D.W. and Lemeshow, S. (2000). Applied Logistic Regression, (2nd ed.). Wiley and Sons, inc. New York.

Jobe, A. H. and Bancalari, E. (2001). Broncho-pulmonary dysplasia. *American Journal of Respiratory and Critical Care Medicine.* 163(7), 1723 - 1729.

Kirkwood, R. B. and Stern, J. C. (2008). Essential Medical Statistics: Blackwell Science, London.

Lange, W., Perking, H., Andersen, R. E., Royce, R., Jewell, N. and Winkel S. W. (1989). Patterns of T–lympphocyte changes with HIV from sero conversion to the developed of AIDS. *Journal of Acquire Immune Defic Syndr.* 2, 63-69.

Levy, A. and Lemeshow, C. (2008). Sampling of Population (2nd ed.). John Wiley, New York.

Mobil, P., Landero, A., De-Antoni, G., Araujo, A. C., Avila-Donoso, H. and Moreno, J. (2010). Multivariate Analysis of Raman Spectra applied to microbiology: Discriminant of Microorganism at the species level. *Revista Mexicana De Fesica* 56(5), 378-385.

Moses, Y. O. (2007). *Econometrics "*A practical approach" Daybis limited, 10 Akinola Maja Avenue, Jericho, Ibadan Nigeria.

Ngwu, E. K. (2009). Birth weight of babies in relation to their mothers' nutritional knowledge and place of residence. *Nigerian Journal of Nutritional Science.* 30(2), 36 – 42.

Onyiriuka, A. N. (2010). Incidence of Low Birth Weight Infants in Twin Gestations. Niger J. Clin Pract; 13, 365 – 370.

Pope, D. P., Mishra, V., Thompson, L. (2010). Risk of low birth weight and Stillbirth associated with indoor air pollution from solid fuel use in developing countries. Epidemiology Rev. 70-80.

Pyng, L., Remco, M., Stephen, L., Adi, F., Katrien, G., Annette, M., Jean, L.,….,Tom, G. (2010). Din Cancer Rearch. NIH Publication Access. 58, 1713-1718

Rencher, A. C. (2002). Method of Multivariate Analysis (2nd ed.) John Wiley & Sons.

New York.

Rees, G., Brook, Z., Doyle, W., and Costeloe, K. (2005). The Nutritional Status of Women in the First Trimester of Pregnancy attending an inner-city Antenatal Department in the UK.Journal of Royal Society for the Promotion of Health; 125(5), 232 – 238.

Shibuya, K., Hoshimo, H. and Chiyo, M. (2003). Sub-epithelial Vascular Patherns in Bronchial Dysplasias Using a High Magnification Bronchovideoscope. Thorax 58, 989-995

Stephenson, B. (2008). Binary response and logistic regression analysis. http://www.public.iastic.edu

Suzanne, M., David, A., Kathleen, H. and John, L. (1997). Outcome of Extremely Low Birth Weight Infants Requiring Rescue High Frequency Ventilation Pediatric Research. 41, 212-212.

Tapia, J. I., Agost, D., Alegria, A. and Standen, J. (2006): Bronihopulmonary Dysplasin, Incidence, Risk Factors and Resource Utilization in a Population of South America Very Low Birth Weight Infants. *Journal de pediatric*. 82(1), 15-20

Thomas, W. (2007). Prevention and Treatment of Bronihopulmonary Dysplasin: Current Status and Future Prospects. *Journal of Perinatology*.3, 26-32

Usman, A. (2011). Statistical Methods for Biometric and Medical Research. Kaduna, Millennium Publisher (New Nigeria Newspaper).

Uthman, O. A. (2008): Effect of Low Birth Weight on Infant Mortality. *The Internet Journal of Epidemiology*. 6, 1-5

Vajpayee, M., Amatya, R. and Kauslik, S. (2004). Lymphocyte Immunophenotypes reference ranges in healthy Indian Adults: Implications for Management of HIV /AIDS in India Clin Immonol. 112, 290 – 295.

Vimalkumar, V. K., Moses, R. C. and Padmanaban, S. (2011). Binary Logistic Model for Detecting Prevalence and Risk Factor of Nephropathy in Type 2 Diabetic Patients. *International Journal of Collaborative Research on Internal Medicine and Public Health* 3(5), 598-615.

Vishwa, N. M., Madaki, U. Y., Vijay, V. S. and Babagana, M. (2015). Application of Discriminant Analysis on Bronch-Plumonary Dysplasia among Infants: *American Journal of Theoretical and Applied Statistics* 4(2), 44-51.

Yu-shun, C., Chiun, L. and Wang, K. (2011). Based on Fuzzy Linear Distribution Analysis for Breast Cancer Mammography Analysis. doi.ieeecomputersociety.org

Appendix I

Prior Probabilities for Groups

Health status	Prior probabilities	Cases Used in Analysis
Unaffected	0.500	257
Affected(BPn)	0.500	179
Total	1.000	433

Show the prior probability of misclassifying Healthy infants to BPn at $\alpha = 0.05$ and prior probability of misclassification, BPn of infants to Healthy baby is also 0.05.

Wilk's Lambda Test

Test of Wilk's Lambda	Chi-square	Df	Sig.
.772	111.125	5	.000

Wilk's Lambda test justifies the significance of the canonical correlation which gives 0.772 with p-value of 0.000 comparing the p-value of Wilk's Lambda of 0.000 with the predefine significance level of $\alpha = 0.05$.