

# MIDLANDS STATE UNIVERSITY



FACULTY OF SCIENCE AND TECHNOLOGY

DEPARTMENT OF APPLIED MATHEMATICS AND STATISTICS

## RESEARCH TOPIC

THE EFFICACY OF PROGNOSTIC RISK FACTORS IN CLASSIFYING MALARIA  
PATIENTS USING DISCRIMINANT ANALYSIS: A CASE STUDY OF SANYATI  
HOSPITAL.

BY

MAZIBISA KUDAKWASHE

R136926T

SUPERVISOR: MR DAMBAZA

This dissertation is submitted in partial fulfillment of the requirements of the Bachelor of  
Science in Mathematics Honours Degree of the Midlands State University

JUNE 2017

---

# ACKNOWLEDGEMENTS

I would like to start by thanking the Lord Almighty for the gift of life and for guiding me in every nook and cranny of my dissertation. I thank God for giving me the strength and power to write this dissertation. Praise be to God.

Special thanks goes to the department of Mathematics for their massive contribution towards my academic achievement. I am indebted to many, greatest appreciation goes to my supervisor Mr. Dambaza for his extensive assistance in coming up with this project. Profound gratitude goes to my work related immediate supervisors Mr. Mtisi (HSA) and Mr. N. Gwacha (HIO) for their assistance in data collection.

I further want to thank my fellow friends Lovemore and Diamond for helping me in applying machine learning algorithms to my dissertation and also my classmates Davie and Martin for assisting me in other areas of my dissertation. I also want to thank my best friend Brenda Chaka for her emotional support and the encouragement that she gave. It will be impossible to acquiesce adequately all the people who have been influential in this dissertation endeavor, to you all I send my intricate gratitude.

Finally, thanks go to my family, especially my parents for their financial support and being a pillar of my strength and inspiration.

# DEDICATION

This dissertation is dedicated to my parents and brothers Kudzai and Takudzwa.

# APPROVAL FORM

The undersigned certify that they have supervised, read and recommend to the Midlands State university for acceptance, a dissertation entitled: the efficacy of prognostic risk factors in classifying malaria patients into low and high risk groups using discriminant analysis: A case study of Sanyati Baptist Hospital, which is submitted in partial fulfilment of the Faculty of Science and Technology for the Bachelor of Science in Mathematics honours degree of Midlands State university.

.....

.....

Supervisor

Date

.....

.....

Chairperson

Date

.....

.....

# DECLARATION

I hereby declare that the work incorporated in this dissertation is original and has not been submitted to any institution for the award of a degree, diploma, or certificate. I further declare that the results in this research and considerations made contribute in general to the advancement of knowledge in education and in medical research.

.....

Name

.....

Signature

# Table of Contents

|   |      |
|---|------|
| <b>ACKNOWLEDGEMENTS</b> .....                                 | i    |
| <b>DEDICATION</b> .....                                       | ii   |
| <b>APPROVAL FORM</b> .....                                    | iii  |
| <b>DECLARATION</b> .....                                      | iv   |
| <b>LIST OF APPENCIES</b> .....                                | viii |
| <b>ABSTRACT</b> .....   | ix   |
| <b>CHAPTER ONE</b> .....                                      | 1    |
| 1.1 Introduction .....  | 1    |
| 1.2 Background of the study .....                             | 1    |
| 1.3 Problem Statement.....                                    | 4    |
| 1.4 Objectives of the study .....                             | 5    |
| 1.4.1 Aim.....  | 5    |
| 1.4.2 Objectives.....   | 5    |
| 1.5 Research questions.....                                   | 5    |
| 1.6 Significance of the study .....                           | 6    |
| 1.7 Assumptions of the study.....                             | 7    |
| 1.8 Delimitations of the study .....                          | 7    |
| 1.9 Limitations of the study .....                            | 8    |
| <b>CHAPTER TWO</b> .....                                      | 9    |
| <b>LITERATURE REVIEW</b> .....                                | 9    |
| 2.0 Introduction .....  | 9    |
| 2.1 Literature on Malaria.....                                | 9    |
| 2.1.1 Global Malaria Situation.....                           | 9    |
| 2.1.2 National Malaria Situation .....                        | 12   |
| 2.1.3 Malaria Situation in Sanyati .....                      | 13   |
| 2.2 Literature on discriminant analysis .....                 | 13   |
| 2.2.1 Discriminant analysis theory.....                       | 17   |
| 2.3 LDA as a classification tool .....                        | 19   |
| 2.3.1 Dimensional reduction in discriminant analysis.....     | 20   |
| 2.3.2 EM algorithm for missing values .....                   | 20   |
| Steps in EM.....  | 21   |
| 2.4 Choosing between logistic and discriminant analysis ..... | 22   |

|  |           |
|--|-----------|
| 2.5 Diagnostics in Linear Discriminant Analysis.....             | 23        |
| 2.6 Stepwise discriminant analysis .....                         | 23        |
| 2.7 Previous research review .....                               | 23        |
| <b>CHAPTER THREE</b> .....                                       | <b>30</b> |
| <b>METHODOLOGY</b> .....   | <b>30</b> |
| 3.0 Introduction.....  | 30        |
| 3.1 Study Area .....   | 30        |
| 3.2 Research design.....   | 31        |
| 3.3 Research instruments.....                                    | 31        |
| 3.4 Population .....   | 32        |
| 3.5 Sampling.....  | 33        |
| 3.5.1 Analysis sample.....                                       | 33        |
| 3.5.2 Holdout Sample.....  | 34        |
| 3.6 Plan for data analysis .....                                 | 34        |
| 3.6.1 Visualization .....  | 34        |
| 3.6.2 Transformation of discrete attributes into continuous..... | 35        |
| 3.6.3 Variable description .....                                 | 35        |
| 3.6.4 Data analysis method.....                                  | 38        |
| 3.6.5 Statistical Software’s.....                                | 38        |
| 3.6.6 Data exploration .....                                     | 39        |
| 3.7 Discriminant analysis .....                                  | 39        |
| 3.7.1 Model applicability.....                                   | 39        |
| 3.7.2 Discriminant analysis model.....                           | 40        |
| 3.7.3 Stepwise discriminant analysis using SPSS .....            | 41        |
| 3.7.4 Discriminant analysis using R. ....                        | 42        |
| 3.7.5 Biplot using Stata.....                                    | 42        |
| 3.7.6 Prediction.....  | 43        |
| 3.7.7 Model strength and fitness.....                            | 45        |
| 3.7.8 Validation of the results .....                            | 45        |
| 3.8 Ethical considerations.....                                  | 46        |
| <b>CHAPTER FOUR</b> .....  | <b>47</b> |
| <b>DATA PRESENTATION AND ANALYSIS</b> .....                      | <b>47</b> |
| 4.0 Introduction.....  | 47        |

|  |           |
|--|-----------|
| 4.1 Stepwise Discriminant Analysis.....                                    | 47        |
| Table 4.1.1: Analysis case processing summary .....                        | 48        |
| Table 4.1.2: Group Statistics .....  | 49        |
| Table 4.1.3 Variables in analysis.....                                     | 51        |
| Table 4.1.4 Variables entered/removed.....                                 | 52        |
| Table 4.1.5 ANOVA.....   | 53        |
| Table 4.1.6 Variables not in analysis .....                                | 54        |
| Table 4.1.7 Standardized canonical discriminant function coefficients..... | 55        |
| Figure 4.1 Biplot.....   | 56        |
| 4.2 Fisher discriminant function.....                                      | 57        |
| Table 4.2.1 Prior probabilities of groups .....                            | 58        |
| Table 4.2.2 Group means.....   | 58        |
| Table 4.2.3 Coefficients of linear discriminants .....                     | 59        |
| Table 4.2.4 Misclassifications table.....                                  | 61        |
| Table 4.2.5 Confusion matrix .....   | 62        |
| 4.3 Classifications and diagnostic testing .....                           | 64        |
| Figure 4.3 Boxplot.....  | 65        |
| 4.3.1 Model validity using SPSS and Stata13 .....                          | 67        |
| 4.3.2 Diagnostics Checking.....  | 68        |
| <b>CHAPTER FIVE .....</b>  | <b>72</b> |
| <b>SUMMARY, CONCLUSION AND RECOMMENDATIONS .....</b>                       | <b>72</b> |
| 5.0 Introduction.....  | 72        |
| 5.1 Summary .....  | 72        |
| 5.2 Recommendations.....   | 74        |
| 5.3 Variable summary and recommendation .....                              | 74        |
| 5.4 Conclusion .....   | 77        |
| APPENDICES .....   | 77        |
| REFERENCE LIST .....   | 80        |



# LIST OF APPENCIES

|  |    |
|--|----|
| Appendix A: Covariance matrix.....               | 78 |
| Appendix B: Test of Equality of group means..... | 78 |
| Appendix C: Pairwise group comparison.....       | 79 |

## **ABSTRACT**

Malaria is endemic in Sanyati, accounting to approximately 3000 patients both outpatient attendance and admissions diagnosed of the disease yearly and 15% of all hospital deaths. The research analyzed the efficacy of prognostic risk factors in classifying malaria patients into low and high risk groups using discriminant analysis: a case study of Sanyati Baptist Hospital. Secondary data was extracted from the inpatient morbidity and mortality register available in the information department from 1<sup>st</sup> of January 2013 to 31<sup>st</sup> December 2015. The results showed that prognostic factors that are age, distance, referral status, disease comorbidity were important predictors of malaria mortality and patient survival. However, it was found that the prognostics risk factors; gender, site of residence and number of reported symptoms were not good predictors of predicting whether a patient is in the high or low risk class. It was recommended, among other things that the hospital executives should implement the patient classification system to reduce malaria related deaths through effective distribution of scarcity drugs and doctors at Sanyati Baptist Hospital.

## **Acronyms**

SBH Sanyati Baptist Hospital

PCA Principal Component Analysis

FDA Fisher Discriminant Analysis

LDA Linear Discriminant Analysis

EM Expectation Maximization

CBU Central Buying Unit

PTC Procurement Tender Committee

WHO World Health Organization

ACT Artemisinin-based combination therapies

RBM Roll Back Malaria

# CHAPTER ONE

## 1.1 Introduction

This chapter will present the following: background of the study, problem statement, objectives of the study, research questions, significance of the study, assumptions of the study and limitations of the study.

## 1.2 Background of the study

The research intends to assess the impact of prognostic factors to enable classification of malaria patients into low and high risk groups using discriminant analysis. The reason of the background of the study is to highlight the reasons for undertaking this research study and give evidence why this project is worthy doing. Malaria continues to be a hyper-endemic disease in Sanyati district owing to the fact that, it remains one of the top devastating disease globally, with a highest morbidity and mortality rate in Africa and beyond.

Sanyati Baptist Hospital was built in 1953 and it was established in 1956 by the Rhodesian Government. A missionary called Reverend Dotson was one of the core founders of the hospital. It was established as a healing and Evangelistic Agency of the Baptist Mission in Zimbabwe. It was organized to minister to the physical, emotional and spiritual needs of the ill or injured. It is situated along Kadoma, Gokwe North road which is 87 km from Kadoma and about 12km from Sanyati growth point.

Many research work has been done in Africa and beyond on malaria, aiming at reducing malaria related deaths. Regardless of all effort, it still remains a top devastating disease in most of the tropical areas like Sanyati district, in Mashonaland West, where infection with plasmodium falciparum remains a major cause of death. In the tropical areas, the annual global mortality due

to plasmodium falciparum stands at 1.2 million and mortality rate due to its prevalence is between 15% to 30% globally (Dondorp,1999). Current WHO report asserts that, malaria threatens the lives of 3.2 billion people globally and leads to over one million deaths annually (WHO, 2016).

A research done in Ghana at Tamale teaching hospital, analyzed the risk factors of malaria mortality among children using logistic regression model and also assessed the interaction effect between age and treatment of malaria patient. The spatial variation of malaria incidences and socioeconomic factors were assessed over a duration of 2 years, 535 children from 9 villages of a small rural area with high plasmodium falciparum transmission in Ghana. The results showed that risk factors such as referral status, age, distance, treatment and length of stay on admission were important predictors of malaria mortality (A.R. Abdul-Aziz et al, 2012).

The variable referral status meant that those who use the hospital as a first point of consultation and referred to the district hospital from peripheral health facilities. The population in study were children of age between 0 to 14 years. Distance was defined as farness or nearest of a malaria patient to the hospital. Patients treatment was through giving artesunate amodiaquine or quinine. It was found that the risk factors like sex and season were not predictors of malaria mortality. Regardless of the perception that malaria transmission is more intense in wet season than dry session it failed to hold water in this study.

In Nairobi, a research was done to explore the risk of perceived malaria and some associated factors in informal tropical settlements. The discriminant analysis approach was used, the prognostic factors were as follows site of residence, age, ethnicity, and number of reported symptoms. Site of residence and number of reported symptoms were powerful discriminating variables to malaria mortality. Site of residence was categorized into two villages that are Viwandani and Korogocho of the same district. The risk of malaria mortality was high in

Viwandani as compared to Korogocho. The research went on to validate the results using a holdout sample. The method used was referred as the split-sample or cross-validation approach (Ye' et al, 2007).

The researches done in Ghana and Nairobi can be used as guides to this project since the areas in study were in the tropical region. The research done in Ghana found referral status, age, distance, treatment, length of stay as significant factors to malaria mortality. The variable length of stay and treatment given are not of relevance in this study since the aim of the research is to classify a patient soon after arrival. The variable sex was insignificant in one of the studies but the researchers' intuition is that this variable can have an effect in group separation.

A research done in Nairobi introduced these additional variables that are site of residence, ethnicity, number of reported symptoms. The variable ethnicity does not apply to my area of study since almost all of the population in Sanyati are Zimbabweans. Critically analyzing these previous researches it has resulted to the researcher choosing the following variables to be the candidate variables: age, gender, distance, referral status, site of residence, number of reported symptoms and disease comorbidity.

### **1.3 Problem Statement**

According to the recent statistical report at Sanyati hospital between 2013 and 2015, reviewed that an average of 2,5% of all confirmed malaria patients die every year. Malaria is one of the top prevalent diseases at Sanyati hospital, with an average of 22 patients dying yearly. This disease remains endemic because of hot climatic conditions. Malaria continues to be a life threatening disease, this study is designed to come up with a model that would help in classifying newly admitted patients in their predicted guilds. Prediction of newly admitted patients is of efficacy importance considering the constrains of drug scarcity and availability of doctors. A continual neglect of studies that analyze the survival of malaria patients will make the mortality rate at Sanyati Baptist hospital to continue to be problematic.

The malaria prevalence statistics at Sanyati hospital are so alarming and it is high time to find out possible ways to reduce the mortality rate through a scientific approach by analyzing prognostic factors and their explanatory power in patient classification at admission. Innocent patients are dying yearly as a result of malaria infections, it is beyond the scope of this research to reduce morbidity, the researchers main focus is to curtail mortality.

Owing to the current economic downturn in Zimbabwe it has resulted to scarcity of medical equipment and drugs at Sanyati hospital. Malaria patients are served on the basis of first come first serve without analysis of those who are at high risk. Some of the malaria drugs run out of stock which results to patients being referred to Kadoma general hospital. If proper distribution of malaria drugs is done giving first priority to those who are critically ill, so many lives will be served. Those patients categorized in the low risk group can be referred to clinics that maybe in position of malaria drugs at that given time. The first come save approach in place at the hospital need to be changed, some of the critically ill transferred patients end up dying along the way as a

result of the hospital having failed to identify those who are critically ill. The study will work towards classifying patients in the low and high risk groups so that those in the high risk group may be given immediate first attention as compared to the other groups.

The increase of malaria related deaths at Sanyati Baptist hospital has pushed the researcher to find a way to curtail mortality through classification of patients into high and low risk groups. This research will enable proper care to be given to those patients in the risk category thereby, increase chances of survival to the victims of malaria.

## **1.4 Objectives of the study**

### **1.4.1 Aim**

To reduce mortality rate through the aid of discriminant analysis to aptly classify newly admitted patients into high and low risk groups.

### **1.4.2 Objectives**

1. To identify prognostic factors that best classify patients into low and high risk group.
2. To apply the Fisher discriminant function to classify patients into lower and high risk category.

## **1.5 Research questions**

1. What are the contributing prognostic factors to classification of patients in the low and high risk groups?
2. Which discriminant model best discriminant between the lower and risk category?



## **1.6 Significance of the study**

This study is relevant to the health sector since most of the researcher work in malaria reduction has been focusing mainly on the spread and disease control. It is also necessary to analyze those who will have been infected by the disease to find best ways of increasing patient survival rate. This research will aid in grouping patients who are at danger of losing their lives and those who are not. This classification criterion will improve efficient drug, resource, service allocation in an optimal way without disadvantaging a certain patient. The implementation of scientific statistical based approaches like the discriminant analysis may explore other hidden solutions to the problem at hand. The researcher's envisaged findings would help not only individuals in the society, but may also help Doctors at large to understand the dynamics of a malaria patient. This will alarm Doctors to treat malaria patients effectively as a result increasing the chances of patient survival.

The application of data science techniques to problem solving in medical research is becoming popular and implementable. In conjunction with data mining which is a collection of analytical techniques many problems in biometry and epidemiology can be solved using mathematical algorithms emanating as a result of the imaging field of machine learning. The data mining techniques helps to uncover new trends and pattern recognition and visualization. Discriminant analysis is one of the data mining techniques used to discriminate a single classification variable using multiple attributes.

Discriminant analysis is highly applications in medical research, the adoption of discriminant function analysis (DFA) techniques has been applied in the medical field where binary or categorical dependent variable exist. Furthermore, the widespread availability of discriminant analysis computer programs has facilitated implementation of these techniques.

The major purpose of this method is to developing a predictive model to classify patients in their predicted classes. The aim of a discriminant analysis in this context is to classify patients, by a set of independent prognostic variables, into one of two mutually exclusive and exhaustive categories that are the low and high risk group. For expository purposes, we will limit this discussion to two classifications that are high and low group.

The increase in mortality at Sanyati hospital calls for analyzing the factors associated to find ways of improve the situation at hand. Noting the assumptions and the dynamics of discriminant analysis will lay a foundation towards lowering malaria deaths. It is of utmost importance to carry research on this topic to ensure that the problem of malaria mortality is lowered and managed.

### **1.7 Assumptions of the study**

- The sample is a true representation of the population.
- The significant identified variables meet the multivariate normality assumption
- Equal Covariance matrix for the low and high risk group.

### **1.8 Delimitations of the study**

The research is only restricted to Sanyati Baptist hospital since it is a case study. The model of classification will only be used at this hospital. The research data used covered a period from 1 January 2013 to 31 December 2015.

### **1.9 Limitations of the study**

The data that was used in this research was not intended originally for this type of study. This study suffered from a major limitation inherent in any retrospective analysis, in that data were not originally recorded with this type of study in mind. Some of the records were missing both in the laboratory and information department.

Confidential information of patients is not easily accessible due to secrecy policies of medical information. Some of the non-sample information that the researcher felt to be of importance to the achievement of the objectives was not available for disclosure.

### **1.10 Conclusion**

This chapter introduced the fundamentals of the dissertation and highlighted the problem being faced by Sanyati Hospital through the background of the study. The chapter outlined and discussed the background of the study, problem statement, objective of the study, research questions, significance of the study, assumptions of the study, delimitations of the study and limitations of the study.

# **CHAPTER TWO**

## **LITERATURE REVIEW**

### **2.0 Introduction**

This chapter is intended to reconnoiter the available literature that is linked with this dissertation. The aim being to appreciate the knowledge that is already present and canvass the available literature. This chapter will form the fundament of the theoretical framework of the study. Historical overview of the disease is of efficacy importance in variable selection. Critically looking at the previous produced documents will help in making abstract analysis of previous findings. This chapter has the following headings literature on malaria, literature on discriminant analysis and previous research review.

### **2.1 Literature on Malaria**

This section looks at the history of malaria globally, nationally and regionally. Literature review of the disease will form a solid background to solicit the spread and dynamics of the vector parasite.

#### **2.1.1 Global Malaria Situation**

Malaria is a parasitic disease that is transmitted not directly from human to human but by a female anopheles' mosquito vector. The mosquito is infected by humans then transmit the disease to the other person. This disease infects the red blood cells, causing recurring fever of sudden onset (F. Brauer,1984; Centre,2012). Malaria remains a devastating 'killer disease, there were approximately 584 000 malaria deaths worldwide of these, 90% in Africa, and 78% where children under 5. It remains responsible for the deaths of approximately 430 000 children in Africa every year (WHO, 2015).

In the history of malaria C.A Laveran was first to discover parasites in patient's blood in 1880. Dr. Ross (1897) went on to demonstrate a dynamic transmission of malaria between anopheles' mosquitoes and humans. Although his work received immediate acceptance in the medical community, his conclusion that malaria could be controlled by controlling mosquitoes was dismissed on the grounds that it would be impossible to rid a region of mosquito completely and that in any case mosquitoes would soon reinvade the region. Dr. Ross advanced his research by formulating a mathematical model that predicts that malaria outbreaks could be avoided if the mosquito population could be reduced below a critical threshold level, field trials supported his conclusions and led to some brilliant successes in malaria control. This research was an eye open towards the reduction of malaria infections (F. Brauer:1984).

G.B. Grassi (1899) demonstrated the life cycle from mosquito to man. The researches in the field continued, The Garki project was put in place to provide a way to eradicate malaria from a region temporarily. People who have recovered from an attack had a temporary immunity against reinfection. Thus, elimination of malaria from a region leaved the inhabitants of this region without immunity when the campaign ends, and the result would be an outbreak there after (F. Brauer, 1984). Global Malaria Eradication Research Agenda is working hard to reduce mortality. New tools and systems are being put in place to accommodate drugs, vaccines, diagnostics and insecticides. High sensitivity test for malaria are being designed some of the machines available are light microscopy and rapid antigen detection. Continued progress in scale-up and elimination will require improved tools for malaria control and surveillance.

The launch of Roll Back Malaria (RBM) commitment to tackle a disease that affects 3.2 billion people and has devastating effects in 1998 was a catalyst for renewed global health and development. This program was supported through the global partnership of WHO, UNICEF, UNDP and World Bank. There are other partners who played a vital role that are the US-President's Malaria initiative (PMI), Bill and Melinda Gates Foundation etc. The global Malaria Action plan was launched in September 25, 2008, by RBM partnership to reduce Malaria morbidity and mortality. There were four major scientific evidence bases for Malaria interventions put in place, that are Artemisinin-based combination therapies (ACTs), insecticide-treated bed nets (ITNs), indoor residual spraying (IRS) and intermittent preventive treatment in pregnancy (IPTP). Malaria exacts its greatest toll on the world's poorest and most marginalized places. It kills at least one million people a year, yet it is treatable and largely preventable with the tools available now (WHO, 2005).

Between the Global Malaria Eradication and the start of the Roll Back Malaria (1975-2000) was time for Science. The Scientist identified: treatment with combined drugs to optimize efficacy and delay resistance, diagnostic that can be deployed close to home and in facilities and can clarify where malaria transmission, illness, and death is occurring. The Scientist are still seeking new improved prevention diagnostics and treatment, new interventions (vaccines, larval control, repellants).

Many kinds of researches were done and some are still in place under research and development, Dr. Ross tried to eliminate the vector completely but it was not possible the population of mosquitoes in an area would start to grow exponentially. This also had a negative effect to the ecosystem as a result affecting the ecological system. The Garki project also came and was successful temporarily. The work of these medical scientist has contributed positively towards

malaria prevalence reduction. Their study created a room for evolvement of new ideas to pop up towards reducing malaria mortality. For the disease to die out it seemingly becoming impossible but researchers are furthering their studies towards zero malaria mortality.

The research in study is not based on malaria eradication on community based control or integrated vector control but to minimize the probability of malaria death at the hospital through severity of disease group classification. Grouping of patients considering the severity of disease will help to allocate drugs and services effectively. The findings of this research will help to curtail malaria death rate Sanyati hospital through patient classification into the low and high risk groups. The hospital executive is going to benefit from this project in the event that they implement the results and involving the problem at hand in their strategic planning and policy formulation.

### **2.1.2 National Malaria Situation**

There are different types of mosquito's species that are plasmodium: ovale, malarie, knowlesi, vivax and falciparum. The plasmodium falciparum is the common one in Zimbabwe and it is life threatening and can cause multiple organ damage, coma and death. The parasite first infects the liver where it begins to multiply. After some days, the parasite is released in the blood stream to infect the red blood cells and the cells eventually burst and infect others. If they reach high numbers they may cause severe disease or even death.

Intervention polices and strategies like direct observed treatment with primaquine, systems on monitoring adverse reactions to antimalarial exists, the sale of oral artemisinin-based monotherapies, patient of all ages receiving diagnostic test and use of larval control have helped in reducing malaria cases in Zimbabwe. These were the statistics available: reports confirmed cases 391 651, confirmed cases at community level: 90,728, reported deaths 200 (Zimbabwe Malaria Statistics:2015).

According to the Mash West province plans they have a goal to reduce crude death rate from 10.2 deaths per 1000 in 2012 to 8 deaths per 1000 in 2022. The objective is to reduce malaria incidence from 35/1000 in 2014 to 10/1000 by 2017 and malaria deaths to near zero by 2017 (MASHONALAND WEST PROVINCE MODO REPORT, 2015). In order for these targets to be achieved in the stipulated years. There are various programs put in place to meet the targets. This research in succor, will work towards zero malaria death through classifying patients into low risk and high risk category. As a result of classification distribution of resources and time in the environment of scarcity will be done effectively and efficiently.

### **2.1.3 Malaria Situation in Sanyati**

Sanyati is an area located in Mashonaland west province and it is in the tropical acreage. It is in the malaria zone and is one of the places that is mostly affected by malaria. The hot climate condition has caused an endemic situation in the area. According to F. Brauer (1984) “An endemic situation is one in which a disease is always present”. In the previous triennial, malaria has been always topping the list both on Out Patient Department (OPD) and Admission department at Sanyati Baptist Hospital.

The system in place is not bringing optimal results for reducing malaria related deaths. All malaria patients are just treated the same of which because of financial instability of the hospital right now calls up for a method that can help the medical practitioners to allocate drugs efficiently to lower the death rate of malaria patient.

## **2.2 Literature on discriminant analysis**



Discriminant analysis was developed as a method for calibrating, it is a tool for correctly classifying cases. For example, scientists having difficulty in classifying closely related species of plants or types of diseases can take a group of cases correctly classified or diagnosed and calibrate classification functions from easily measurable aspects of the specimens. Once the classification functions are developed on known cases, they can be used to classify unknown cases. This indeed, was the original use for discriminant analysis as it was developed by Fisher (1936). Fisher's goal was the ability to classify two species of plants by easily observable aspects. Theory suggested sepal length and petal width would be among the relevant distinguishing characteristics. Since with some labor it was possible to know to which species a particular plant belonged, the problem was to calibrate on known plants a classification function based on easily measured characteristics that could be used in the field on plants of unknown species (M.R Daniels and R. Darcy, 1983).

According to J.D. Knoke (1982) discriminant analysis can be used when the response variable is categorical and unordered, hence denoting group membership, the statistical problem is variously termed, classification analysis, or risk analysis. "It can be used in research areas where the dependent variable (Y) consist of categories rather than a continuous metric scale (interval or ratio). It is a tool for classification of new observational units, especially new respondents, into groups or categories in which it must probably belongs. Also, discriminant analysis as a result gives the probability of group membership. Classification is conducted on the basis of measured value for group characteristics for each observational unit separately" (M. Savic et al, 2008).

"The goal of discriminant analysis is to construct the model on the basis of observational unit's variation. On the basis of the discriminant model the classification of new observational units into the groups or categories will be conducted. Some authors like Timm (2002) indicate that goals of

discriminant analysis are to construct a set of discriminants that may be used to describe or characterize group separation based upon a reduced set of variables, to analyze the contribution of the original variables to the separation, and to evaluate the degree of separation” (M. Savic and et al, 2008).

The mathematical objective of discriminant analysis is to weight and linearly combine information from a set of p-independent variables in a manner that forces the k groups to be as distinct as possible (M.T. Brown, 2014). In the usual approach to develop predictive models, discriminant function analysis (DFA) is used to assign a test site to a group of matched reference sites. These groups typically are established by classification. DA identifies a set of habitat variables that show the strongest relationship with the classification. DFA then establishes a set of coefficients for the selected set of habitat variables that allows an exposed test site to be assigned a probability of belonging to each of the discharge destination groups created by the classification (T.B. Reynoldson et al, 2014).

The discovery bases of the Fisher discriminant analysis make it possible to classifying patients into the high and low risk. It is focused on classifying objects into predicted groups and it is applicable in different areas such as politics, education, biometry and finance. The Fisher linear discriminant analysis will be applied in the health sector specifically at SBH to classify patients into low and high risk groups.

Discriminant analysis evaluates the degree to which such variables differentiate the groups, hence the name "discriminator variables." The effectiveness of the discriminant analysis depends on the extent to which the groups differ significantly on these variables. Therefore, the decision to select

certain variables as potential candidate discriminator variables is critical to the success of discriminant analysis and the dissertation at large. Agresti (1996) pointed out that independent variables are usually selected in two ways either from previous research or from intuition. Variable selection can be based on previous studies or an investigator's professional opinion also can be relied upon when selecting potential discriminator variables. Variables should be chosen that are believed to represent dimensions on which the groups are expected to differ (M.T Brown et al, 2009).

The dependent variable is risk level of patient with two classes that are the [lower risk; high risk]. An admitted patient can either be discharged or die as a result of malaria infection. There are other possibilities that can happen to admitted patients that is a patient can abscond or be transferred to Kadoma General Hospital. Therefore, because of insufficiency of the data relating to transfers and absconds. The researcher opted for death and discharge class only as indicators of lower and high risk. The dependent variable is categorical in nature, which is seconded by J.D. Knoke (1982) when he suggested that discriminant analysis can be used when the response variable is categorical and unordered.

The explanatory variables are gender, age, distance, referral status, site of residence, number of reported symptoms and disease comorbidity. The nature of the independent variables is categorical and continuous. Although the probability statements used in discriminant analysis assume that these variables are continuous (and normal), the technique is robust enough that it can tolerate

discrete variables assuming that they are numeric (NCSS, 2015). Which applies that both continuous and discrete variables can be used as explanatory variables in multivariate analysis(DA).

### 2.2.1 Discriminant analysis theory

The basic problem in discriminant analysis is to assign an unknown subject to one of two or more groups on the basis of a multivariate observation. It is important to consider the costs of assignment, the priori probabilities of belonging to one of the groups, and the number of groups involved. The allocation rule is selected to optimize some function of the costs of making an error and the priori probabilities of belonging to one of the groups. Denote by  $P_i$  a priori probability of belonging to  $\pi_i$  the  $i$ th group; by  $c_{ji}$  the cost of assigning an observation to the  $j$ th group which the individual belongs to the  $i$ th group; and by  $D_i$  the region for which the assignment is made to population  $i$ .  $P(D_j | \pi_i)$  is the probability that an observation from  $\pi_i$  falls in  $D_j$ .

The classification is based upon the covariance matrix. LDA is based on the concept of searching for a linear combination of variables(predictors) that best separates two classes(targets) that are the high and the low risk group. To capture the notion of reparability Fisher, define the following score function.

$$Z = \beta X_1 + \beta_2 X_2 + \dots + \beta_d X_d$$

$$S(B) = \beta^T \mu_1 - \beta^T \mu_2 / \beta^T C \beta \quad \text{score function}$$

$\mu_1$  - mean for the discharge group

$\mu_2$ - mean for the death group

Given the score function, the score function is there to estimate the linear coefficients that maximize the score which can be solved by the following equation.

$\beta = C^{-1}(\mu_1 - \mu_2)$  model coefficients

$C = 1/n_1 + n_2(n_1 C_1 + n_2 C_2)$

Where  $\beta$ = linear model coefficients

$C_1 C_2$ : covariance matrices

$\mu_1 \mu_2$ : mean vectors

$n_1$ : number of observation in the low risk class.

$n_2$ : number of observation in the high risk class.

If we find  $C$  then it becomes possible to find the vector matrix of coefficients. Finally, a new patient admitted at Sanyati Baptist Hospital is classified by projecting it into the maximally separating direction and classify it as  $C_1$  if...

The main assumption of discriminant analysis to be met are homogeneity of variance, normality assumptions and equal class covariance. The purpose of linear discriminant analysis (LDA) in this project is to find the linear combinations of the seven prognostic factors variables that gives the best possible separation between the groups high and low. If we want to separate the malaria patients by level of risk, the end result of a patient is survival or death, so the number of groups is  $G=2$ . The number of explanatory variables is 7 (7 prognostic factors;  $p=7$ ). The maximum number of useful discriminant functions that can separate the patients by risk level prognostic factors is the minimum of  $G-1$  and  $p=7$ , and so in this case it is the minimum of 1 and 7, which is 1. Thus,

we can find at most one useful discriminant function to separate the patients by prognostic factors, using the successful candidate explanatory variables.

### **2.3 LDA as a classification tool**

Discriminant analysis is one of the statistical algorithms of machine learning and is frequently used in classification problems. These are some of the examples of classifiers decision trees, decision rules, naïve Bayesian classifiers, Bayesian belief networks, nearest neighbor classifiers, logistic regression, support vector machines and artificial networks.

I. Kononenko and M. Kukar (2007) brings out the point that in the medical field there is a lot of data that can be utilized to improve health standards and patient survival. A typical classification task is medical diagnosis where a patient is described with continuous variables such as age, height, weight, body temperature, heart rate, blood pressure and discrete attributes such as sex, skin discoloration, location of pain. These mentioned variables may be used to determine the health status of a patient and associated diseases through designing a classification model.

The availability of past medical records, including diagnoses of all patients that had been treated at the hospital can be used to improving hospital standards. Data utilization through learning and analysis from the set of patients with known diagnoses can be used to improve the treatment, service of new patients.

The task of the discriminant function learning algorithm is to calculate coefficients whose structure is fixed in advance. A discriminant function is actually a hypersurface, dichotomizing between two classes in the attribute space. As a hypersurface can be defined only in a continuous hyperspace, most of the attributes need to be continuous. If there are more than two classes, a separate hypersurface is necessary for each pair of classes. Discriminant functions can be linear,

quadratic, polynomial, etc. In case of linear discriminant function the corresponding hypersurface is a hyperplane dichotomizing between two classes (I. Kononenko and M. Kukar, 2007).

In linear discriminant problems Fisher's linear discriminant function is frequently used. It assumes normal distribution of learning example within each class. It maximizes the Euclidean distance between averaged examples from both classes. It accounts for potentially different class variances and finds an optimal classification boundary between classes (I. Kononenko and M. Kukar, 2007).

### **2.3.1 Dimensional reduction in discriminant analysis**

Fisher Discriminant Analysis (FDA) is a linear method of dimensionality reduction from the expression space comprising all selected discriminatory prognostic factors to just a few dimensions where the separation of sample classes is maximized. FDA is similar to Principal Component Analysis (PCA) (Alter, 2000; Holter et al., 2000) in the linear reduction of data (Johnson and Wichern, 1992; Dillon and Goldstein, 1984). The major difference is that the discriminant axes of the FDA space are selected such as to maximize class separation in the reduced FDA space, instead of variability as in the case of PCA. The discriminant axes of FDA, termed as discriminant weights ( $V$ ), maximizing the separation of sample classes in their projection space (D. Hwang et al, 2007).

### **2.3.2 EM algorithm for missing values**

The expectation maximization (EM) is a description of a family of related algorithms, not a specific algorithm. Therefore, EM is a meta-algorithm which is used to devise particular algorithms. There are two main applications of the EM algorithm. The first is when the data indeed has missing values, due to problems with or limitations of the observation process. The second (more frequent) is when optimizing the likelihood function is analytically intractable but can be simplified by assuming the existence of additional but missing (or unobservable) variables (I. Kononenko and M. Kukar, 2007).

## **Steps in EM**

### **E-step**

Since the  $Z$  values are unobservable, the EM first finds the expected value of the complete-data log-likelihood for  $\log f(T/Z|\theta)$  with respect to the unobservable data  $Z$  given the observed data  $T$  and the current parameter estimates.

### **M-Step**

In this step EM iteratively improves the initial estimate  $\theta_0$  and constructs new estimates  $\theta^{(1)}, \dots, \theta^{(N)}$ .

These two steps are repeated as necessary.



## 2.4 Choosing between logistic and discriminant analysis

Classifying an observation into one of several populations is discriminant analysis, or classification. Relating qualitative variables to other variables through a logistic cdf functional form is logistic regression. Estimators generated for one of these problems are often used in the other. If the populations are normal with identical covariance matrices, discriminant analysis estimators are preferred to logistic regression estimators for the discriminant analysis problem. In most discriminant analysis applications, however, at least one variable is qualitative (ruling out multivariate normality). Under nonnormality, logistic regression model is preferred as compared to discriminant analysis (S. J and S. Wilson, 1978).

Discriminant analysis model is one of the mostly used multivariate method in classification problems. This is applicable when the dependent variable is dichotomous and is an appropriate statistical technique for testing the hypothesis that the group means of a set of explanatory variables for two groups. The group mean is referred to as a centroid and it indicates the typical location of any patient from a particular group.

Logistic regression is preferred as another method of classification and it does the same job as the discriminant analysis. The logistic regression is equivalent to a two-group discriminant analysis. The researcher chose to use discriminant analysis because the data set met the assumptions and also because of having a better experience with discriminant analysis as compared to the logistic regression.

Discriminant analysis relies on strictly meeting the assumptions of multivariate normality and equal variance covariance matrices across groups. Many researchers opt for logistic regression because it does not have to face strict assumptions and its more applicable in many situations. The

other reason is that logistic regression is similar to multiple regression, it only differs in the sense that it predicts the probabilities of an event occurring. It is mostly preferred also because of straightforward statistical tests and due to the reason that it has got the ability to incorporate nonlinear effects.

## **2.5 Diagnostics in Linear Discriminant Analysis**

Although discriminant coefficients can be determined under a regression model, regression diagnostic measures are shown to be inappropriate for detecting influential observations in linear discriminant analysis. The temptation of applying regression diagnostic measures in linear discriminant analysis must be resisted. When multivariate normal distribution is assumed it is usually also assumed that the covariance matrices are the same. (W.K Fung, 1995).

## **2.6 Stepwise discriminant analysis**

The widely used stepwise discriminant analysis procedure selects one variable at a time. However, instances occur in which variables occur naturally paired, such as the real and imaginary parts of a Fourier transformed signal. The existing stepwise discriminant procedure selects variables one at a time and might select the real part of the  $i$ th harmonic and the imaginary part of the  $j$ th harmonic. In discriminant analysis, an aspect of pattern recognition, one is generally given a large number of variables and the objective is to select a subset that best classifies the data into their correct groupings. The most popular technique is to select the single variable that can discriminate among different groups. Then the next best variable is selected and so on. The chosen variables are then used to form multivariate discriminant functions to classify the patterns or signals (C.F Lam and M.Cox, 1981).

## **2.7 Previous research review**

S.J. Lewis et al (1992) investigated on the effects of antimalarial chemoprophylaxis and other variables on the severity of falciparum malaria. Forward stepwise logistic regression analysis was performed to model the odds of having severe malaria compared with mild malaria given six potential explanatory variables: age, sex, chemoprophylaxis ethnic origin, time to presentation, and where the malaria was acquired.

Chemoprophylaxis is the prevention of infectious disease by the use of chemical agents. Prior chemoprophylaxis led to a reduction in the severity of falciparum malaria. Ethnic origin, time to presentation, and sex were also associated with the severity of malaria. The explanatory variables are almost the same with the one used in this research. Logistic regression was used in Lewis' research, the researcher is going to use discriminant analysis since it has strengths in classification.

Several guidelines can be followed in selecting discriminator variables. First, relevant theory should be consulted to determine which variables have the greatest potential for distinguishing among the groups of interest. Second, investigators should consider variables that have been shown to be relevant to group discrimination by previous research or theory (Brown & Tinsley, 1983). The research mentioned above has helped the researcher in choosing variables even though candidate variables vary depending on factors like geographical location, type of parasite.

According to P.G Kremsner et al (2009) appreciated that plasmodium falciparum malaria is a common cause of morbidity in African children, but identifying those who are likely to die is problematic. Previous studies suggested that circulating malarial useful predictor of severity, but none were large enough to detect any association with mortality. Most of the researches focused on severity of the disease as the dependent variable, but the researcher will focus on the end destine point to be able to classify newly admitted malaria patient in lower and high risk group.

The goal of this research is to find the axis of greatest discrimination between the low and the high risk group. The researcher will test whether the means of the two groups along the axis are significantly different, and attempt to assign patients into their predicted groups. Any variable that will be included in the prognostic model is going to help in classifying new patients in their respective class. The main focus of the researcher is prediction of group membership and classification of new admitted patients into the lower and high risk group.

The research done in Ghana analyzed the risk factors of malaria mortality among children using a logistic regression model and also assessed the interaction effect between age and treatment of malaria patient. Secondary data was obtained from the inpatient morbidity and mortality returns register at Tamale Teaching Hospital. In Ghana, malaria is a significant cause of adult morbidity and the leading cause of workdays lost to illness. Malaria is hyper-endemic in Ghana, accounting for 44% of outpatient attendance, 13% of all hospital deaths, and 22% of mortality among children less than five years of age. Malaria presents a serious health problem in Ghana. It is also the leading cause of workday loss due to illness in the country. For instance, it accounts for 3.6 ill days in a month, 1.3 workdays absent and 6.4% of potential income loss to Ghana's economy (A.R. Abdul-Aziz et al :2012).

The model used in this research in Ghana was done using Logistic regression and secondary data was used. The research at Sanyati Baptist Hospital is going to use discriminant analysis since the main objective of my research work is to classify patients. Many researchers have opted for logistic regression because of its flexibility of not being subjected to multinomial assumptions. The data set used in this case study is from January 2013 to December 2015.

The results of the study showed that risk factors such as referral status, age, distance, treatment and length of stay on admission were important predictors of malaria mortality.

However, it was found that the risk factors; sex and season were not good predictors of malaria mortality. Finally, the interaction effect between age and treatment was found to be significant.

Analysis of previous work is important in variable selection, the variables referral status, distance, gender were included as candidate variables in this research work. Treatment and length of stay on admission were not part of the candidate variables because the nature of the variables are not in line with achieving the research objectives. Factors that were considered are only those available at the time of diagnosis since the research is based on classifying patients to either high or low risk group to help medical practitioners to allocate drugs and services taking into consideration the severity status of a malaria patient.

Tamale, despite many years of prevention and control measures, malaria still remains a public health problem in low lying and water logged areas. In some areas across the metropolis, the transmission persistently occurs throughout the year. It is interesting to note that Tamale Teaching Hospital (TTH) possesses large amount of data on diseases, in particular malaria, on its hospital register. The data is usually compiled and submitted to the district and the regional Ghana Health Service directorate for preparation of quarterly and annual reports. Though records on malaria disease and its risk factors are usually not studied, yet it serves as a rich source of information for the stakeholders in the field (A.R. Abdul-Aziz et al, 2012).

The problem highlighted by the research at Tamale Teaching Hospital is the same problem at Sanyati hospital. The hospital possesses large amount of patient's data. Reports on weekly, monthly, quarterly and yearly are compiled and send to the district and regional Zimbabwe health services. There is no data analysis that is being done at the hospital to improve the welfare and survival rate of malaria patients. Yet the hospital is in position of rich data set that can bring solutions to some of the problems arising at the hospital.

The study provides evidence of the risk factors which influence in-hospital malaria mortality among children, from 0 – 14 years, at Tamale Teaching hospital in the northern region of Ghana. The model indicates that distance contributes more, among other factors, in terms of influencing malaria in hospital deaths. Distant villages or areas with ill resourced health centers or none at all suggest problems of access to health care, which does translate into high mortality rate. Thus the further the village is from the health Centre, the more disadvantaged the households are in terms of getting early health care. The study showed that patients within 5 km of hospital were less likely to die in hospital than those beyond 5 km, and does reflect the fact that nearness to the hospital improved early access to care, thus reducing the risk of mortality. It was also observed that referral children were at higher chance of dying in hospital, after adjusting for distance and other risk factors. This seems to suggest that delayed effective treatment, in the process of being transferred to the Tamale Teaching hospital, increased the severity of the disease. This could be because most referring health facilities may often be faced with stock-out of effective drugs or may not have prompt access to ambulatory support when needed.

This also suggests inadequate care being available at primary facilities, regardless of whether they are distant from the hospital or not. It is also possible that referring hospitals are referring the more severe cases which are expected to have higher mortality case. Although, there is the perception that malaria transmission is more intense in the wet season than the dry season, yet the study showed that there were 1306 (57%) cases in dry season and 987 (43%) cases in wet season. At least, in Tamale, season surprisingly was not even significant predictor in the model. This could be due to the fact that all 3 years were combined in this analysis, implying that an interaction between year and season could improve the understanding of the ‘season’ effect. Also, the predictor sex was not significant in the model, which suggests that it does not contribute significantly to

predicting deaths, though the converse was shown in Malawi. This study provides evidence that mortality for malaria among children, in and around, Tamale metropolis could be described as high (A.R. Abdul-Aziz et al, 2012).

The study indicated that many more children used the Tamale Teaching hospital as their first call for consultation. Most of the children were aged between 0 – 5 years and just few were above 10 years. The findings showed risk factors such as referral status, age, distance, treatment type and length of stay on admission were significantly contributing to mortality of children administered as malaria patients at the Tamale Teaching hospital. However, risk factors such as sex and season were not significantly contributing to malaria mortality. Finally, the interaction effect between age and treatment was found to be significant, which may imply that the ACT treatment is more effective at certain age group compared to another age group.

These were the proposed strategies: a strategic plan to build poly-clinics in every district capital and cheap-compound health facilities, at least in every community. This could help curb the long distance villagers staying around the metropolis have to travel to access health care; other stakeholders such as the World Health Organization (WHO), should step up effective campaign to discourage, if not ban entirely, the use of quinine in treating malaria and rather strengthen the use of ACTs drugs in the fight to reduce malaria mortality; and the expansion of ambulance services as well as improving more assessable roads in and around the Tamale metropolis to facilitate timely transportation of referral cases. It was recommended, among other things, that the government should provide more assessable roads and expand ambulance services to the various Districts/communities in and around the Tamale metropolis to facilitate referral cases (A.R. Abdul-Aziz et al, 2012).

The hospital has got a major role to play towards making a critical analysis on the attributes being possessed by a patient to identify those who will be approaching death. The likelihood of dying at a hospital varies from one country to another and from hospital to hospital. By classifying patients to low and high risk class will help in increasing patient survival rate. Most of the hospitals have a likelihood of dying based on an exhaustive population of all admitted patients. The researcher will focus on admitted malaria patients only, since this disease is amongst the top disease that affects people in Sanyati district.

### **Conclusion**

The purpose of this chapter was to highlight and discuss what other scholars wrote concerning discriminant analysis and malaria. This literature will help the researcher to come up with a simplified and reliable model for evaluation of new observational patients. Thereby, results helping the hospital executive to proactive decision making and knowledge discovery.



# **CHAPTER THREE**

## **METHODOLOGY**

### **3.0 Introduction**

The main focus of this chapter is to explore fundamental ground that is going to make data analysis results reliable. All the subsections of this chapter are the building blocks towards inferential analysis of the data generated. Pathak (2010) asserts that the reason of research methodology is to highlight the information pertaining to the procedures employed in conducting the research. It covers the following study area, research design, research instruments, population and sample, plan for data analysis, ethical considerations.

### **3.1 Study Area**

Sanyati is located in Mashonaland West Province, it is generally classified as malaria endemic area. Sanyati Baptist Hospital has an average of 300 admitted malaria patient per year (Hospital report:2015). It is situated along Kadoma- Gokwe North road which is 87 km from Kadoma and approximately 12km from Sanyati growth point. It began as an Evangelical mission station in 1940, ministering Christianity and providing spiritual healing doctrine to the community. It was later on upgraded to hospital standard in 1953. It serves as a referral center for more than 10 health clinics.

### **3.2 Research design**

Freeman (1984) defines research design as a plan or strategy for conducting an empirical study. On the other hand, Mohsin (1984) views a research design as a plan that enables one to reason from observed facts and events to logically sound conclusions. In this work, the researchers aim is to solicit information on malaria to facilitate reduction of mortality rate at Sanyati Hospital by scientifically classifying new patients into low and high risk groups. The classification would facilitate rational and economic use of the scarce medical resources at the hospital at any given time.

The researcher is going to use discriminant analysis to classify patients in the low and high risk groups. This kind of classification will help to answer the general objective of the research. LDA is mathematically robust and often produces models whose accuracy is trustworthy. It is a highly reliable statistical technique in classification problems. There are other methods of data classification like logistic regression, neural networks that may have been used but however, the researcher chose to use discriminant analysis.

### **3.3 Research instruments**

Kumar (2013) points out that secondary data source of information is information which is collected from existing records which were previously used for other purposes. Secondary data was used since the research is not based on peoples thought, opinions but on natural occurrence. The data used in this study were obtained in the archival records and extracted from discharge records charts of all admitted malaria patients at Sanyati Baptist Hospital over a period of January 2013 to December 2015. The registration chart included patients` age, sex, disease comorbidity,

date of admission and discharge outcome (i.e. death, discharged home, or absconded), village or location of residences, cost (i.e. for treatment), referral status and treatment given.

The use of secondary data is economical in the sense that cost of getting the data is minimum as compared to the use of questionnaires which is costly. The secondary data is readily available at the information department. The data meet some of the characteristics that are expected to be available: that is relevance, accuracy, reliability and sufficiency.

### **3.4 Population**

The data is going to cover a period from 1<sup>st</sup> January 2013 to 31<sup>st</sup> December 2015. The information will be extracted from patient charts. Chisnall (1997) defined population as a group of people or objects which are similar in one or more ways and which form the subjects of the study area extracted, in a particular survey. For this study, the population consists of all admitted malaria patients as confirmed by the laboratory to be malaria diseased during the period of study. The total population of patients in the time frame of study is approximately 1 000 malaria patients. In the period of investigation 900 patients were discharged and 100 patients died within the three-year period. The ratio between the low risk and high risk is 9:1 respectively, that is for every nine patients discharged there is one malaria death.

### **3.5 Sampling**

Proportional stratified sampling was used to divide the population into two strata of low risk and high risk. This sampling technique was used since there is heterogeneity among the groups and homogeneity within the groups. Simple random sampling was then used to select sufficient number of patients in each stratum. The population was stratified in order to have a good representation of the low risk and high risk classes and to be able to get meaningful results in data analysis.

Brown and Tinsley (1983) recommended that the total sample size should be at least ten times the number of discriminator variables. Huberty (1975) stated that the number of cases in the smallest group must be at least three times the number of variables. According to Huberty's assertion the sample size in this research is justifiably. A ratio of 2:1 was used to come up with 100 patients from the analysis and 50 for validation. The whole sample size selected is 150 patients taken from the two strata in the proportion of 105 from the low risk strata and 45 from high risk strata.

#### **3.5.1 Analysis sample**

An analysis sample of 100 patients was taken, taking 70 from discharged and 30 from the dead group. In discriminant analysis many at times the sample is divided into two subsamples. The analysis set also called the training set is used for estimation of the discriminant function and the holdout sample for validation process. According to various literature that the researcher has found there are no definite guidelines that have been established in dividing the sample into analysis and holdout sample. The most popular split up approach is to divide the sample equally, but some researchers prefer a 75-25 or 60-40 split over.

### **3.5.2 Holdout Sample**

The researcher followed a proportionately stratified sampling procedure that is analysis sample consists of 70 discharged and 30 dead, and the holdout sample consist of 35 discharged and 15 dead. The holdout sample was used to avoid giving an inflated predictive accuracy of the function. This method of validation is referred as the cross-validation or split-sample approach.

### **3.6 Plan for data analysis**

Analytic strategies were mapped out after successful sampling following the statistical procedures that are data cleaning, data transformation, data visualization, data summary in preparation of eventual inferences from the sample survey. Planning for data analysis is of efficacy importance in coming up with valid, trustworthy, unbiased statistical findings. Having a clear plan is important for research quality and integrity. It forms a firm foundation to the exploration of the next chapter

#### **3.6.1 Visualization**

Data visualization is a transformation of the raw data into a more presentable (and understandable) shape (I. Kononenko and M. Kukar, 2007). The data was reviewed, transformed in most appropriate format during data preparation. Visualization of each single and paired attributes were made through the use of scatterplots, histograms and biplot.

### **3.6.2 Transformation of discrete attributes into continuous**

Several statistical packages like STATA, SPSS cannot deal with discrete variables. All the discrete variables are implicitly assumed continuous for analysis purposes. Amongst the seven explanatory variables gender, referral status, site of residence and disease comorbidity were dichotomous. Gender and referral status are two valued discrete variables: site of residence, disease comorbidity are multi-valued discrete variables. These variables were coded using the codes 0,1,2,3 as indicated on the heading variable description.

### **3.6.3 Variable description**

The number of variables at start of the analysis were eight variable inclusive of the dependent variable. The dependent variable is risk level of patient at admission, which is binary, that is the patient is either classified as low or high risk. The explanatory variables are the prognostic factors that are age, gender, distance, referral status, site of residence, number of reported symptoms, and disease comorbidity.

#### **Risk level of patient**

The dependent variable is risk level of patient and this is determined by the discharge and death of malaria patients. The variable was coded as follows ( $0 = Low\ risk, 1 = High\ risk$ ). When a patient is admitted there are four end results discharged, death, abscond and patient transfer but due to lack of adequacy of the factors abscond and patient transfer there were excluded in this analysis and in the sampling frame.

## **Age**

Age is a continuous variable with an age distribution that emanates from 4 to 85 years. The researcher chose to exclude those who are less than 4 years and more than 85 years. This was done to avoid biasness caused by these age groups. Neonatal death were excluded in analysis because of the challenges and complexity of determining whether it is malaria or other child birth associated factors that may have contributed to the death of the child.

## **Gender**

Gender is a dummy variable and the following codes were used 1 if sex of respondent is male and 2 if female. Recent and previous studies in malaria mortality has shown that gender is not a significant factor in determining the malaria related death risk but the researcher has got an intuition that it can be a contributing factor towards group separation.

## **Distance**

Distance is the radius value between the hospital (center) and the location or residence of a malaria patient. Distance to the hospital in many previous researches has been treated as a dichotomous variable instead of continuous and this has got a drawback of losing some important attributing information. Some of the reasons why so many researches qualifies the variable it's because of the measures instrument used that would have already treated the variable as qualitative. Available in the information department is a scale of estimated distances to measure the actual distance of a patient. Therefore, the variable distance was treated as a continuous variable.

### **Site of residence**

Site of residence is the location or place of residence of a malaria patient. It is classified into 3 categories based on urban, semi urban, and rural based on the classifications criteria used in the information department, taking (1 = *urban*, 2 = *semi rural*, 3 = *rural*). In other previous researches this nominal variable has been coded based on different site of residence allocation approach.

### **Referral status**

Referral status is a situation where a malaria patient is transferred from a clinic to the hospital in other words using the hospital as a second health referral point. It is a dummy variable taking (1 = *YES* or 0 = *NO*). A “YES” answer means that the patient has been referred to the hospital and “NO” otherwise.

### **Disease comorbidity.**

Disease comorbidity is a situation whereby a malaria patient suffers from another certain type of disease at the time of admission. This variable was coded as 1 if a patient at the time of admission is suffering from malaria and HIV/AIDS, 2 if suffering from malaria and other diseases, 3 if infected with malaria only.



### **3.6.4 Data analysis method**

The nature of the data and variables in analysis are the ones that determine the statistical method employed. The dependent variable that is risk level of patient narrowed the statistical method that have to be employed in this analysis. There are many methods that are applicable to my research topic but the researcher opted for discriminant analysis a multivariate approach. Discriminant analysis in a broad sense is a very powerful statistical tool technique for many types of analyses including epidemiology and biometry. It is a long-standing method for deriving the dimensions along in which groups differ.

### **3.6.5 Statistical Software's**

R programming, Stata and SPSS were the software's that were used in data analysis. Johnson and Wichern (2007, p. 649) state that software support is very useful in graphical presentation of discriminant models. Moreover, since discriminant analysis is a robust statistical method there is need of advanced statistical software like R programming that helps in data exploration and analysis. Visual displays are important aid in discrimination and classification. Sophisticated computer graphics allow us to visually examine multivariate data in different dimensions. Other Microsoft packages like excel, publisher were used not for analysis but for graphs and charts editing.

### **3.6.6 Data exploration**

The explanatory variables were checked for partial correlation before the onset of using a stepwise discriminant analysis. Scatter plots were drawn and the calculation of the covariance matrix was made. This was done to check validity of the assumption of multicollinearity. Descriptive statistics were made to identify outliers, missing values etc. Outlier may exist as a result of incorrect data entry or statistical software malfunction.

The data set extracted for study in this research were having some missing attributes. A machine learning algorithm called EM algorithm (expectation maximization algorithm) was used to fill-in the missing values. EM algorithm uses the maximum likelihood estimation method. The algorithm predicted all the missing values from our two split samples that are the analysis and the holdout sample.

### **3.7 Discriminant analysis**

Discriminant analysis was developed as a method for calibrating, a tool for correctly classifying cases. Once the classification functions are developed on known cases, they can be used to classify unknown cases. This, indeed, was the original use for discriminant analysis as it was developed by Fisher (1936) before it become a powerful classification tool in different areas of study.

#### **3.7.1 Model applicability**

Discriminant analysis model is one of the used multivariate method in classification problems. This is applicable when the dependent variable is dichotomous and is appropriate statistical technique for testing the hypothesis that the group means of a set of explanatory variables for two groups. The group mean is referred to as a centroid, it indicates the typical location of any patient from a particular group.

Logistic regression is preferred as another method that do the same job as the discriminant analysis. The logistic regression is equivalent to two-group discriminant analysis. The researcher chose to use discriminant analysis because of having a better experience with the method as compared to the logistic regression.

Discriminant analysis relies on strictly meeting the assumptions of multivariate normality and equal variance covariance matrices across groups. So many researchers opt for logistic regression because it does not have to face strict assumptions and its more applicable in many situations. Other reason is that, logistic regression is similar to multiple regression, it only differs in the sense that it predicts the probabilities of an event occurring. It is mostly preferred also because of straightforward statistical tests and due to the reason that it has got the ability to incorporate nonlinear effects.

### **3.7.2 Discriminant analysis model**

The equation for the discriminant analysis is as follows:

$$Z = W_1X_1 + W_2X_2 + \dots + W_nX_n$$

Where

Z= discriminant Z score of discriminant function

W<sub>i</sub> = discriminant coefficient for independent variable i

X<sub>i</sub>= independent variable.

### **3.7.3 Stepwise discriminant analysis using SPSS**

There are different kinds of variable selection used in discriminant analysis like the simultaneous estimation which involves computing the discriminant function where all variables are considered concurrently. This kind of approach does not take into consideration the discriminating power of each explanatory variable. The researcher opted for the stepwise estimation approach which is an alternative to the simultaneous approach. Stepwise criteria involve entering explanatory variables into the discriminant function one at a time on the bases of their discriminating power.

The first step is to regress the dependent variable (discharge destination) with each of the explanatory variables that is age, gender, distance, referral status, site of residence, number of reported symptoms and disease comorbidity. The initial variable is then paired with each of the other explanatory variable one at a time, and the variable that is best improving the discriminating power of the function in combination with the first variable is chosen. The variables that are highly contributing to group separation are included and put in the model depending on the discriminating power of the variable. This procedure is done until all the significant variable are entered. The insignificant variables are removed in each step basing on the F to enter criteria. Different statistical methods that are the Roy's greatest characteristic root, Wilks' lambda, Pillai's criteria, Hotelling's trace all evaluate the statistical significance of the discriminatory power but in this research Wilks' lambda was used.

### **3.7.4 Discriminant analysis using R.**

To achieve the second objective of the study, the R software was used and data entry was done using SPSS and to export the data to R, the data was converted first to an excel.csv file for readability purpose. The two samples that are the analysis and the holdout sample were imported into the R software and before running any discriminant analysis in R a package called the “library mass” have be activated first. This package is necessary to run discriminant analysis using R. Discriminant analysis was done using the analysis data to come up with a model that best classify the patient into low and high risk category. The holdout sample data was used to validate the model prediction accuracy.

### **3.7.5 Biplot using Stata**

Dimensionality reduction is important when analyzing data using a two stage discriminate analysis. Computational and simulations are made to represent more than two explanatory variables that is hyperplane to a linear plan. The discriminant analysis method will create an equation which will minimize the possibility of misclassification of newly admitted patients into their predicted predicaments. The method maximizes the distance between the low and high risk group. In this research there are four explanatory variables. That means we need to have a 4 dimensionality graph of which in practical sense it is difficulty to visualize. The method of dimensionality reduction will be applied through the use of a biplot. not necessarily remaining with a single predictor variable. All the variables contribution will be this is what is called reduction by no loss of precision.

### 3.7.6 Prediction

After coming up with a discriminant analysis model the holdout sample was used to test the classification accuracy of the model. The Press's Q statistic was used as a simple measure that compares the number of correct classifications with the total holdout sample and the number of groups. The calculated value is then compared with the critical value from the Chi-Square distribution table with 1 degree of freedom.

The Q statistic is calculated using the following formula:

$$\text{Press's } Q = \frac{[N - (nK)]^2}{N(K-1)}$$

Where

N= total sample size

n =number of observations correctly classified

K= number of groups

## Confusion matrix

|                                   | Actual (correct class)            |                                  | $\Sigma$      |
|-----------------------------------|-----------------------------------|----------------------------------|---------------|
| Predicted (Classified as)         | “0” High Risk<br>(Negative class) | “1” Low Risk<br>(Positive class) |               |
| “0” High Risk<br>(Negative class) | True negative(TN)                 | False positive(FP)               | PP=TN+FP      |
| “1” Low Risk<br>(Positive class)  | False negative(FP)                | True positive(TP)                | PP=FN+TP      |
| $\Sigma$                          | NEG=TN+FP                         | POS=FP+TP                        | n=TN+FN+TP+FP |

## Sensitivity and specificity

Sensitivity and specificity are frequently used in analyzing the confusion matrix.

Sensitivity is a relative frequency of correctly classified positive examples

$$Sens = \frac{TP}{TP + FN} = \frac{TP}{POS}$$

$$Spec = \frac{TN}{TN + FP} = \frac{TN}{NEG}$$

Classification accuracy is therefore:

$$Acc = \frac{TP + TN}{TN + FP + FN + TN} = \frac{TP + TN}{n}$$

Where

n= total number of the holdout sample, TN= number of true negative examples

TP= number of true positive examples, FN= number of false negative examples

FP= number of false positive examples, NEG= number of negative examples

PP =number of predicted negative examples, POS =number of positive examples

### **3.7.7 Model strength and fitness.**

Wilk's Lambda was used to indicate the significance of the discriminant function. The canonical correlation table was used to assess the multiple correlation between the predictors and the discriminant function. It provides an index of the model fit which is interpreted as being the proportion of variance explained ( $R^2$ ). The larger the eigenvalue, the more of the variance in the dependent variable that is explained by the function.

### **3.7.8 Validation of the results**

The final stage of discriminant analysis involves validation of discriminant results to provide assurances that the results has external and internal validity. With propensity of discriminant analysis to inflate the hit ratio if evaluated only on the analysis sample, cross-validation is an essential step.

Several ways can be used to test the validity of the model. A holdout sample was set aside to test model accuracy. The whole sample size consisted of 150 malaria patients 66.6% was for the analysis sample (the sample on which the model was developed) and 33.3% was for the holdout sample to check for model prediction accuracy. The holdout sample was applied using R to avoid



the inflation of the hit ratio value. In this manner, validity is established in classifying observations that were not used in the estimation process.

Stata and SPSS were also used to classify new patients basing on the analysis sample of 100 patients. This kind of classification has got the disadvantage of causing upward bias that occurs in the prediction accuracy of the discriminant function. This is caused by using same observations used in developing the classification matrix with those used in computing the discriminant function.

### **3.8 Ethical considerations**

According to Bryman and Bell (2007) in his ten principles of ethical considerations says research participants should not be subjected to harm in any ways whatsoever. Respect for the dignity of research participants should be prioritized. The classification of patients in the low and high risk category is very sensitive. The results of this study will be used only by the medical practitioners. No patient should know his/her survival chances as this may result to patient panic.

The information pertaining to a patient is confidential in the medical field. This rightly protects the anonymity of the patients. No names of patients were used in this research, the results are based on group analysis and also the classification process of newly admitted patients is kept secret.

### **Conclusion**

The methodology chapter is the back born of coming up with a sound and organized analysis. This chapter presented the following: research design, population and sample, plans for data analysis and ethical consideration.

# **CHAPTER FOUR**

## **DATA PRESENTATION AND ANALYSIS**

### **4.0 Introduction**

Data presentation and analysis will present statistical output to help attain the objectives of the study. The first phase will identify explanatory variables that are significant. The second phase will build a discriminating analysis model then lastly classification and testing the discriminant analysis assumptions.

### **4.1 Stepwise Discriminant Analysis**

Variable selection is of efficacy important in multivariate analysis and regression, to find the important variables the researcher used SPSSv20. Stepwise discriminant analysis was used to select variables that maximize group separation between means of projected classes and variables that minimize variance within each class (between class scatter).

The use of stepwise methodology has been sharply criticized by several researchers, yet its popularity continues unabated. This method is of conspicuous importance in variable selection. In order to come up with better predicting variables there is need to put several variables in analysis to see which one(s) contribute to group discrimination. The researcher chose to use stepwise because of the advantages and flexibility that it offers. This technique involves entering the independent variables into the discriminant function one at a time on the basis of their discriminant power. Specifically, at each step all variable are reviewed and evaluated to determine which one will contribute most to the discrimination between groups (Kachigan, 1986).

In a stepwise discriminant analysis, it limits the interpretation of relationships between independent variables and groups defined by the dependent variable to those independent variables that met the statistical test for inclusion in the analysis.

**Table 4.1.1: Analysis case processing summary**

| Analysis Case Processing Summary |   |            |              |
|----------------------------------|---|------------|--------------|
| Unweighted Cases                 | ...   | N          | Percent      |
| Valid                            |   | 100        | 100.0        |
| Excluded                         | Missing or out-of-range group codes   | 0          | 0.0          |
|                                  | At least one missing discriminating variable  | 0          | 0.0          |
|                                  | Both missing or out-of-range group codes and at least one missing discriminating variable | 0          | 0.0          |
|                                  | Total   | 0          | 0.0          |
| <b>Total</b>                     |   | <b>100</b> | <b>100.0</b> |

The analysis case processing summary gives us the dataset output in terms of the validity, excluded cases and the total cases. The excluded row verifies the data set to check for values out of range group codes. All qualitative variables were converted to numeric through coding and expected ranges were put in place for continuous variables to detect and eradicate outliers. There is no case excluded otherwise SPSS output would have given the reason for exclusion. The minimum ratio of valid cases to independent variables for discriminant analysis is 5 to 1. In this analysis, there are 100 valid cases and 7 independent variables. The ratio of cases to independent variables is 14.29 to 1, which satisfies the minimum requirement.

**Table 4.1.2: Group Statistics**

| Group Statistics      |                             |       |                |                    |          |
|-----------------------|-----------------------------|-------|----------------|--------------------|----------|
| Discharge Destination | ...                         | Mean  | Std. Deviation | Valid N (listwise) | ....     |
|                       |                             |       |                | Unweighted         | Weighted |
| Dead                  | Gender                      | 1.60  | .498           | 30                 | 30.000   |
|                       | Patient Age                 | 61.93 | 15.299         | 30                 | 30.000   |
|                       | Distance                    | 28.07 | 9.674          | 30                 | 30.000   |
|                       | Referral status             | .73   | .450           | 30                 | 30.000   |
|                       | Site of residence           | 2.37  | .765           | 30                 | 30.000   |
|                       | Disease comorbidity         | 1.47  | .681           | 30                 | 30.000   |
|                       | Number of reported symptoms | 2.77  | 1.478          | 30                 | 30.000   |
| Discharged            | Gender                      | 1.56  | .500           | 70                 | 70.000   |
|                       | Patient Age                 | 30.39 | 14.447         | 70                 | 70.000   |
|                       | Distance                    | 11.83 | 7.993          | 70                 | 70.000   |
|                       | Referral status             | .19   | .392           | 70                 | 70.000   |
|                       | Site of residence           | 2.13  | .700           | 70                 | 70.000   |
|                       | Disease comorbidity         | 2.46  | .879           | 70                 | 70.000   |
|                       | Number of reported symptoms | 3.30  | 1.334          | 70                 | 70.000   |
| Total                 | Gender                      | 1.57  | .498           | 100                | 100.000  |
|                       | Patient Age                 | 39.85 | 20.619         | 100                | 100.000  |
|                       | Distance                    | 16.70 | 11.308         | 100                | 100.000  |
|                       | Referral status             | .35   | .479           | 100                | 100.000  |
|                       | Site of residence           | 2.20  | .725           | 100                | 100.000  |
|                       | Disease comorbidity         | 2.16  | .940           | 100                | 100.000  |
|                       | Number of reported symptoms | 3.14  | 1.393          | 100                | 100.000  |

The group statistics gives the measures of central tendency of explanatory variables, unweighted and weighted categorized by factors of the dependent variable. Examination of whether there is any difference between groups on each of the independent variables using group means is of substantial contribution towards detection of variable discriminating power. The group statistics helps us to see whether it is worthwhile proceeding any further with the analysis. A rough idea of important variables is detected as a result of looking at the group means and standard deviations. If group size is equal, the cut-off is mean score but since group size is unequal the cut-off is calculated from weighted means.

In this research categorization is based on two groups death as “0” and discharged as “1”. The mean and the standard deviation were used as measures of central tendency, the mean show group separation in the factors of the dependent variable. Since gender, referral status, disease comorbidity, site of residence are dichotomous variables, the mean is not directly interpretable. Its interpretation takes into account the coding done during data entry. The mean for gender, site of residence, number of symptoms are not significantly different between the two groups of the variable discharge destination.

Age, distance, referral status, disease comorbidity shows a great difference in the mean between the groups. To interpret referral status and disease comorbidity, visualization of the coding makes it easy to interpret the data. Referral status was based on whether a patient has been referred by a clinic to the hospital or not. It was coded as follows “0” for NO and “1” for YES. The mean values 0.73 and 0.17 shown respectively indicates that most of the patients who were referred to the hospital died ( $0.73 \approx 1$ ) and those coming to the hospital directly from their homes where discharged ( $0.17 \approx 0$ ). For disease comorbidity it was coded as “1” Malaria and HIV/AIDS, “2” Malaria and other disease, “3” Malaria only. The mean values of 1.47 and 2.46 respectively shows that most of the patients who died had a combination of malaria an HIV/AIDS even though the interpretation looks robust.

**Table 4.1.3 Variables in analysis**

| Variables in the Analysis |                     |           |             |               |
|---------------------------|---------------------|-----------|-------------|---------------|
| Step                      | ...                 | Tolerance | F to Remove | Wilks' Lambda |
| 1                         | Patient Age         | 1.000     | 96.667      |               |
| 2                         | Patient Age         | 1.000     | 55.485      | .563          |
|                           | Distance            | 1.000     | 39.449      | .503          |
| 3                         | Patient Age         | .994      | 49.368      | .458          |
|                           | Distance            | .999      | 32.514      | .405          |
|                           | Referral status     | .994      | 17.614      | .358          |
| 4                         | Patient Age         | .994      | 36.973      | .347          |
|                           | Distance            | .979      | 33.784      | .339          |
|                           | Referral status     | .950      | 22.432      | .309          |
|                           | Disease comorbidity | .938      | 19.858      | .302          |

The table above shows the steps SPSS went through and variables that were included in the model at each step to improve group separation. The criteria that was used to include a variable in the model: a variable with an F value less than 2.71 to be excluded while variables with  $F > 3.84$  to be included in the model. Tolerance is the proportion of a variable's variance not accounted for by other independent variables in the equation. A variable with low tolerance contributes little information to a model and can cause computational problems that is why other variables were not in analysis. F to remove values are useful for describing what happens if a variable is removed from the current model given that the other variables remain. Patient age was the first best single predictor to be entered in the model in step 2 distance was added as the next best one then referral status and finally disease comorbidity.

**Table 4.1.4 Variables entered/removed**

| Variables Entered/Removed <sup>a,b,c,d</sup> |                     |               |     |     |        |           |     |        |      |  |
|--|---------------------|---------------|-----|-----|--------|-----------|-----|--------|------|--|
| Step   | Entered             | Wilks' Lambda |     |     |        |           |     |        |      |  |
|  |                     | Statistic     | df1 | df2 | df3    | Exact F   |     |        |      |  |
|  |                     |               |     |     |        | Statistic | df1 | df2    | Sig. |  |
| 1  | Patient Age         | .503          | 1   | 1   | 98.000 | 96.667    | 1   | 98.000 | .000 |  |
| 2  | Distance            | .358          | 2   | 1   | 98.000 | 87.021    | 2   | 97.000 | .000 |  |
| 3  | Referral status     | .302          | 3   | 1   | 98.000 | 73.822    | 3   | 96.000 | .000 |  |
| 4  | Disease comorbidity | .250          | 4   | 1   | 98.000 | 71.207    | 4   | 95.000 | .000 |  |

At each step, the variable that minimizes the overall Wilks' Lambda is entered.

a. Maximum number of steps is 14.

b. Minimum partial F to enter is 3.84.

c. Maximum partial F to remove is 2.71.

d. F level, tolerance, or VIN insufficient for further computation.

In stepwise discriminant analysis, the interpretations are limited to independent variable predictors listed as statistically significant otherwise non-significant variables are removed in each iteration. The table of variables entered/removed shows all the variables entered where significant as shown above. The stepwise method of variable selection identified at each step variables that satisfied the level of significance of 0.05.

**Table 4.1.5 ANOVA***Univariate ANOVA summaries*

| <i>Variable</i>    | <i>Model MS</i> | <i>Resid MS</i> | <i>Total MS</i> | <i>R-sq</i> | <i>Adj.<br/>R-sq</i> | <i>F</i> | <i>Pr &gt; F</i> |
|--------------------|-----------------|-----------------|-----------------|-------------|----------------------|----------|------------------|
| <i>gender</i>      | .03857143       | 24.471429       | 24.224632       | 0.0016      | -0.0086              | .15447   | 0.6952           |
| <i>age</i>         | 20900.298       | 21188.452       | 21185.542       | 0.4966      | 0.4914               | 96.667   | 0.0000           |
| <i>distance</i>    | 5537.1905       | 7121.8095       | 7105.8033       | 0.4374      | 0.4317               | 76.195   | 0.0000           |
| <i>referral_~s</i> | 6.297619        | 16.452381       | 16.349808       | 0.2768      | 0.2694               | 37.512   | 0.0000           |
| <i>Site_of_r~e</i> | 1.1904762       | 50.809524       | 50.308321       | 0.0229      | 0.0129               | 2.2962   | 0.1329           |
| <i>disease_c~y</i> | 20.601905       | 66.838095       | 66.371063       | 0.2356      | 0.2278               | 30.207   | 0.0000           |
| <i>Number_of~s</i> | 5.9733333       | 186.06667       | 184.24754       | 0.0311      | 0.0212               | 3.1461   | 0.0792           |

*Number of obs = 100      Model df = 1      Residual df = 98*

An anova table was run using Stata to check on the candidate variables in analysis, The R square value shows the amount of variation in the dependent that is being explained by the variables. Age, distance, referral status, disease comorbidity had a greater contribution to the R square value. The variables gender, site of residence were excluded in analysis because of not being statistically significant and their weak R square values.



**Table 4.1.6 Variables not in analysis**

| Variables Not in the Analysis |                             |           |                |            |               |
|-------------------------------|-----------------------------|-----------|----------------|------------|---------------|
| Step                          | Variable                    | Tolerance | Min. Tolerance | F to Enter | Wilks' Lambda |
| 0                             | Gender                      | 1.000     | 1.000          | .154       | .998          |
|                               | Patient Age                 | 1.000     | 1.000          | 96.667     | .503          |
|                               | Distance                    | 1.000     | 1.000          | 76.195     | .563          |
|                               | Referral status             | 1.000     | 1.000          | 37.512     | .723          |
|                               | Site of residence           | 1.000     | 1.000          | 2.296      | .977          |
|                               | Disease comorbidity         | 1.000     | 1.000          | 30.207     | .764          |
|                               | Number of reported symptoms | 1.000     | 1.000          | 3.146      | .969          |
| 1                             | Gender                      | .999      | .999           | .002       | .503          |
|                               | Distance                    | 1.000     | 1.000          | 39.449     | .358          |
|                               | Referral status             | .994      | .994           | 23.628     | .405          |
|                               | Site of residence           | .998      | .998           | .589       | .500          |
|                               | Disease comorbidity         | 1.000     | 1.000          | 14.017     | .440          |
|                               | Number of reported symptoms | .988      | .988           | 4.039      | .483          |
| 2                             | Gender                      | .990      | .990           | .274       | .357          |
|                               | Referral status             | .994      | .994           | 17.614     | .302          |
|                               | Site of residence           | .998      | .998           | .389       | .356          |
|                               | Disease comorbidity         | .981      | .981           | 15.123     | .309          |
|                               | Number of reported symptoms | .983      | .983           | 4.263      | .343          |
| 3                             | Gender                      | .990      | .990           | .136       | .302          |
|                               | Site of residence           | .990      | .985           | .858       | .300          |
|                               | Disease comorbidity         | .938      | .938           | 19.858     | .250          |
|                               | Number of reported symptoms | .983      | .983           | 3.248      | .292          |
| 4                             | Gender                      | .973      | .923           | .748       | .248          |
|                               | Site of residence           | .978      | .927           | .161       | .250          |
|                               | Number of reported symptoms | .968      | .924           | 1.292      | .247          |

The table above shows variables not in analysis at any given step, the stepwise method starts with a model that does not include any of the predictors (step 0). At each step the predictors with the smallest F to Enter value is removed from the model. In step 1 the complement of patient age was not in analysis and some of the variables were included in the analysis based on their contribution to group discrimination until the final step 4 where gender, site of residence, number of reported symptoms were not in analysis and automatically rejected as significant variables.

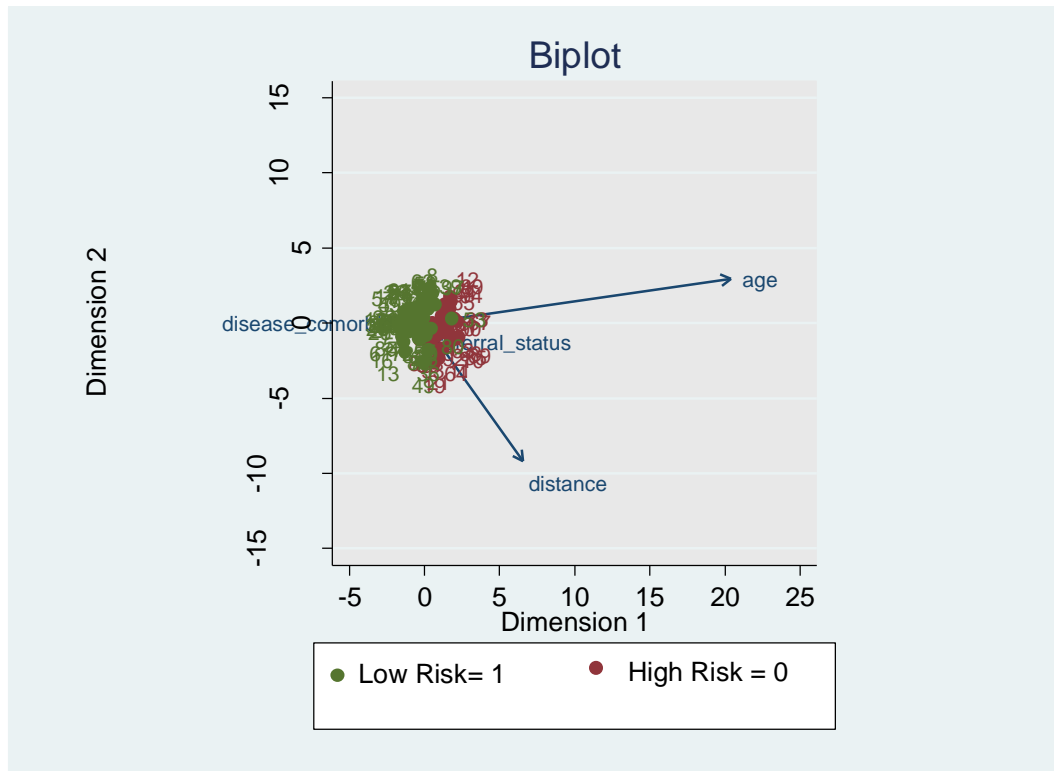
**Table 4.1.7 Standardized canonical discriminant function coefficients.**

**Standardized Canonical  
Discriminant Function  
Coefficients**

|                     | Function |
|---------------------|----------|
|                     | 1        |
| Patient Age         | .613     |
| Distance            | .598     |
| Referral status     | .518     |
| Disease comorbidity | -.496    |

The standardized canonical discriminant function coefficients show the ratings of explanatory successful candidate variables. The interpretation of weights is based on finding the absolute function value for each variable. The sign shows the direction of the relationship. Patient age score was the strongest predictor while distance was next in importance and referral status and disease comorbidity were less successful as predictors.

**Figure 4.1 Biplot**

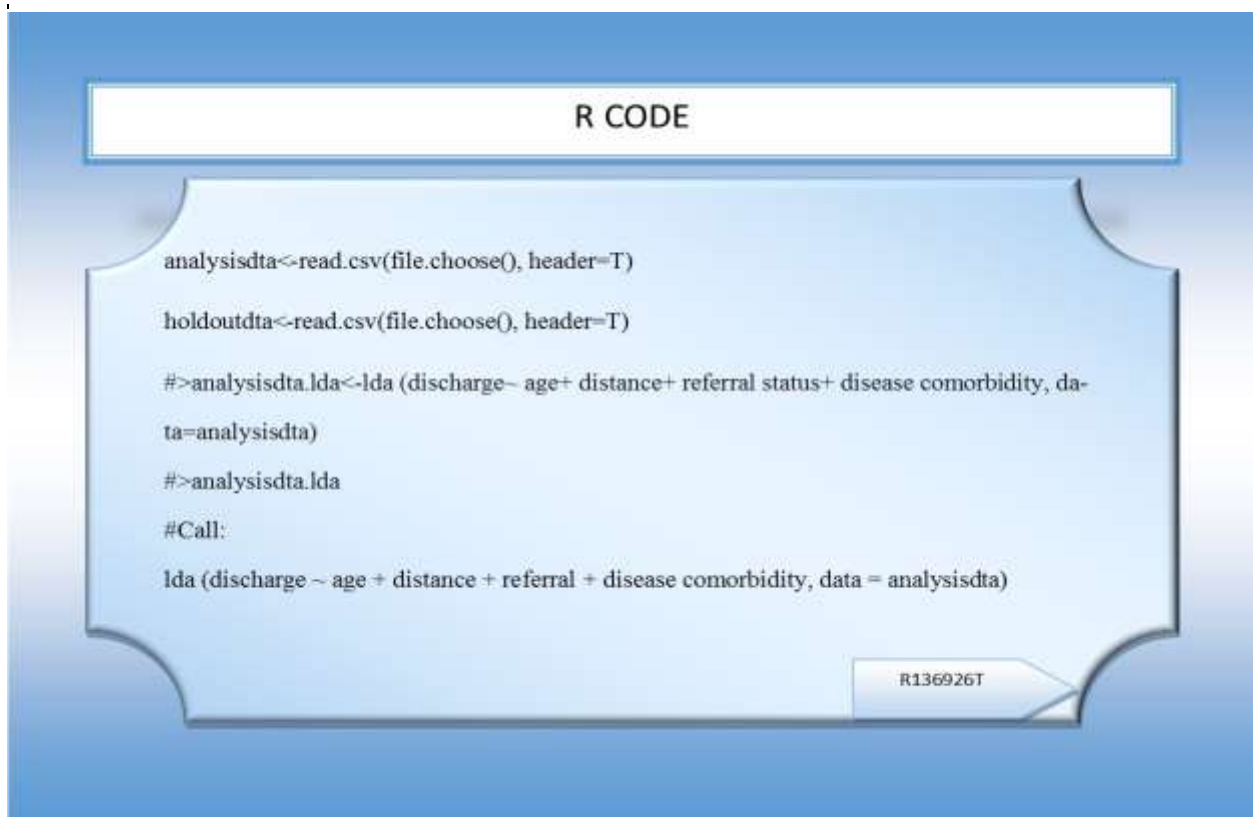


Biplot is an enhanced scatterplot that uses vectors to represent a hyper dimensional structure on (x, y) plain. It is extensively used in discriminant analysis to visualize group separation. The axis are a pair or combination of explanatory variables. The explanatory variables used to separate the risk groups shows an acceptable distinction even if there are some kind of overlapping's. This a clear indication that some observations will be misclassified but having a small proportion of wrong group prediction. The variables where first standardized to ensure that scale difference between variables are limited. Standardization is of efficacy importance in ranking variables based on their discriminating capability. Explanatory variables with high weights are mostly those that contribute to group separation that is age and distance.

## 4.2 Fisher discriminant function.

The fisher discriminant function was used to come up with a classification model using the R programming. Data entry and data cleaning was done using SPSS. The data was then converted to an CSV file for readability purpose in R. The sample consisted of 150 patients: 100 for estimation (analysisdta) and 50 for model validation (holdoutdta).

### Code snippet for lda.

A screenshot of an R code editor window with a blue background. At the top, there is a white header bar with the text "R CODE" in black. Below the header, the R code is displayed in a light blue box with rounded corners. The code consists of several lines: two lines for reading CSV files, a line for creating an LDA model, a line for displaying the model, a comment line, and a line for calling the LDA function. In the bottom right corner of the code box, there is a small white box with the text "R136926T" and a right-pointing arrow.

```
analysisdta<-read.csv(file.choose(), header=T)

holdoutdta<-read.csv(file.choose(), header=T)

#>analysisdta.lda<-lda (discharge~ age+ distance+ referral status+ disease comorbidity, da-
ta=analysisdta)

#>analysisdta.lda

#Call:

lda (discharge ~ age + distance + referral + disease comorbidity, data = analysisdta)
```

**Table 4.2.1 Prior probabilities of groups**

| Prior probabilities of groups: |              |
|--------------------------------|--------------|
| "0":High risk                  | "0":Low risk |
| 0.3                            | 0.7          |

The prior probabilities of groups are calculated from the sample values. The analysis sample of c onsisted of 100 admitted malaria patients, 30 from the death group also named high risk and 70 fr om the discharge group also named low risk.

**Table 4.2.2 Group means**

| Group means:  |          |          |           |                     |
|---------------|----------|----------|-----------|---------------------|
| Risk          | age      | distance | referral  | disease comorbidity |
| "0" High risk | 61.93333 | 28.06667 | 0.7333333 | 1.466667            |
| "1" Low risk  | 30.38571 | 11.82857 | 0.1857143 | 2.457143            |

The table above shows descriptive statistics of the mean, that where produced after running the code discriminant analysis. The group means highlighted above are for significant variables that where selected as being more discriminating than other using stepwise discriminant analysis.

The four successful candidate variables show great mean difference between the groups, which is one of the greatest contributing factors that made the explanatory variables to be included in the model.

**Table 4.2.3 Coefficients of linear discriminants**

| Coefficients of linear discriminants: ▾ |             |
|---|-------------|
| LD1                                     |             |
| age                                     | -0.04169461 |
| distance                                | -0.07012943 |
| referral                                | -1.26351693 |
| disease comorbidity                     | 0.60017641  |

The coefficients table is of high importance in discriminant analysis because it gives coefficients value of the model. The coefficient of linear discriminants produces discriminate equations (functions). The maximum number of discriminating functions produced is the number of groups minus one. In this research the dependent variable is binary namely 'low risk' and 'high risk' thus only LD1 is displayed in the table.

#### **Discriminant analysis model**

$$\begin{aligned} \text{Risk level} = & -0.04\text{Age} - 0.07\text{Distance} - 1.26\text{Referral status} \\ & + 0.6\text{Disease comorbidity} \end{aligned}$$

#### **Classification of new patients**

The other crucial part in discriminant analysis stage is classification of patients into predicted groups at the time of arrival at the hospital. Two newly malaria patients were selected at Sanyati Baptist Hospital to predict whether a patient is in the low or high risk category. R software was used to make the predictions, since the data was coded as "0" =high risk, "1" =low risk.

### **The code used in R to generate output:**

For classification purpose two newly admitted malaria patients were taken from the data set of 2017. To see the validity of the model and the classification power of the model to the current newly admitted patients. The results were impressive, all the patients were correctly classified to their actual discharge destination

### **Code snippet for classifying new patients**

A patient of age 31, who lives 21km away from the hospital, who had been referred by a certain clinic, with disease comorbidity of 3 that is infected by Malaria only was predicted to the low risk group:

```
##patient=lda (discharge~ age+ distance+ referral+ disease comorbidity, data=analysisdta)

##predict(patient,newdata=data.frame(age=31,distance=21,referral=1,diseascom=
3))$class

#The first patient was predicted to be in the low risk class.
#R output
[1] 1
Levels: 0 1
```

A patient of age 64, who lives 41km away from the hospital, who had been referred by a certain clinic, with disease comorbidity of 2 that is infected by Malaria and other disease was predicted to the high risk group:

### **Code snippet for R**

```
## predict(patient,newdata=data.frame(age=64,distance=41,referral=1,diseascom
=2))$class

#The second patient was predicted to be in the high risk class.
#R output

[1] 0
Levels: 0 1
```

### Code snippet for model fitness and validation

```
##model fitness and cross validation  
  
#analysisdta.lda.p<-predict(analysisdta.lda, holdoutdta[,c(3,4,5,7)])$class  
  
table(analysisdta.lda.p, holdoutdta[,1])
```

**Table 4.2.4 Misclassifications table**

| Index | Actual | Predicted |
|-------|--------|-----------|
| [1]   | 0      | 0         |
| [2]   | 0      | 0         |
| [3]   | 1      | 1         |
| [4]   | 1      | 1         |
| [5]   | 1      | 1         |
| [6]   | 1      | 1         |
| [7]   | 1      | 1         |
| [8]   | 1      | 1         |
| [9]   | 1      | 1         |
| [10]  | 1      | 1         |
| [11]  | 1      | 1         |
| [12]  | 0      | 1         |
| [13]  | 0      | 0         |
| [14]  | 0      | 0         |
| [15]  | 0      | 0         |
| [16]  | 0      | 0         |
| [17]  | 0      | 0         |
| [18]  | 1      | 0         |
| [19]  | 0      | 1         |
| [20]  | 1      | 1         |
| [21]  | 1      | 1         |
| [22]  | 1      | 1         |
| [23]  | 1      | 1         |
| [24]  | 1      | 1         |
| [25]  | 1      | 1         |
| [26]  | 1      | 1         |
| [27]  | 1      | 1         |
| [28]  | 1      | 1         |
| [29]  | 1      | 1         |
| [30]  | 1      | 1         |
| [31]  | 1      | 1         |
| [32]  | 1      | 1         |
| [33]  | 1      | 1         |
| [34]  | 1      | 1         |
| [35]  | 1      | 0         |

KEY: Levels: "0" High risk "1" Low risk

One the aspects to consider after completing the procedure of model building is to examine misclassified observations. By examining the table above observations 12, 18 and 39 were actually in the “0” high risk but where predicted to be in “1” low risk. The opposite goes for observation 29 that was predicted to group 0 when actually in group 1.



**Table 4.2.5 Confusion matrix**

| CONFUSION MATRIX |           | Actual        |              |           |
|------------------|-----------|---------------|--------------|-----------|
|                  |           | “0” High Risk | “1” Low Risk | $\Sigma$  |
| “0” High Risk    | Predicted |               |              |           |
|                  |           | <b>12</b>     | <b>1</b>     | <b>13</b> |
| “1” Low Risk     | Predicted |               |              |           |
|                  |           | <b>3</b>      | <b>34</b>    | <b>37</b> |
| $\Sigma$         |           | <b>15</b>     | <b>35</b>    | <b>50</b> |

In order to test the significance of the model a holdout sample of 50 observations was taken to classify malaria patients. This matrix shows how the model correctly and misclassify patients. The diagonal matrix shows the correct classification and the off-diagonal represents misclassification. The high frequency in the main diagonal shows that the forecasting power of our model is good. The output in R has got the columns as “Actual” and rows as “Predicted”. Sensitivity and specificity are frequently used in medical terminology for assessing the quality of diagnostic test.

**Calculations**       $Sensitivity = \frac{TP}{TP+FN} = \frac{TP}{POS} = \frac{34}{35} = 97,14\%$

$$Specificity = \frac{TN}{TN+FP} = \frac{TN}{NEG} = \frac{12}{15} = 80\%$$

If a patient is in the lower risk, then the test will confirm the severity of the illness with probability of 0.9714.

If the patient is in the high risk, then the test will with probability of 0.2 wrongly classify a patient in the high risk category.

The application of sensitivity and specificity is important in evaluating the classification and misclassification of observations in the confusion matrix. Highly sensitive test means that there are few false negative results (Type II error). Low risk patients had a few false negative result of 2.86%, that is there is a probability of 0.0286 of wrongly predicting that a patient is going to survive when in fact he/she is at high risk of losing life. High specificity test means that there are few false positive results (Type I error). This is when a high risk patient is classified to the low risk group with a false positive probability of 0.2 of wrongly predicting that a patient is going to be discharged.

### **Press's Q statistic**

Press's Q statistic compares the number of correct classifications with the total sample size and number of groups. The calculated value is then compared with the critical value from the Chi-Square distribution table with 1 degree of freedom and  $\alpha=0.01$ .

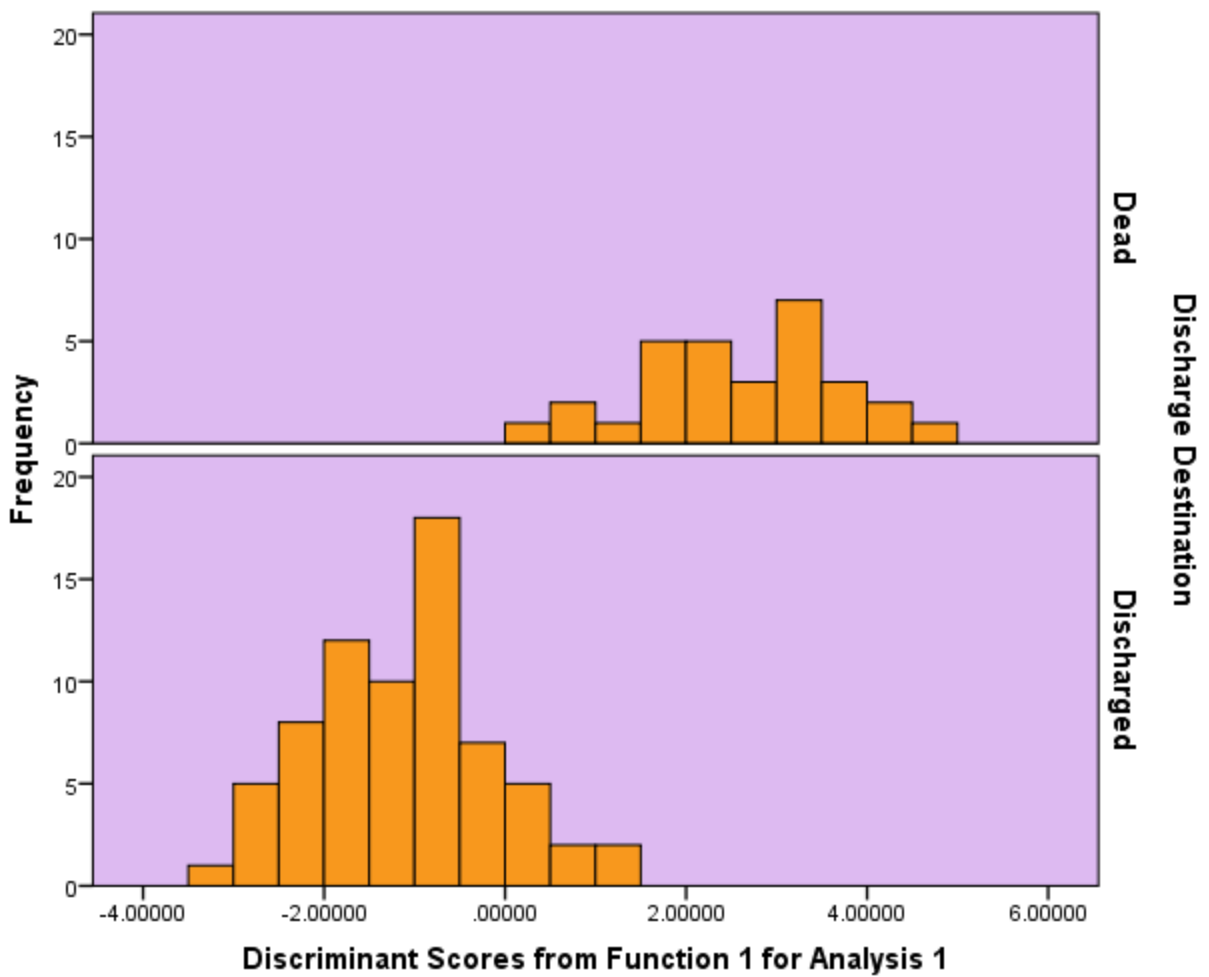
### Calculation of the statistic

$$\text{Press's } Q_{\text{holdout sample}} = \frac{[50 - (46 \times 2)]^2}{50 \times (2 - 1)} = 35.28$$

Which is greater than our critical value of 6.65. Therefore, our results exceed the classification accuracy expected by chance at a statistically significant level.

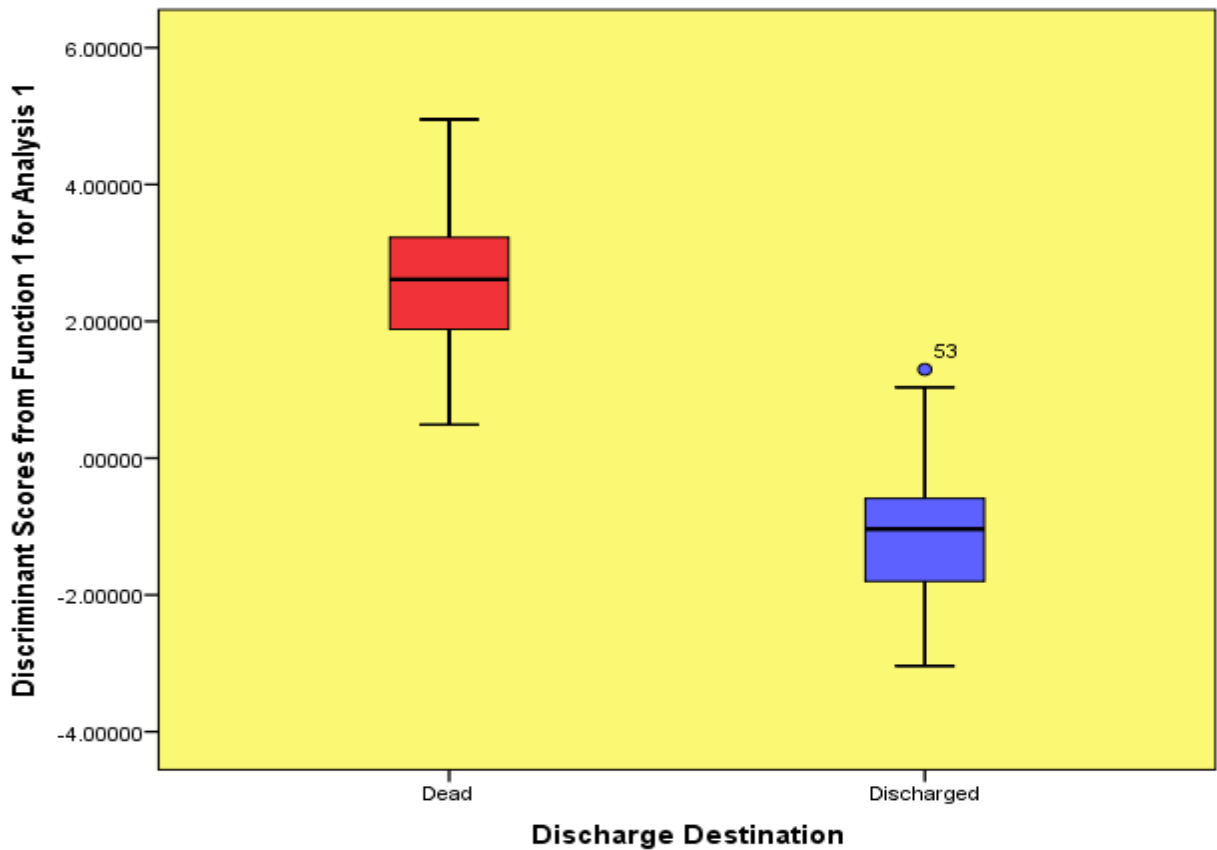
### 4.3 Classifications and diagnostic testing

Figure 4.2 Histogram showing the distribution of discriminant scores.



The graph above shows separate group plots of the dependent variable and the length of overlapping. The visualization of the graph shows that the two classes do not overlap too much which is an indication of a good discriminant function. These histograms show the distribution of the discriminant scores for the dead named high risk and discharged patients named low risk. In order to interpret the axis noting of group centroid table is important the means where as follows dead = 2.618 and discharged -1.122.

**Figure 4.3 Boxplot**



Box plot illustrating the distribution of discriminant scores for the dead and discharge group.

The box plot is useful in visual demonstration of the effectiveness of the discriminant function. It is an alternative to the histogram in Figure 4.2 in illustrating the distribution of the discriminant function score each group.

**Table 4.3.1 Eigenvalues table**

**Eigenvalues**

| Function | Eigenvalue         | % of Variance | Cumulative % | Canonical Correlation |
|----------|--------------------|---------------|--------------|-----------------------|
| 1        | 2.998 <sup>a</sup> | 100.0         | 100.0        | .866                  |

a. First 1 canonical discriminant functions were used in the analysis.

The canonical correlation is the multiple correlation between the predictors and the discriminant function. It provides an index of the model fit which is interpreted as being the proportion of variance explained ( $R^2$ ). The larger the eigenvalue, the more of the variance in the dependent variable that is explained by the function. The canonical correlation is the measure of association between the discriminant function and the variable of interest. The table shows a canonical correlation of 0.866 which suggest that 75% of the variation in the grouping variable.

**Table 4.3.2 Wilks' Lambda table**

**Wilks' Lambda**

| Test of Function(s) | Wilks' Lambda | Chi-square | df | Sig. |
|---------------------|---------------|------------|----|------|
| 1                   | .250          | 133.041    | 4  | .000 |

Wilk's Lambda indicates the significance of the discriminant function. It is a measure of how well each function separates cases into groups and mostly used in variable section when using stepwise discriminant analysis. The Wilks' Lambda of 25% shows the variation unexplained by the model and this smaller value of 0.25 indicate greater discriminatory ability of the function. The associated

chi-square statistic tests the hypothesis that the means of the functions listed are equal across groups. The p value is less than 0.05 which shows that the model is statistically significant.

#### 4.3.1 Model validity using SPSS and Stata13

**Table 4.3.3 Classification results table using SPSS.**

**Classification Results<sup>a,c</sup>**

|                              |       |            | Predicted Group Membership |            | Total |
|------------------------------|-------|------------|----------------------------|------------|-------|
|                              |       |            | Dead                       | Discharged |       |
| Original                     | Count | Dead       | 29                         | 1          | 30    |
|                              |       | Discharged | 3                          | 67         | 70    |
|                              | %     | Dead       | 96.7                       | 3.3        | 100.0 |
|                              |       | Discharged | 4.3                        | 95.7       | 100.0 |
| Cross-validated <sup>b</sup> | Count | Dead       | 29                         | 1          | 30    |
|                              |       | Discharged | 3                          | 67         | 70    |
|                              | %     | Dead       | 96.7                       | 3.3        | 100.0 |
|                              |       | Discharged | 4.3                        | 95.7       | 100.0 |

a. 96.0% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

c. 96.0% of cross-validated grouped cases correctly classified.

Maximum likelihood technique was used to assign a patient to a group. When SPSS runs the discriminant analysis it produces a table of classification of results, a new data set is generated using the same sample to classify observations. It gives information about actual group membership versus predicted group membership.

Overall % correctly classified (hit ratio) = 96%.

Sensitivity =  $29/30 = 96.7\%$ .

Specificity =  $67/70 = 95.7\%$

**Table 4.3.4 Classification table using Stata.**

| True discharge | Classified  |             | Total         |
|----------------|-------------|-------------|---------------|
|                | 0           | 1           |               |
| 0              | 29<br>96.67 | 1<br>3.33   | 30<br>100.00  |
| 1              | 2<br>2.86   | 68<br>97.14 | 70<br>100.00  |
| Total          | 31<br>31.00 | 69<br>69.00 | 100<br>100.00 |
| Priors         | 0.5000      | 0.5000      |               |

Stata also produced similar classifications as that in SPSS, only two patients were wrongly classified out of 70 patients. In other words, we can say that 1 in every 30 newly admitted malaria patients is classified to the wrong class, that is a high risk patient may be predicted to be in the low risk group. Also 2 in every 70 low risk patients may be wrongly classified to the high risk category.

#### 4.3.2 Diagnostics Checking

**Table 4.3.5 Correlation for table for multi collinearity test**

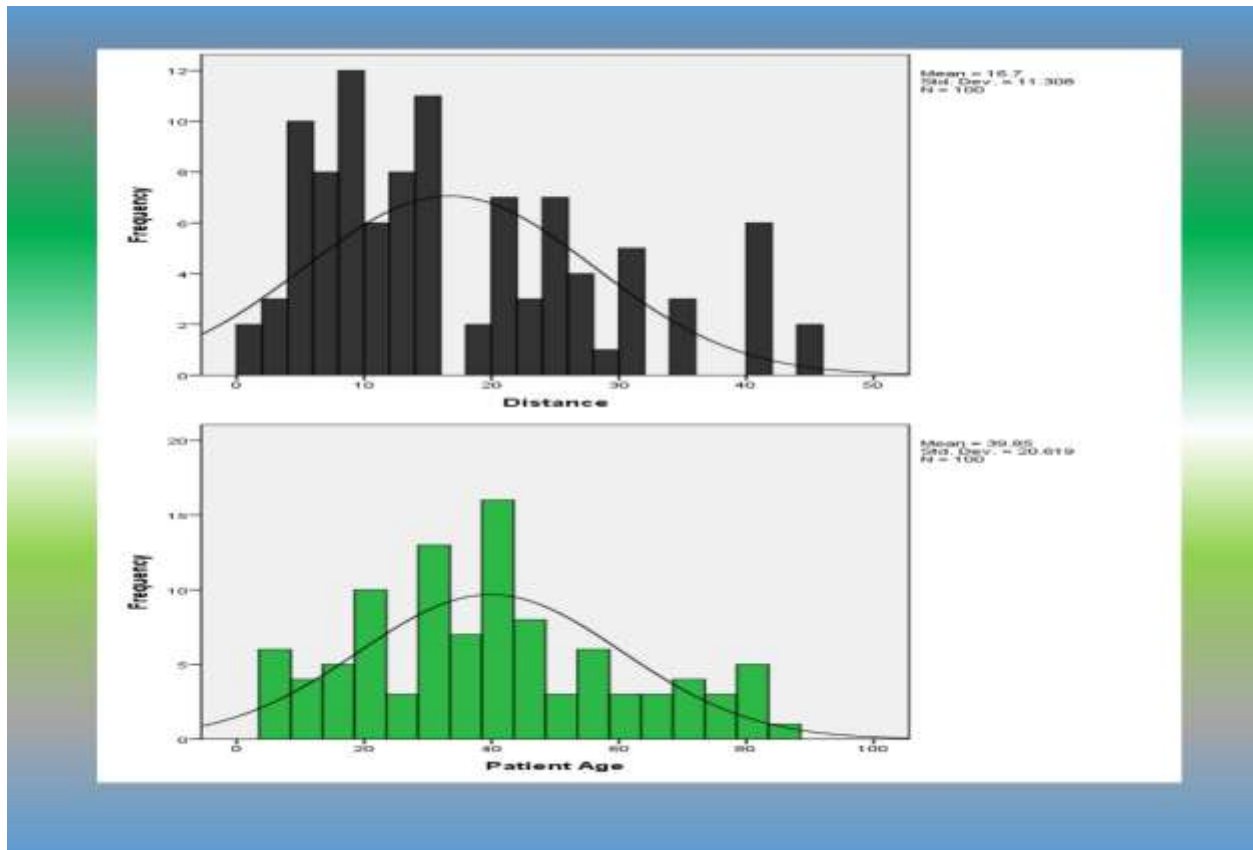
|                   |
|-------------------|
| Key               |
| Correlation       |
| Two-sided p-value |

|           | gender              | age                 | distance            | site                | referral            | diseas-m           | symptoms |
|-----------|---------------------|---------------------|---------------------|---------------------|---------------------|--------------------|----------|
| gender    | 1.00000             |                     |                     |                     |                     |                    |          |
| age       | 0.06809<br>0.57543  | 1.00000             |                     |                     |                     |                    |          |
| distance  | 0.02061<br>0.86555  | 0.08003<br>0.51016  | 1.00000             |                     |                     |                    |          |
| site      | -0.20740<br>0.08493 | 0.12251<br>0.31231  | -0.01154<br>0.92446 | 1.00000             |                     |                    |          |
| referral  | 0.05600<br>0.64522  | -0.04102<br>0.73602 | -0.11931<br>0.32525 | -0.08831<br>0.46726 | 1.00000             |                    |          |
| diseascom | 0.13739<br>0.25671  | 0.00075<br>0.99509  | 0.16181<br>0.18080  | -0.19092<br>0.11337 | 0.21276<br>0.07700  | 1.00000            |          |
| symptoms  | -0.10210<br>0.40034 | 0.06312<br>0.60370  | 0.13135<br>0.27842  | 0.14432<br>0.23327  | -0.05272<br>0.66469 | 0.00494<br>0.96761 | 1.00000  |

One of the assumptions that needs to be met in discriminant analysis is multi collinearity. It shows the correlation within the predictors. When two explanatory variables are heading in the same direction it causes the coefficient to be unreliable. The correlation table above shows that multi collinearity is not present to a large extent as shown by correlations less than +/- 0.5.



**Table 4.5 Normality test**



The variables show that there are approximately normally distributed despite some a low level of departure.

**Table 4.3.6 Test for equality of variance**

**Log Determinants**

| Discharge Destination | Rank | Log Determinant |
|-----------------------|------|-----------------|
| Dead                  | 4    | 7.510           |
| Discharged            | 4    | 7.260           |
| Pooled within-groups  | 4    | 7.425           |

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

The table of log determinants test for homogeneity of covariance matrices. The rank column shows the number of explanatory variables in this analysis. Log determinants are 7.51 and 7.26 which are relatively equal, this is a clear indication that there is equality of variance.

#### **4.4 Conclusion**

This chapter analyzed data using a statistical technique called discriminant analysis approach. The software's that were utilized in this research are R programming, STATA and IBM SPSS.

# CHAPTER FIVE

## SUMMARY, CONCLUSION AND RECOMMENDATIONS

### 5.0 Introduction

The previous chapter analyzed data and the researcher came up with a discriminating model. To this effect having presented and analyzed data in chapter four. This final and last chapter will draw conclusion from the findings and make recommendations which will contribute to mortality reduction at Sanyati Baptist Hospital (SBH).

### 5.1 Summary

Hospitals are clearly an important context for end-of-life care, yet there are still difficulties in making the transition to palliative care and in implementing interventions for the imminently dying. In order for appropriate care plans to be made and delivered to malaria patients, there is a need for SBH to adopt a more vigorous approach to identifying patients who are entering last moments of their lives and do the best to increase the chances of patient survival.

The classification results showed that the model does richly classifying malaria patients with accuracy. A holdout sample of 50 was used to test the validity of the model the results was as follows 92% correct classification of patient. This percentage shows that the classification model is reliable as far as predictions are concerned. SPSS was also used to validate the model with an overall classification of 96% and 97% using Stata. This two software's produced a high inflated classification percentage as compared with the that from R. The major reason being that SPSS and Stata used the same sample (analysis) to classify new cases whilst R used the holdout sample.

The probabilities found in this project is a clear indication that we cannot predict group membership with certainty due to some factors beyond our control. If an interested part tries to predict a patient to the correct group and fails obtain the presaged result, one should not be quick to turn a deaf ear to the hypothesis since there is a probability of less than 0.08 of wrong classification.

The pharmacy department is responsible for dispensing drugs to the hospital wards for inpatient use. It also pays regular visits to wards to strengthen drug management but there is no criteria in place to efficiently dispense those malaria drugs to patients. A randomized method is being used in distributing drugs haphazardly. As a result of such kind of dispensing stock out will be occur result to high risk patient not getting any drug. Provincial medical stores used to supply drugs constantly but now they no longer supply the drugs.

The Central Buying Unit (CBU) is the one responsible for sourcing quotations through the tendering process to look for the cheapest supplier of medicine but with maximum quality so that the procurement of drugs might be facilitated smoothly. The upper board that is Procurement Tender Committee PTC is responsible for the procurement of the drugs and there are falling to by adequate malaria drugs sometimes the reason being limited financial capacity. The Stores department is responsible for receiving supplies from different suppliers including drugs and then issue them to the pharmacy. If the central buying unit, pharmacy, stores department work together diligently better drug allocation and management may help low drug shortage thereby contributing to malaria mortality reduction.

Due to the financial difficulties being faced by the hospital drugs are bought in short quantities to the extent that stock may run out for about 3 days without any drugs and patients end up being given pain killers and other temporary drug treatment. Scarcity of resources is a key driving problem that has aroused the writing of this report.

## **5.2 Recommendations**

The hospital has to implement a new strategic system for priority scheduling and distribution of patients. Due to scarcity of drugs at Sanyati hospital there is need to distribute drugs and services to patients first preference being given to high risk patient then to the low risk group. The system currently in place have to be changed, that is first come first serve because most of the high risk patients admitted at a later stage may quickly die. When a patient is admitted it is necessary to first categorize patients in the low or high risk group. This will help health practitioner to device best treatment to give to such kind of patients.

The model is there to optimize drug allocation efficiently and find an optimal solution in this environment of limited resources. High risk patients' needs to be given the first preference in drug distribution as compared to low risk patients. Currently there is only one doctor at the hospital. Identification of those who are in the high risk group or critically ill will help in allocating the health practitioners effectively, letting the doctor take care of those patients first. Most of the time nurses are delegated to look after critical malaria patients without a close supervision by the doctor.

## **5.3 Variable summary and recommendation**

Prognostic factors that are age, distance, referral status, disease comorbidity are the candidate variables that were statistically significant. A discriminant model was found that predicts patients

to high or low risk group. This model will act as an aiding tool towards mortality reduction by proper and efficient way of allocating patients in their predicted predicaments. Drugs that are mostly available for treatment at Sanyati hospital are quinine sulfate and chloroquine some of the times there will not be available.

The discriminant model will classify patients into one of two predefined "risk level" groups, based on prognostic information from each malaria patient. Through the output of the structure matrix, identification of explanatory variables that are most useful for group prediction was made.

This research has shown that there are variables that contribute or discriminate between the high risk and the low risk groups. Age, distance, referral status, disease comorbidity were significant variables to group separation. Gender, site of residence and number of reported symptoms were insignificant.

The age of a patient was the greatest discriminating variable amongst all the variables. The high risk group showed a mean of 61,93 and 30,39 for the low risk class. This indicates that as the person grows older he/she is more likely to die as compared to someone of lower age. For every one additional year of a patient there is an associated probability of dying.

Distance was second in the list showing that those patients who stay furthest were more likely to die as compared with those within a close radius. In Sanyati there is poor road networking, the roads are full of potholes to the extent that transfer or movement of patients become problematic. Transport availability to move malaria patients to the hospital is in short supply. In other rural areas there will be only one vehicle of which in the event of vehicle breakdown it will result to difficulties of patient's movement especially to those patients that lives far away from the main

road. Sanyati is a rural area of which there is only one hospital in the district most of the people are small scale farmers and they are not financial stable as a result, monetary problems may delay patients quick access to the hospital.

The hospital has only two ambulances of which most of the times the other vehicle will not be readily available. It is used for carrying out hospital logistics such as out scheduled meetings, purchasing of drugs, equipment etc. The ambulances are few to the extent that they cannot suffice the whole district. Purchase of additional ambulances can be a major resolver towards coming with solutions to the long distance problem. The hospital mostly relies on donations for acquiring big assets like a motor vehicle. If it was possible to have a donation of additional ambulances, it would resolve the long distance problem. In the event that a patient gets sick of malaria it can quickly reach there in time and transport patients to the hospital. The cost of hiring an ambulance is high most probably within a continuum of \$20 to \$40 depending on distance to be covered of which most of the residences in Sanyati are not able to afford that price.

Referral status is also one of the prognostic factors selected as highly contributing to group separation. All the clinics in Sanyati uses the hospital as their immediate referral point. Some of the malaria patients first seek medical assistance from the nearby clinic. When a patient's condition is seen as critical or if the malaria drugs are not available the patient is then transferred to the hospital. The data analysis results reviewed that mostly those who are referred to the hospital are more likely to die as compared with those who uses the hospital as their first referral point.

Disease comorbidity was also crucial contributing factor towards group discrimination. When a patient is admitted of malaria there is a tendency that he/she may be suffering from other disease like HIV/AIDS, Hypertension, Gonorrhoea, Asthma, Tuberculosis etc. The researcher categorized this variable into three attributes that are "Malaria only". "Malaria and HIV/AIDS" and "Malaria

and other disease". This variable was seen as contributing factor in this research because it showed positive effect on the model. Malaria patients with HIV/AIDS were more likely to die as compared with those with malaria and other disease. Since the malaria parasite plasmodium falciparum damages the red blood cells sometimes because of reduced CD4 count or weak immune system malaria would capitalize and take advantage of that result to cerebral.

#### **5.4 Conclusion**

One of the aim of mathematics is to understand what is going on around us and since it was realized that mathematics is a language in which nature speaks to us. Even if we may try to predict patient survival not for the self-glorification of the mind but to try to make Sanyati Baptist hospital a better health facility. Sometimes because of our limited mathematical capacity there are some things we may not review because of restricted wisdom and also due to factors beyond our control. The Creator is the one who knows the end of life of every person.

## **APPENDICES**

### **Appendix A: Covariance matrix**



**Covariance Matrices<sup>a</sup>**

| Discharge Destination |                             | Gender | Patient Age | Distance | Referral status | Site of residence | Disease comorbidity | Number of reported symptoms |
|-----------------------|-----------------------------|--------|-------------|----------|-----------------|-------------------|---------------------|-----------------------------|
| Dead                  | Gender                      | .248   | -.338       | -1.524   | -.007           | -.021             | .021                | -.234                       |
|                       | Patient Age                 | -.338  | 234.064     | -29.202  | -.984           | -1.389            | -.830               | 4.467                       |
|                       | Distance                    | -1.524 | -29.202     | 93.582   | .639            | .216              | .520                | -.605                       |
|                       | Referral status             | -.007  | -.984       | .639     | .202            | -.037             | .060                | .005                        |
|                       | Site of residence           | -.021  | -1.389      | .216     | -.037           | .585              | .030                | .123                        |
|                       | Disease comorbidity         | .021   | -.830       | .520     | .060            | .030              | .464                | .457                        |
|                       | Number of reported symptoms | -.234  | 4.467       | -.605    | .005            | .123              | .457                | 2.185                       |
| Discharged            | Gender                      | .250   | .492        | .082     | .011            | -.073             | .060                | -.068                       |
|                       | Patient Age                 | .492   | 208.704     | 9.241    | -.232           | 1.240             | .010                | 1.216                       |
|                       | Distance                    | .082   | 9.241       | 63.883   | -.373           | -.065             | 1.137               | 1.400                       |
|                       | Referral status             | .011   | -.232       | -.373    | .153            | -.024             | .073                | -.028                       |
|                       | Site of residence           | -.073  | 1.240       | -.065    | -.024           | .490              | -.118               | .135                        |
|                       | Disease comorbidity         | .060   | .010        | 1.137    | .073            | -.118             | .773                | .006                        |
|                       | Number of reported symptoms | -.068  | 1.216       | 1.400    | -.028           | .135              | .006                | 1.778                       |
| Total                 | Gender                      | .248   | .531        | -.241    | .011            | -.055             | .039                | -.121                       |
|                       | Patient Age                 | .531   | 425.139     | 106.551  | 3.215           | 2.051             | -6.865              | -1.413                      |
|                       | Distance                    | -.241  | 106.551     | 127.869  | 1.813           | .838              | -2.467              | -1.038                      |
|                       | Referral status             | .011   | 3.215       | 1.813    | .230            | .000              | -.046               | -.080                       |
|                       | Site of residence           | -.055  | 2.051       | .838     | .000            | .525              | -.123               | .103                        |
|                       | Disease comorbidity         | .039   | -6.865      | -2.467   | -.046           | -.123             | .883                | .250                        |
|                       | Number of reported symptoms | -.121  | -1.413      | -1.038   | -.080           | .103              | .250                | 1.940                       |

a. The total covariance matrix has 99 degrees of freedom.

**Appendix B: Test of Equality of group means**

**Tests of Equality of Group Means**

|                             | Wilks' Lambda | F      | df1 | df2 | Sig. |
|-----------------------------|---------------|--------|-----|-----|------|
| Gender                      | .998          | .154   | 1   | 98  | .695 |
| Patient Age                 | .503          | 96.667 | 1   | 98  | .000 |
| Distance                    | .563          | 76.195 | 1   | 98  | .000 |
| Referral status             | .723          | 37.512 | 1   | 98  | .000 |
| Site of residence           | .977          | 2.296  | 1   | 98  | .133 |
| Disease comorbidity         | .764          | 30.207 | 1   | 98  | .000 |
| Number of reported symptoms | .969          | 3.146  | 1   | 98  | .079 |

**Appendix C: Pairwise group comparison**

**Pairwise Group Comparisons<sup>a,b,c,d</sup>**

| Step | Discharge Destination |      | Dead   | Discharged |
|------|-----------------------|------|--------|------------|
| 1    | Dead                  | F    |        | 96.667     |
|      |                       | Sig. |        | .000       |
|      | Discharged            | F    | 96.667 |            |
|      |                       | Sig. | .000   |            |
| 2    | Dead                  | F    |        | 87.021     |
|      |                       | Sig. |        | .000       |
|      | Discharged            | F    | 87.021 |            |
|      |                       | Sig. | .000   |            |
| 3    | Dead                  | F    |        | 73.822     |
|      |                       | Sig. |        | .000       |
|      | Discharged            | F    | 73.822 |            |
|      |                       | Sig. | .000   |            |
| 4    | Dead                  | F    |        | 71.207     |
|      |                       | Sig. |        | .000       |
|      | Discharged            | F    | 71.207 |            |
|      |                       | Sig. | .000   |            |

- a. 1, 98 degrees of freedom for step 1.
- b. 2, 97 degrees of freedom for step 2.
- c. 3, 96 degrees of freedom for step 3.
- d. 4, 95 degrees of freedom for step 4.

## REFERENCE LIST

- 1) Agresti, A. 1996. An introduction to Categorical Data Analysis. John Wiley and Sons.
- 2) Ramayah et al 2010. Discriminant analysis: An illustrated example.
- 3) Paul F. Cook, University of Colorado Denver, Center for Nursing Research.
- 4) McGill, R., Tukey, J. W., and Larsen, W. A. (1978), "Variations of Box Plots," The American Statistician, 32, 12–16.
- 5) Tukey, J. W. (1977), Exploratory Data Analysis, Reading, MA: Addison-Wesley.
- 6) Freemantle, N. Et al (2012) Weekend hospitalization and additional risk of death.
- 7) Cooley, W.W. and P. R. Lohnes. 1971. Multivariate Data Analysis. John Wiley & Sons, Inc.
- 8) George H. Dunteman (1984). Introduction to multivariate analysis.
- 9) Oaks, CA: Sage Publications. Classification procedures and discriminant analysis.
- 10) Klecka, William R. (1980). Discriminant Analysis. Quantitative Applications in the
- 11) Social Sciences Series, No. 19. Thousand Oaks, CA: Sage Publications.
- 12) Lachenbruch, P. A. (1975). Discriminant Analysis. NY: Hafner. For detailed notes on computations.
- 13) Morrison, D.F. 1967. Multivariate Statistical Methods. McGraw-Hill: New York. A general textbook explanation.
- 14) Overall, J.E. and C.J. Klett. 1972. Applied Multivariate Analysis. McGraw-Hill: New York.
- 15) Cox DR, 1958. Two further applications of a model for binary responses. Biometric, 45: 562–565.
- 16) Chan F. Lam and Michael Cox (1981) A Discriminant Analysis Procedure for Paired Variables.

- 17) Mark R. Daniels and R. Darcy (1983) Notes on the Use and Interpretation of Discriminant Analysis.
- 18) S. James Press and Sandra Wilson (1978) Choosing Between Logistic Regression and Discriminant Analysis.
- 19) Wing K. Fung (1995) Diagnostics in Linear Discriminant Analysis.
- 20) D. Hwang (2007) Determination of minimum Sample Size Discriminatory Expression Patterns in Microarray Data.
- 21) F. Brauer (1984) Malaria Disease Infection in adults.
- 22) G.B. Grassi (1899) Global Malaria Eradication Research Agenda.
- 23) I. Kononenko and M. Kukar (2007) Application of Machine Learning.