PREDICTION OF BASELINE ACUTE RESPIRATORY DISTRESS IN SEVERE MALARIA AFRICAN CHILDREN

MSc (BIOSTATISTICS) THESIS

INNOCENT HARVEY GONDWE

UNIVERSITY OF MALAWI

MAY, 2024

PREDICTION OF BASELINE ACUTE RESPIRATORY DISTRESS IN SEVERE MALARIA AFRICAN CHILDREN

MSc (BIOSTATISTICS) THESIS

By

INNOCENT HARVEY GONDWE

BScE (Mathematics and Physics) - Mzuzu University

Submitted to the Department of Mathematical Sciences, School of Natural and Applied Sciences, in partial fulfilment of the requirements for the degree of Master of Science (Biostatistics)

UNIVERSITY OF MALAWI

MAY, 2024

DECLARATION

I, the undersigned hereby declare that this thesis is my own original work which has not been submitted to any other institution for similar purposes. Where other people's work has been used acknowledgements have been made.

INNOCENT HARVEY GONDWE

Full Legal Name

Signature

Date

CERTIFICATE OF APPROVAL

The undersigned certifies that this thesis represents the student's own work and effort and has been submitted with my approval.

Signature: _____Date: _____

Mavuto Mukaka, Ph.D. (Professor)

Supervisor

DEDICATION

This thesis is dedicated to my late parents Richard and Agness Gondwe. Thank you for teaching me the importance of school.

ACKNOWLEDGEMENTS

I would like to thank God for granting me the opportunity to do this research. Secondly, I would like to thank my supervisor, Professor Mavuto Mukaka, for his unwavering support and guidance through the whole process of conducting this research. His support and guidance made me complete this work in time and with less difficulties. I would also like to acknowledge Dr Tsirizani Kaombe, the Master of Science in Biostatistics programme coordinator and the entire leadership of the programme, for their encouragement through this process.

I would also like to thank my wife Juliana Manda Gondwe and my son Richard Gondwe for their understanding, encouragement and support throughout the process of doing this research.

It would not be complete if I fail to thank Tony Matimati, Joachim Nyirongo, Boswell Munthali, Madalitso Mtika, Brave Mwanza, Potiphar Damiano, Chikumbutso Wasera and the entire 2022 Master of Science in Biostatistics cohort for their support and encouragement throughout the journey of my studies.

To all my friends and relatives, thank you for the encouragement and all your support and guidance till the completion of my research work. God bless you all.

ABSTRACT

Acute respiratory distress (ARD) is a global health concern due to its high rates of morbidity and mortality in children. Early identification of the predictors of baseline ARD is very vital to necessitate timely interventions and improved clinical management of the condition. This study aimed at establishing a predictive model for predicting baseline ARD in African children with severe malaria. This retrospective cohort study used secondary data from 'African Quinine-Artesunate Malaria Trial' (AQUAMAT) that was conducted from 2005 to 2010, among children (<15 years) who had been hospitalized for severe malaria. The predictors of baseline ARD were determined using univariable and multivariable binary logistic regression models. A nomogram was constructed to visualise the predictive model. The Receiver Operating Characteristic (ROC) curve was plotted to evaluate the discriminative power of the predictive model. Classification tree analysis was done to classify patients at a higher or lower risk of developing baseline ARD. The outcome of interest was baseline ARD. The study included 5,426 children admitted with severe malaria. The multivariable binary logistic regression model revealed that the major predictors of baseline ARD were pneumonia [Odds Ratio (OR): 2.49, CI: 1.99 - 3.13, p-value < 0.001], severe acidosis (OR: 2.49, CI: 2.09 - 2.97, p-value < 0.001), patient is currently treated for chronic illness (OR: 2.32, CI: 1.05-5.14, p-value $= 0.038$), hyperparasitaemia (OR: 1.96, CI: 1.21 - 3.16, p-value $= 0.006$), sepsis (OR: 1.46, CI: 1.18 - 1.82, p-value $= 0.001$), respiratory rate (OR: 1.03, CI: 1.03 - 1.04, p-value < 0.001). The predictive model was valuable in predicting baseline ARD with overall correct classification of 68.72% and area under the ROC curve of 0.75 (95% CI: 0.73 - 0.77). Classification tree ranked pneumonia, severe acidosis, hyperparasitaemia, sepsis, respiratory rate as well as severe prostration as major conditions classifying a patient of being at high risk of developing baseline ARD. These findings will help medical practitioners in early identification of severe malaria children who are at high risk of developing baseline ARD. This will necessitate improved management and timely interventions provided to such patients in order to prevent development of baseline ARD.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS AND ACRONYMS

ARD	Acute Respiratory Distress
ARDS	Acute Respiratory Distress Syndrome
PARDS	Paediatric Acute Respiratory Distress Syndrome
ICU	Intensive Care Unit
WHO	World Health Organisation
ROC	Receiver Operating Characteristic
AUC	Area Under the ROC Curve
GLM	Generalised Linear Model
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative
CART	Classification and Regression Tree
AECC	American European Consensus Conference
PALICC	Paediatric Acute Lung Injury Consensus Conference
PARDIE	Paediatric Acute Respiratory Distress Incidence and Epidemiology
PICU	Paediatric Intensive Care Unit
PACCMAN	Paediatric Acute and Critical Care Medicine Asian Network
BMI	Body Mass Index
CI	Confidence Interval
AQUAMAT	African Quinine-Artesunate Malaria Trial
MORU	Mahidol Oxford Tropical Medicine Research Unit

CHAPTER 1

INTRODUCTION

1.1 Background

Since its initial description in 1967 by Ashbaugh et al. (1967), acute respiratory distress (ARD) has received widespread recognition as a significant clinical issue with a high morbidity and mortality burden (Confalonieri et al., 2017). Acute respiratory distress is a dangerous lung problem that is characterized by inadequate oxygenation and non-compliant lungs. This disease puts critically ill patients' lives in danger (Feliciano & Mahapatra, 2017). In ARD, surfactant disintegrates due to disruption of alveolar epithelial-endothelial permeability barrier leading to accumulation of protein-rich fluid inside the alveoli which results into hypoxemia (Heidemann et al., 2017).

The most common causes or risk factors of acute respiratory distress are pneumonia, non-pulmonary sepsis, aspiration of gastric contents, non-cardiogenic shock, pancreatitis, severe trauma, drug overdose and ischaemia-reperfusion injury (Bos & Ware, 2022; Feliciano & Mahapatra, 2017; Sweeney & McAuley, 2016). The chance of developing ARD from an underlying disorder can be increased by certain exposures, such as alcohol consumption, smoking, and exposure to pollution in the air (Calfee et al., 2015; Moazed et al., 2022; Reilly et al., 2019; Simou et al., 2018; Ware et al., 2016). According to Gong et al. (2005) and Toy et al. (2022), transfusion of blood products can lead to ARD and

raise risk when there is an underlying cause. Genetic variability may also enhance the risk; however, the majority of discovered variations are infrequent and have a low attributable risk (Reilly et al., 2017).

Acute respiratory distress is a more common health burden worldwide. The incidence of ARD ranges from 1.5 cases per 100,000 person-years to nearly 79 cases per 100,000 person-years (Brun-Buisson et al., 2004; Rubenfeld et al., 2005). Studies from Brazil reported incidence rates ranging from 1.8 to 31 per 100,000 person-years (Azevedo et al., 2013; Caser et al., 2014). A study of patients in 459 intensive care units (ICUs) from 50 countries in 2016 found that 23% of patients on mechanical ventilation and 10% of ICU patients met the criteria for ARD (Bellani et al., 2016). A total of 10.4% of all ICU admissions and 23.4% of patients needing mechanical ventilation among 4499 patients who had acute hypoxemic respiratory failure had ARD. In comparison to South America, Asia and Africa, higher incidence rates were found in North America, Oceania and Europe.

In terms of lung maturation with age, developmental stages, epidemiology, comorbidities, and prognosis, there are significant distinctions between paediatric acute respiratory distress (PARD) and adult acute respiratory distress (ARD) (Flori et al., 2005; Quasney et al., 2015; Schouten et al., 2016; Thomas et al., 2013). Children and adults have different risk factors for developing ARD, and the particular impacts of bacterial or viral causative agents may contribute to the variability of PARD. In developing nations, including Africa, one in five instances of acute respiratory infection in children leads in a lower

respiratory tract infection, which accounts for between 11 and 20 million hospitalizations and 2 million paediatric fatalities annually (Bryce et al., 2005; Williams et al., 2002). This accounts for 20% of the 10.8 million deaths of children under the age of five that occur annually in the world (Bryce et al., 2005). According to Jeena (2008), pneumonia is one of the common causes of acute respiratory distress in children.

In Africa, malaria continues to be a leading cause of illness and mortality in children. An estimated 619,000 individuals died from malaria worldwide in 2021 (WHO, 2022). There are four species of the protozoa in the genus Plasmodium, including; *Plasmodium malaria, Plasmodium vivax, Plasmodium ovale,* and *Plasmodium falciparum,* which cause the acute illness known as malaria. According to Hviid and Jensen (2015), *Plasmodium falciparum* is the most virulent malaria-causing species that affects humans, partly because of its wide range of antigenic diversity and capacity to dwell in the host tissues' microvasculature. It is believed that the etiology of the disease is largely influenced by the accumulation of mature *Plasmodium falciparum* infected erythrocytes in various tissues, which might cause circulation problems and inflammation (Hviid & Jensen, 2015).

All deaths are caused by the *Plasmodium falciparum* infection, which is most common in Sub-Saharan Africa. About 90% of the world's 300-500 million malaria cases and 1.5-2.7 million annual deaths occur in Sub-Saharan Africa (Breman et al., 2004; Helegbe et al., 2007). Cerebral malaria, severe malarial anaemia, and respiratory distress are all part of the clinical symptoms of severe malaria caused by *Plasmodium falciparum*. Children

from Africa who have respiratory distress (RD), a consequence of severe malaria, have a very high risk of dying (Shah et al., 2021). A study conducted by Oduro et al. (2007) in Ghana revealed that severe anaemia (36.5%), respiratory distress (24.4%), and cerebral malaria (5.4%) are the three most prevalent symptoms of severe malaria. By adolescence, those who reside in areas with high rates of transmission usually acquire clinical immunity to severe *falciparum* malaria (Marsh & Kinyanjui, 2006). However, the burden of morbidity and mortality in the paediatric population is disproportionately high because children are at high risk for developing severe malaria.

A study by Blumberg et al. (1996) on "predictors of mortality in severe malaria: a two-year experience in a non-endemic area" reviewed the clinical profiles and therapy of 28 consecutive patients with severe and complicated malaria admitted to Baragwanath Hospital ICU in Johannesburg, South Africa over a two-year period from January 1993 to December 1994. The study found that 13 patients were diagnosed with acute respiratory distress syndrome (ARDS) out of which eight (8) patients died.

Marsh et al. (1995) studied 1844 children who were admitted to the paediatric ward of Kilifi District Hospital in Kenya with a primary diagnosis of malaria. It was revealed that the mortality rate was 3.5% and 84% of the deaths occurred within 24 hours of admission. These fatalities were associated with four prognostic indicators namely; impaired consciousness, respiratory distress, hypoglycaemia and jaundice. Despite the fact that many patients had overlapping symptoms, RD had the greatest mortality rate of the three syndromes in children.

In the original description of "acute respiratory distress syndrome in adults" by Ashbaugh and colleagues in 1967, special attention was paid to the inciting illness or injury which included; severe trauma, viral infection, acute pancreatitis and possible contributing factors which included; hypotension, acidosis, and fluid overload (Ashbaugh et al., 1967). That initial description has, over the years, evolved into the American European Consensus Conference (AECC) definition in 1994 and then the Berlin definition of ARDS for adults and the Paediatric Acute Lung Injury Consensus Conference (PALICC) definition of Paediatric Acute Respiratory Distress Syndrome (PARDS) (see, Bernard et al., 1994; PALICC, 2015; Ranieri et al., 2012). In these definitions, much attention continues to be paid towards understanding what conditions place patients at particular risk for ARDS development and what conditions contribute to worse ARDS clinical outcomes.

Identifying risk factors and understanding which patients are at risk for developing acute respiratory distress is significantly important to be able to develop preventative and early interventions. In a bid to identify patients at risk of developing paediatric acute respiratory distress, Kuhne and Flori (2020), assessed risk factors and etiologies associated with the development of paediatric acute respiratory distress. It was determined that paediatric patients with pre-existing immunodeficiencies, for example, HIV were at an increases risk of both development of acute respiratory distress and worse outcomes after acute respiratory distress. It was also reported that increased Body Mass Index (BMI) has been shown to be associated with increased risk of acute respiratory distress development. Exposure to environmental factors such as smoke, air pollution,

nitrogen dioxide, sulphur dioxide, and particulate matter < 2.5 micrometres were reported to be significantly associated with ARDS development (Lin et al., 2018).

The Paediatric Acute Respiratory Distress Incidence and Epidemiology (PARDIE) study, an international observational study, surveyed over 23,000 Paediatric Intensive Care Unit (PICU) admissions and 12,000 patients requiring mechanical ventilation (Khemani et al., 2008). Of those patients, 744 (3.2%) were identified as having PARDS based on PALICC criteria. Among PARDS patients, the most common risk factor was pneumonia or lower respiratory tract infection (63%), followed by sepsis (19%), aspiration (8%), trauma (4%), other (3%), drowning (1%), and non-septic shock (1%). The Paediatric Acute and Critical Care Medicine Asian Network (PACCMAN) published a study in 2018 which compared "pulmonary" versus "extrapulmonary" ARDS (Kallet et al., 2017). The "extrapulmonary" group included patients with sepsis, massive transfusions, burns, multi-trauma, and haemorrhagic shock and comprised 41 (13.4%) of the 307 patients with PARDS. In this cohort, the extrapulmonary group had higher mortality, higher proportion of multiple organ dysfunction, and higher median oxygenation index.

Studies have found that acute respiratory distress is a major prognostic indictor of morbidity and mortality in children with severe malaria (Marsh et al., 1995; Mitran et al., 2023). Other studies have generated predictive models for predicting acute respiratory distress in patients with sepsis or pneumonia (Lin et al., 2023; Lv et al., 2024; Watkins et al., 2012; Xul et al., 2023). However, no predictive model has been developed to predict baseline acute respiratory distress in children who have severe malaria in Africa. Thus,

this study will embark on predicting baseline acute respiratory distress in severe malaria African children using appropriate statistical prediction models.

1.2 Problem Statement

Malaria is still a pressing public health concern in the African region despite a 2% decline in the number of annual deaths since 2015. There were 257,950 paediatric deaths in Africa in 2019, accounting for 67.2% of all malaria-related deaths across the board (WHO, 2021). This equates to a daily death toll of almost 707 children under the age of five (WHO, 2021). The prevalence is very high in Africa due to inadequate health care services as well as limited resources in the available health facilities. Studies have indicated that respiratory distress is a major prognostic indicator of morbidity and mortality in children with severe malaria (Marsh et al., 1995; Mitran et al., 2023). In developing nations, Africa in particular, one in five instances of acute respiratory infection in children accounts for between 11 and 20 million hospitalizations and 2 million paediatric fatalities annually (Bryce et al., 2005; Williams et al., 2002). This accounts for 20% of the 10.8 million deaths of children under the age of five that occur annually in the world (Bryce et al., 2005).

Studies conducted in Africa have focused on prognostic indicators associated with severe malaria, that would help predict mortality, of which respiratory distress was found to be a major prognostic indicator (Blumberg et al., 1996; Marsh et al., 1995; Mitran et al., 2023; Mzumara et al., 2021). Other studies have generated predictive models for predicting acute respiratory distress in patients with sepsis or pneumonia (Lin et al., 2023; Lv et al., 2024; Watkins et al., 2012; Xul et al., 2023). However, no predictive model has been

developed to predict baseline acute respiratory distress in children who have severe malaria in Africa. For that reason, this study will embark on predicting baseline acute respiratory distress in severe malaria African children using appropriate statistical prediction models.

1.3 Study Objectives

1.3.1 General objective

The general objective of the research is to generate a prediction model to predict baseline acute respiratory distress in children who have severe malaria in Africa.

1.3.2 Specific objectives

i. Predict baseline acute respiratory distress in a cohort of severe malaria children using binary logistic regression model and nomogram.
ii. Calculate measures of goodness-of-fit of the predictive model such as Hosmer-Lemeshow test, sensitivity, specificity, positive predictive values and negative predictive values.
iii. Determine the predictive power of the model using Area Under the ROC Curve (AUC).
iv. Classify patients as having a high or low risk of developing baseline ARD using classification tree.

1.3.3 Study justification

This study is worth conducting because it will:

i. Assist medical practitioners in early identification of severe malaria children who are at high risk of developing baseline ARD.

ii. Help improve management and timely interventions provided to malaria patients in order to prevent development of baseline ARD.

iii. Help WHO and/or Ministries of Health in different countries to come up with health policies and guidelines that guide diagnosis and management of acute respiratory distress in severe malaria children.

iv. Add to the body of knowledge in the scientific academic world.

CHAPTER 2

LITERATURE REVIEW

2.1 Generalized Linear Models

The theory of Generalized Linear Models (GLMs) was first introduced by Nelder and Wedderburn in 1972 (Nelder & Wedderburn, 1972). The aim of GLMs is to define the relationship between the observed response/outcome variable and a set of covariates/explanatory/predictor variables. The outcome variable is seen as a realization from a random variable. This class of models include those whose single response variable was assumed to have the variance reflected by a one-parameter exponential probability distribution. This family of distributions includes the Gaussian or normal, binomial, Poisson, gamma, inverse Gaussian, geometric, and negative binomial (Hardin & Hilbe, 2018; Kutner et al., 2004). The most common GLMs in medical applications are the logistic regression models (for categorical response variable) and Poisson regression models (for count response variable). There are different forms of logistic regression models which include: binary (categorical response variable with two possible outcomes), multinomial (categorical response variable with more than two unordered outcomes) and ordinal (categorical response variable with more than two ordered outcomes) logistic regression. The mathematical details of these three forms of logistic regression models will be discussed in the sections below. This study, in particular, applies binary logistic regression in predicting baseline acute respiratory distress. That is,

the response is either a patient has acute respiratory distress or not based on the presented predictors.

2.2 Logistic Regression Models

A logistic regression model is in the binomial family of generalized linear models. Logistic regression is a powerful statistical model used for modelling the relationship between a set of predictor variables and a categorical outcome variable. There are various forms of logistic regression model, such as binary logistic regression model (with logit or probit or complementary log-log link), multinomial logistic regression model and ordinal logistic regression model (with proportional odds) (Kutner et al., 2004; McCullagh & Nelder, 1989). The choice of a particular type of logistic regression model depends on the type of the categorical outcome of interest. Various types of logistic regression models are discussed in the following sections.

2.2.1 Binary logistic regression model

This model becomes handy when modelling events that have binary responses. Examples of such events, in medical research, include; treatment failure or success, disease status (yes or no), hospitalisation (yes or no), mortality (dead or alive), among others. This outcome can be represented by a binary indicator variable taking on dummy values 0 and 1. A binary logistic regression model is based on the following key assumptions:

1. Dichotomous outcome variable. The outcome variable should be measured on a dichotomous scale, meaning it has only two nominal/categorical values.

2. Mutually exclusive and exhaustive categories. The categories of the outcome variable should be mutually exclusive (no overlap) and exhaustive (cover all possible outcomes).
3. Independence of observations. The observations (data points) should be independent of each other. This assumption ensures that the model is not influenced by repeated measurements or correlated data.
4. No outliers. There should be no extreme outliers in the data that significantly affect the model's performance.
5. Linear relationship. Logistic regression assumes a linear relationship between the log-odds of the outcome and each predictor variable. However, the relationship is modelled using the logit function (S-shaped curve) rather than a straight line.
6. Linearity of independent variables. The independent variables should have a linear effect on the log-odds of the outcome. If the relationship is nonlinear, transformations such as polynomial terms may be needed (Kutner et al., 2004; McCullagh & Nelder, 1989).

If the outcome variable of a generalized linear regression model has two possible outcomes such that the probability, $P(Y_i = 1) = p$ and $P(Y_i = 0) = 1 - p$, then we need a transformation that will bound the values in the range of 0 to 1 as probability outside this range is invalid. Such transformation is necessitated by use of a known monotonic, one-to-one, differentiable link function $g(\cdot)$ relating the linear predictor to the fitted values. (Kutner et al., 2004; McCullagh & Nelder, 1989). Because the function is one-to-

one, there is an inverse function relating the mean expected response, $E[y] = \mu$, to the linear predictor such that

$$\mu = g^{-1}(\eta) = E[y].$$

The commonly used link functions in a binary logistic regression model are as follows:

1. When the outcome is binary and the interest is on assessing odds ratios, the logit link function is commonly used. The logit link function is given by

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right).$$

2. If the outcome variable is considered as obtained by thresholding a normally distributed latent variable, then a probit link function is appropriate. That is, if normality is involved in the linear relationship and the interest is in the predictive and classification value of the model. The probit or inverse normal link function is given by

$$g(\mu_i) = \Phi^{-1}(\mu_i).$$

3. Unlike logit and probit link functions, the log-log function approaches 1 more sharply than it approaches 0. The log-log link function is particularly useful when dealing with rare events. These are situations where the outcome (success or event occurrence) is extremely infrequent, for example, survival after cardiac arrest. It is also used if the outcome variable exhibits extreme probabilities (either very low or very high), for instance, success of glaucoma surgery. The log-log link function is given by

$$g(\mu_i) = -\log[-\log(\mu_i)].$$

4. Complementary log-log link function is the complement of the log-log link function. It is used when dealing with rare events as well as events whose outcome variable exhibits extreme skewness (either unusually very low or very high probabilities). The complementary log-log link function is given by

$$g(\mu_i) = -\log[-\log(1-\mu_i)].$$

Let Y be a column vector of length N where each element Y_i is a random variable representing the number of successes for population i. Let the column vector y contain elements y_i representing the observed counts of the number of successes for each population. Let p be a column vector of length N with elements $p = P(Y_i = 1)$, i.e., the probability of "success" for any given observation in the i^{th} population. Suppose the outcome of interest

$$Y_i \sim \text{Binomial}(n, p)$$

for a particular observation i, with a probability mass function

$$f(y|\beta) = \binom{n}{y} p^y (1-p)^{n-y}, \quad \text{for } y = 0, 1, 2, \dots, n$$

where n is the number of trials. Then, a plausible link function is the logit given by

$$g(p) = \log\left(\frac{p}{1-p}\right).$$

The linear component of the model contains the design matrix and the vector of parameters to be estimated (Hardin & Hilbe, 2018; Kutner et al., 2004; McCullagh & Nelder, 1989). The design matrix of independent variables $\boldsymbol{X}$ is composed of N rows and $K+1$ columns, where K is the number of independent variables specified in the model. The parameter vector $\boldsymbol{\beta}$ is a column vector of length $K+1$. There is one parameter

corresponding to each of the K columns of independent variable settings in $\boldsymbol{X}$, plus one β_0, for the intercept. The logistic regression model equates the logit transform i.e., the log-odds of the probability of a success, to the linear component as follows

$$\log\left(\frac{p}{1-p}\right) = \boldsymbol{X}^T\boldsymbol{\beta} = \sum_{j=0}^{K} x_{ij}\,\beta_j \qquad i = 1, 2, \dots, N \qquad (1)$$

$$p = \frac{\exp(\sum_{j=0}^{K} x_{ij}\,\beta_j)}{1 + \exp(\sum_{j=0}^{K} x_{ij}\,\beta_j)} = \frac{1}{1 + \exp(-\sum_{j=0}^{K} x_{ij}\,\beta_j)}$$

where p is the probability of the "outcome of interest" and the ratio $\frac{p}{1-p}$ is called the odds (Hardin & Hilbe, 2018; Kutner et al., 2004; McCullagh & Nelder, 1989).

2.2.1.1 Maximum likelihood estimation of parameters

The goal of logistic regression is to estimate $K + 1$ unknown parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ in Equation (1). This is achieved using maximum likelihood estimation which entails finding the set of parameters for which the probability of the observed data is greatest. Since each y_i represents a binomial count of i^{th}population, then, the joint probability mass function (likelihood function) of the outcome variable $\boldsymbol{Y}$ is

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{N} \binom{n_i}{y_i} p^{y_i}(1-p)^{n_i - y_i} \qquad (2)$$

where N is the sample size. For each population, there are $\binom{n_i}{y_i}$ different ways to arrange y_i successes from among n_i trials. Since the probability of a success for any one of the n_i trials is p, then, the probability of y_i successes is p^{y_i}. Likewise, the probability of $n_i - y_i$ failures is $(1-p)^{n_i - y_i}$.

The maximum likelihood estimates are the values for $\boldsymbol{\beta}$ that maximize the likelihood function in Equation (2). The critical points of a function (maxima and minima) occur when the first derivative equals 0. If the second derivative evaluated at that point is less than zero, then the critical point is a maximum. However, attempting to take the derivative of Equation (2) with respect to $\boldsymbol{\beta}$ is a difficult task due to the complexity of multiplicative terms. So, a log-likelihood function is used (Czepiel, 2016; Kutner et al., 2004; Kutner et al., 2005; McCullagh & Nelder, 1989).

The term $\binom{n_i}{y_i}$ in Equation (2) does not include p, so it is a constant that can be ignored. After rearranging terms, Equation (2) becomes

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{N} \left(\frac{p}{1-p}\right)^{y_i} (1-p)^{n_i}. \qquad (3)$$

Substituting the relation

$$\frac{p}{1-p} = \exp\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right)$$

and

$$p = \frac{\exp(\sum_{j=0}^{K} x_{ij}\,\beta_j)}{1 + \exp(\sum_{j=0}^{K} x_{ij}\,\beta_j)}$$

in the first and second term, respectively, of Equation (3) yields

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{N}\left(\exp\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right)\right)^{y_i}\left(1 - \frac{\exp(\sum_{j=0}^{K} x_{ij}\,\beta_j)}{1+\exp(\sum_{j=0}^{K} x_{ij}\,\beta_j)}\right)^{n_i}. \qquad (4)$$

Replacing 1 in the second bracket of Equation (4) by $\frac{1+\exp(\sum_{j=0}^{K} x_{ij}\beta_j)}{1+\exp(\sum_{j=0}^{K} x_{ij}\beta_j)}$ and simplifying yields

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{N}\exp\left(y_i\sum_{j=0}^{K} x_{ij}\,\beta_j\right)\left(1+\exp\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right)\right)^{-n_i}. \qquad (5)$$

Equation (5) is the kernel of the likelihood function to maximize. However, it is still difficult to differentiate (Czepiel, 2016; Kutner et al., 2005; McCullagh & Nelder, 1989). However, since the logarithm is a monotonic function, any maximum of the likelihood function will also be a maximum of the log-likelihood function and vice versa. Thus, taking the natural logarithm of Equation (5) yields the log-likelihood function as

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N} y_i\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right) - n_i\cdot\log\left(1+\exp\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right)\right). \qquad (6)$$

The first-order partial derivative of Equation (6) with respect to each β_j is found as

$$\begin{aligned}\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^{N} y_i x_{ij} - n_i\cdot\frac{1}{1+\exp\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right)}\cdot\frac{\partial}{\partial\beta_j}\left(1+\exp\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right)\right)\\ &= \sum_{i=1}^{N} y_i x_{ij} - n_i\cdot\frac{1}{1+\exp\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right)}\cdot\exp\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right)\cdot\frac{\partial}{\partial\beta_j}\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right)\\ &= \sum_{i=1}^{N} y_i x_{ij} - n_i\cdot\frac{\exp\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right)}{1+\exp\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right)}\cdot x_{ij}\end{aligned}$$

$$= \sum_{i=1}^{N} y_i x_{ij} - n_i p x_{ij}. \quad (7)$$

Using Newton-Raphson method to determine the critical values of the derivative of the log-likelihood function, the values of the estimates for $\boldsymbol{\beta}$ are obtained by setting each of the $K+1$ equations in the derivative of the log-likelihood function in Equation (7) to zero and solving for each β_j (Blei, 2015; Czepiel, 2016; Kutner et al., 2005). Each such solution specifies a critical point i.e., either a maximum or a minimum. The critical point will be a maximum if the matrix (Hessian matrix) of second-order partial derivatives is negative definite. That is, if every element on the diagonal of the matrix is less than zero. It is formed by differentiating each of the $K+1$ equations in Equation (7) a second time with respect to each element of $\boldsymbol{\beta}$ denoted by $\beta_{j'}$. The general form of the matrix of second-order partial derivatives is

$$\begin{aligned}
\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_{j'}} &= \frac{\partial}{\partial \beta_{j'}} \left(\sum_{i=1}^{N} y_i x_{ij} - n_i p x_{ij} \right) \\
&= \frac{\partial}{\partial \beta_{j'}} \left(\sum_{i=1}^{N} -n_i p x_{ij} \right) \\
&= -\sum_{i=1}^{N} n_i x_{ij} \frac{\partial}{\partial \beta_{j'}} \left(\frac{\exp(\sum_{j=0}^{K} x_{ij}\, \beta_j)}{1 + \exp(\sum_{j=0}^{K} x_{ij}\, \beta_j)} \right) \\
&= -\sum_{i=1}^{N} n_i x_{ij} p(1-p)\, x_{ij'}.
\end{aligned}$$

which is negative definite (Blei, 2015; Czepiel, 2016; Kutner et al., 2005). Therefore, the estimates of $\boldsymbol{\beta}$ obtained by setting Equation (7) to zero maximize the log-likelihood function in Equation (6) and hence maximize the likelihood function in Equation (2).

2.2.1.2 Odds ratio and interpretation

The most common interpretable measure of effect from logistic regression model is the odds ratio. For example, considering a logistic regression model given in Equation (1), the odds of having acute respiratory distress given a particular predictor variable are

$$\frac{P(Y_i = 1|X_i)}{1 - P(Y_i = 1|X_i)}.$$

In order to obtain the effect of a one-unit increase in the predictor variable on the outcome of interest, a measure known as odds ratio is used and it is calculated as follows:

$$Odds\ ratio = \frac{P(Y_i = 1|X_i + 1)\Big/1 - P(Y_i = 1|X_i + 1)}{P(Y_i = 1|X_i)\Big/1 - P(Y_i = 1|X_i)} = e^{\beta_j} \qquad j = 0, 1, \dots, K.$$

That is, for a one-unit increase in the predictor variable X_i, we expect e^{β_j} times odds of obtaining the outcome of interest (having acute respiratory distress).

2.2.2 Multinomial logistic regression model

In a multinomial logistic regression model, the response variable has three or more categories and there is no natural ordering among these categories. An example, in medical research, could be predicting the type of disease a patient has, among three diseases, namely; diabetes, hypertension and renal failure based on age and gender. The binary logistic model is therefore a special case of the multinomial logistic regression

model. The link function is the generalized logit and the random component is the multinomial distribution. The model differs from the standard logistic model in that the comparisons are all estimated simultaneously within the same model (Czepiel, 2016; Kutner et al., 2004).

The key assumptions of a multinomial logistic regression model are as follows:

1. There should be a linear relationship between the log-odds and the predictor variables.
2. The model assumes that there are no extreme outliers or influential observations in the dataset.
3. Cases should be independent.
4. There should be no multicollinearity between the independent variables (Czepiel, 2016; Kutner et al., 2004).

Let J represent the number of discrete categories of the outcome variable. Consider a random variable $\boldsymbol{Y}$ that can take on one of J possible values. If each observation is independent, then each Y_i is a multinomial random variable. The column vector $\boldsymbol{n}$ contains elements n_i which represent the number of observations in population i and that $\sum_{i=1}^{N} n_i = M$, the total sample size (Czepiel, 2016; Kutner et al., 2004).

Since each observation consist one of J possible values for the outcome variable $\boldsymbol{Y}$, let $\boldsymbol{y}$ be a matrix with N rows and $J-1$ columns. For each population, y_{ij} represents the observed counts of the j^{th}value of Y_i. Also, $\boldsymbol{\pi}$ is a matrix with N rows and $J-1$ columns where each element π_{ij} is the probability of observing the j^{th} value of the response

variable for any given observation in the i^{th}population (Czepiel, 2016; Kutner et al., 2004). The design matrix of predictor variables $\boldsymbol{X}$ contains N rows and $K+1$ columns where K is the number of predictor variables. Let $\boldsymbol{\beta}$ be a matrix with $K+1$ rows and $J-1$ columns. For the multinomial logistic regression model, the linear component is equated to the log of the odds of a j^{th} observation compared to the b^{th} observation (the baseline category). The model can then be written as

$$\log\left(\frac{\pi_{ij}}{\pi_{ib}}\right) = \log\left(\frac{\pi_{ij}}{1-\sum_{j=1}^{J-1}\pi_{ij}}\right) = \sum_{k=0}^{K} x_{ik}\beta_{kj} \qquad (8)$$

for $i = 1, 2, \dots, N$ and $j = 1, 2, \dots, J-1$.

Solving Equation (8) for π_{ij} and π_{ib}, respectively, yields

$$\pi_{ij} = \frac{\exp(\sum_{k=0}^{K} x_{ik}\beta_{kj})}{1+\sum_{j=1}^{J-1}\exp(\sum_{k=0}^{K} x_{ik}\beta_{kj})} \qquad j < J$$

and

$$\pi_{ib} = \frac{1}{1+\sum_{j=1}^{J-1}\exp(\sum_{k=0}^{K} x_{ik}\beta_{kj})}.$$

2.2.2.1 Maximum likelihood estimation of parameters

For each population, the outcome variable follows a multinomial distribution with J levels. That is, the joint probability mass function is

$$f(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{N}\left[\frac{n_i!}{\prod_{j=1}^{J} y_{ij}!}\cdot\prod_{j=1}^{J}\pi_{ij}^{y_{ij}}\right]. \qquad (9)$$

The factorial terms in Equation (9) do not contain any terms with π_{ij} as such they are treated as constants (Czepiel, 2016; Kutner et al., 2004). Therefore, the kernel of the likelihood function for multinomial logistic regression model is

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{N}\prod_{j=1}^{J} \pi_{ij}^{y_{ij}}. \tag{10}$$

Replacing the b^{th}terms, Equation (10) becomes

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{N}\prod_{j=1}^{J-1} \pi_{ij}^{y_{ij}} \cdot \pi_{ib}^{n_i - \sum_{j=1}^{J-1} y_{ij}} \tag{11}$$

which simplifies to

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{N}\prod_{j=1}^{J-1} \pi_{ij}^{y_{ij}} \cdot \frac{\pi_{ib}^{n_i}}{\pi_{ib}^{\sum_{j=1}^{J-1} y_{ij}}}$$

$$= \prod_{i=1}^{N}\prod_{j=1}^{J-1} \pi_{ij}^{y_{ij}} \cdot \frac{\pi_{ib}^{n_i}}{\prod_{j=1}^{J-1} \pi_{ib}^{y_{ij}}}. \tag{12}$$

Grouping together the terms that are raised to the y_{ij} power in Equation (12) gives

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{N}\prod_{j=1}^{J-1} \left(\frac{\pi_{ij}}{\pi_{ib}}\right)^{y_{ij}} \cdot \pi_{ib}^{n_i}. \tag{13}$$

Since

$$\frac{\pi_{ij}}{\pi_{ib}} = \exp\left(\sum_{k=0}^{K} x_{ik}\beta_{kj}\right) \tag{14}$$

and

$$\pi_{ib} = \frac{1}{1 + \sum_{j=1}^{J-1} \exp(\sum_{k=0}^{K} x_{ik}\beta_{kj})}, \tag{15}$$

substitute Equation (14) and (15) in the first and second terms, respectively, of Equation (13) to get

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{N}\prod_{j=1}^{J-1}\left(\exp\left(\sum_{k=0}^{K} x_{ik}\beta_{kj}\right)\right)^{y_{ij}} \cdot \left(\frac{1}{1+\sum_{j=1}^{J-1}\exp\left(\sum_{k=0}^{K} x_{ik}\beta_{kj}\right)}\right)^{n_i}$$

$$= \prod_{i=1}^{N}\prod_{j=1}^{J-1}\exp\left(y_{ij}\sum_{k=0}^{K} x_{ik}\beta_{kj}\right)\cdot\left(1+\sum_{j=1}^{J-1}\exp\left(\sum_{k=0}^{K} x_{ik}\beta_{kj}\right)\right)^{-n_i}. \quad (16)$$

Taking the natural log of Equation (16) gives the log-likelihood function as

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N}\sum_{j=1}^{J-1}\left(y_{ij}\sum_{k=0}^{K} x_{ik}\beta_{kj}\right) - n_i\log\left(1+\sum_{j=1}^{J-1}\exp\left(\sum_{k=0}^{K} x_{ik}\beta_{kj}\right)\right). \quad (17)$$

The goal is to find the values of $\boldsymbol{\beta}$ which maximise Equation (17). This will be done using Newton-Raphson method which involves calculating the first and second-order partial derivatives of the log-likelihood function (Czepiel, 2016; Hossain et al, 2014; Kutner et al., 2004; Rasha, 2021). The first-order partial derivative of Equation (17) is

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_{kj}} = \sum_{i=1}^{N} y_{ij}x_{ik} - n_i\cdot\frac{1}{1+\sum_{j=1}^{J-1}\exp\left(\sum_{k=0}^{K} x_{ik}\beta_{kj}\right)} \cdot\frac{\partial}{\partial\beta_{kj}}\left(1+\sum_{j=1}^{J-1}\exp\left(\sum_{k=0}^{K} x_{ik}\beta_{kj}\right)\right)$$

$$= \sum_{i=1}^{N} y_{ij}x_{ik} - n_i\cdot\frac{1}{1+\sum_{j=1}^{J-1}\exp\left(\sum_{k=0}^{K} x_{ik}\beta_{kj}\right)}\cdot\exp\left(\sum_{k=0}^{K} x_{ik}\beta_{kj}\right)\cdot\frac{\partial}{\partial\beta_{kj}}\left(\sum_{k=0}^{K} x_{ik}\beta_{kj}\right)$$

$$= \sum_{i=1}^{N} y_{ij} x_{ik} - n_i \cdot \frac{1}{1 + \sum_{j=1}^{J-1} \exp\left(\sum_{k=0}^{K} x_{ik}\beta_{kj}\right)} \cdot \exp\left(\sum_{k=0}^{K} x_{ik}\beta_{kj}\right) \cdot x_{ik}$$

$$= \sum_{i=1}^{N} y_{ij} x_{ik} - n_i \pi_{ij} x_{ik}. \qquad (18)$$

There are $(J-1)\cdot(K+1)$ equations in Equation (18) which are set equal to zero and solved for each β_{kj}. The general form of the matrix of second-order partial derivatives is given by

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_{kj} \partial \beta_{k'j'}} = \frac{\partial}{\partial \beta_{k'j'}} \left(\sum_{i=1}^{N} y_{ij} x_{ik} - n_i \pi_{ij} x_{ik} \right) = \frac{\partial}{\partial \beta_{k'j'}} \left(\sum_{i=1}^{N} -n_i \pi_{ij} x_{ik} \right)$$

$$= -\sum_{i=1}^{N} n_i x_{ik} \frac{\partial}{\partial \beta_{k'j'}} \left(\frac{\exp\left(\sum_{k=0}^{K} x_{ik}\beta_{kj}\right)}{1 + \sum_{j=1}^{J-1} \exp\left(\sum_{k=0}^{K} x_{ik}\beta_{kj}\right)} \right)$$

$$= -\sum_{i=1}^{N} n_i x_{ik}\, \pi_{ij}(1 - \pi_{ij}) x_{ik'}, \qquad for\ j' = j$$

which is negative definite (Czepiel, 2016; Hossain et al, 2014; Kutner et al., 2004; Rasha, 2021). Therefore, the estimates of $\boldsymbol{\beta}$ obtained by setting Equation (18) to zero maximize the log-likelihood function in Equation (17) and hence maximize the likelihood function in Equation (10).

2.2.2.2 Odds ratio and interpretation

Odds ratios for each coefficient (for predicting the difference of one category response from the baseline category) are computed as

$$Odds\ ratio = \frac{P(Y_i = j | x_i + 1) \Big/ P(Y_i = b | x_i + 1)}{P(Y_i = j | x_i) \Big/ P(Y_i = b | x_i)} = e^{\beta_{kj}}$$

and represent the odds of increase (or decrease) for category j compared with the baseline category for each unit increase in the predictor variable x_i.

2.2.3 Ordinal logistic regression model

An ordinal logistic regression is used to predict an ordinal outcome variable given one or more predictor variables. In ordinal logistic regression model, the outcome variable has three or more categories. Unlike, multinomial logistic regression model, there is ordering among categories in ordinal logistic regression model (Abreu et al., 2008; Hardin & Hilbe, 2018). An example, in medical research, could be predicting the level of pain (low, mild, high) one hour after taking a particular type of pain-relieving drug. Other ordered categories include; tumour stage (local, regional, distant), disability severity (none, mild, moderate, severe), Likert items (strongly disagree, disagree, agree, strongly agree), weight status (underweight, normal, overweight, obese), among others.

An ordinal logistic regression model is based on the following key assumptions:

1. The outcome variable is measured on an ordinal scale.
2. One or more of the predictor variables are either continuous, categorical or ordinal.
3. There is no multicollinearity. There is no correlation between two or more predictor variables.
4. The response is determined as proportional odds. An ordinal outcome with three or more categories, the odds ratio for the logistic model represents the odds of the higher category as compared to all lower categories combined. In other words, it

is a cumulative odds ratio representing the increased likelihood to the next highest category relative to the lower categories for each unit increase in the predictor (Abreu et al., 2008; Hardin & Hilbe, 2018).

Let y_i denote the response outcome category for subject i. That is, $y_i = j$ means that the response category for that particular subject is j, where $j = 1, 2, \dots, c$.

The cumulative probabilities are modelled as

$$P(y_i \leq j) = \pi_{i1} + \pi_{i2} + \cdots + \pi_{ij}, \qquad j = 1, 2, \dots, c$$

where $\pi_{ij}$is the probability of subject i to choose category j. Also

$$P(Y_i \leq j) = \frac{\exp(\alpha_j + \boldsymbol{x}_i^T \boldsymbol{\beta})}{1 + \exp(\alpha_j + \boldsymbol{x}_i^T \boldsymbol{\beta})}, \qquad j = 1, 2, \dots, c-1 \qquad (19)$$

is a proportional odds model where α_j is the intercept for category j, $\boldsymbol{x}_i^T$is a vector of predictor variables and $\boldsymbol{\beta}$ is a vector of coefficients whose effects are the same for each cumulative logit. That is, the predictor variables have the same effect on the odds of all levels of the response. This is called the proportional-odds assumption or parallel-lines assumption (Abreu et al., 2008; Hardin & Hilbe, 2018).

Taking the logit transformation of both sides of Equation (19) yields cumulative logit link given by

$$logit[P(Y_i \leq j)] = \log\left[\frac{P(Y_i \leq j)}{1 - P(Y_i \leq j)}\right] = \log\left(\frac{\pi_{i1} + \pi_{i2} + \cdots + \pi_{ij}}{\pi_{i,j+1} + \cdots + \pi_{ic}}\right) \qquad j = 1, 2, \dots, c.$$

The cumulative logit link function is, therefore, given by

$$\log\left[\frac{P(Y_i \leq j)}{1 - P(Y_i \leq j)}\right] = \alpha_j + \boldsymbol{x}_i^T \boldsymbol{\beta}, \qquad j = 1, 2, \dots, c-1.$$

2.2.3.1 Maximum likelihood estimation of parameters

Since $P(Y_i \leq j) = F(\alpha_j + \boldsymbol{x}_i^T\boldsymbol{\beta})$, the likelihood function is given by

$$L(\alpha, \beta) = \prod_{i=1}^{N}\left(\prod_{j=1}^{c}(\pi_{ij})^{y_{ij}}\right) = \prod_{i=1}^{N}\left(\prod_{j=1}^{c}[P(Y_i \leq j) - P(Y_i \leq j-1)]^{y_{ij}}\right)$$

where N is the total number of subjects (Abreu et al., 2008; Hardin & Hilbe, 2018). The log-likelihood function is given by

$$l(\alpha, \beta) = \sum_{i=1}^{N}\sum_{j=1}^{c} y_{ij}\log[F\left(\alpha_j + \boldsymbol{x}_i^T\boldsymbol{\beta}\right) - F(\alpha_{j-1} + \boldsymbol{x}_i^T\boldsymbol{\beta})]. \qquad (20)$$

The Newton-Raphson method for estimating the parameters is used to determine the roots of the derivative of the log-likelihood function (Czepiel, 2016). The first derivatives of Equation (20) with respect to α_j and β_k are, respectively, given as

$$\frac{\partial l}{\partial \beta_k} = \sum_{i=1}^{N}\sum_{j=1}^{c} y_{ij}x_{ik}\frac{f\left(\alpha_j + \boldsymbol{x}_i^T\boldsymbol{\beta}\right) - f(\alpha_{j-1} + \boldsymbol{x}_i^T\boldsymbol{\beta})}{F\left(\alpha_j + \boldsymbol{x}_i^T\boldsymbol{\beta}\right) - F(\alpha_{j-1} + \boldsymbol{x}_i^T\boldsymbol{\beta})} \qquad (21)$$

and

$$\frac{\partial l}{\partial \alpha_j} = \sum_{j=1}^{N}\left\{\frac{y_{ik}f\left(\alpha_j + \boldsymbol{x}_i^T\boldsymbol{\beta}\right)}{F\left(\alpha_j + \boldsymbol{x}_i^T\boldsymbol{\beta}\right) - F(\alpha_{j-1} + \boldsymbol{x}_i^T\boldsymbol{\beta})} - \frac{y_{i,j+1}f(\alpha_k + \boldsymbol{x}_i^T\boldsymbol{\beta})}{F\left(\alpha_{j+1} + \boldsymbol{x}_i^T\boldsymbol{\beta}\right) - F(\alpha_j + \boldsymbol{x}_i^T\boldsymbol{\beta})}\right\}. \qquad (22)$$

The values of interest α_j and β_k are obtained by setting the first derivatives in Equations (21) and (22) to zero and solving. Each such root specifies a critical point (either a maximum or a minimum). The critical point will be a maximum if the matrix of second-order partial derivatives is negative definite (Golub & Van Loan, 1996). It is formed by

differentiating Equations (21) and (22) a second time with respect to each element α and β denoted by α_j and β_k (Abreu et al., 2008; Hardin & Hilbe, 2018).

2.2.3.2 Cumulative odds ratio

For subject i, if $\boldsymbol{x}_i$ changes from $\boldsymbol{a}$ to $\boldsymbol{b}$, then

$$logit[P(Y_i \leq j|x_i = \boldsymbol{b})] - logit[P(Y_i \leq j|x_i = \boldsymbol{a})]$$

$$= \log\left[\frac{P(Y_i \leq j|x_i = \boldsymbol{b}) \Big/ P(Y_i > j|x_i = \boldsymbol{b})}{P(Y_i \leq j|x_i = \boldsymbol{a}) \Big/ P(Y_i > j|x_i = \boldsymbol{a})}\right]$$

$$= (b-a)^T\boldsymbol{\beta}$$

That is, the cumulative odds ratio is given by

$$Cumulative\ odds\ ratio = \frac{P(Y_i \leq j|x_i = \boldsymbol{b}) \Big/ P(Y_i > j|x_i = \boldsymbol{b})}{P(Y_i \leq j|x_i = \boldsymbol{a}) \Big/ P(Y_i > j|x_i = \boldsymbol{a})} = e^{(b-a)^T\boldsymbol{\beta}}.$$

2.3 Diagnostic Accuracy

Diagnostic accuracy measures the ability of a predictive model to detect a disease when it is present and to detect the absence of a disease when it is absent. This discriminative ability is assessed by measures of diagnostic accuracy such as sensitivity, specificity, positive predictive values, negative predictive values and area under the Receiver Operating Characteristic (ROC) curve (Šimundić, 2009). These are discussed in the following sections.

2.3.1 Sensitivity and specificity

In 1947, American biostatistician Jacob Yerushalmy coined the words “sensitivity” and “specificity” (Yerushalmy, 1947). An ideal predictive model could totally distinguish between those who have an illness and those who do not. Perfect predicted results that are above a certain threshold are always indicative of the disease, while those that are below the threshold are always negative for the disease. Unfortunately, such a flawless prediction does not exist in reality, and as a result, prediction processes can only partially distinguish between those who have disease and those who do not. Since patients without disease might occasionally have above threshold readings of a certain parameter of interest, then, values above the threshold are not always suggestive of a disease. This implies a false positive (FP) result. Similarly, patients with the disease may present readings of the parameter of interest below threshold. This implies a false negative (FN) result (Šimundić, 2009). In light of parameter values of interest, the threshold divides the population of investigated participants with and without disease into four categories. The model prediction results are compared with the gold standard results i.e., assumed accurate results, as presented in a confusion matrix in table 1.

Table 1: *A* 2 × 2 *confusion matrix*

		Gold standard	
		Disease	**No disease**
Model classification	**Positive**	TP	FP
	Negative	FN	TN

In table 1, true positive (TP) represents a positive model predicted result given that the subjects have the disease. False positive (FP) represents a positive model predicted result given that the subjects do not have the disease. True negative (TN) represents a negative model predicted result given that the subjects do not have the disease and false negative (FN) represents a negative model predicted result given that the subjects have the disease (Šimundić, 2009; Swift et al., 2020).

Sensitivity refers to the probability that a predictive model will return a positive result when a disease is actually present. Sensitivity is the percentage or proportion of patients who are truly positive for the disease among all patients who have the disease. From table 1, we calculate sensitivity as

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

or if positive model predicted result is denoted by T and having a disease on the gold standard is denoted by D, then

$$\text{Sensitivity} = P(T|D).$$

Specificity refers to the probability that a predictive model will return a negative result when a disease is actually absent. Specificity is the percentage or proportion of subjects who are truly negative for the disease among all subjects who do not have the disease (Šimundić, 2009; Swift et al., 2020). From Table 1, we calculate specificity as

$$\text{Specificity} = \frac{TN}{TN + FP}$$

or if negative model predicted result is denoted by $\bar{T}$ and not having a disease on the gold standard is denoted by $\bar{D}$, then

$$\text{Specificity} = P(\bar{T}|\bar{D}).$$

Remarks:

i. *A predictive model with high sensitivity will detect some individuals without the condition. The predictive model will identify everyone who has the condition as well as many people who do not. This is crucial if the condition's consequences for not treating it are severe and/or if there is a treatment that is readily available, highly effective, and has few negative side effects. However, for healthy people, this will lead to stress and unneeded follow-up.*

ii. *A predictive model with high specificity will result in a high number of true negatives and smaller number of false positives. In this case, subjects identified as having a disease may be subjected to more testing.*

iii. *The prevalence of the disease has no effect on sensitivity or specificity, therefore findings from one study might readily be applied to another context with a variable prevalence of the condition in the population. However, depending on the disease spectrum in the examined population, sensitivity and specificity can vary significantly.*

2.3.2 Positive predictive value

Positive predictive value refers to the percentage or proportion of patients with positive predicted results among all subjects with positive test results. From table 1, we have

$$\text{Positive predictive value} = \frac{TP}{TP + FP}.$$

Positive predictive value is also defined as the probability of having a disease of interest in a patient with positive predicted result (Šimundić, 2009; Swift et al., 2020). Mathematically,

$$\text{Positive predictive value} = P(D|T).$$

2.3.3 Negative predictive value

Negative predictive value refers to the percentage or proportion of subjects with negative predicted results among all subjects with negative test results. From table 1, we have

$$\text{Negative predictive value} = \frac{TN}{TN + FN}.$$

Negative predictive value is also defined as the probability of not having a disease of interest in a subject with negative predicted result (Šimundić, 2009; Swift et al., 2020). Mathematically,

$$\text{Negative predictive value} = P(\bar{D}|\bar{T}).$$

Remarks:

i. *Predictive values are strongly influenced by the prevalence of the disease in the population under study. Because of this, predictive estimates from one study should not be used in a situation where the population's prevalence of the disease is different.*

ii. *With an increase in the disease's prevalence in a population, the positive predictive value rises and the negative predictive value falls.*

2.3.4 Area Under ROC Curve (AUC)

The Area Under the Curve (AUC) score is the area under the Receiver Operating Characteristic (ROC) curve, and it measures the ability of the predictive model to accurately predict classes i.e., a child having baseline acute respiratory distress on not. A ROC curve plots sensitivity against specificity at different possible classification thresholds (Šimundić, 2009; Swift et al., 2020). AUC is defined as the likelihood that the predictive model will give a higher probability to a random positive observation than to a random negative observation (Hanley & McNeil, 1982). The AUC score represents the predictive model's ability to accurately categorize classes on a scale of 0 to 1, with 1 being the best and 0.5 being as good as random choosing. This is a measure used to assess the accuracy or performance of a predictive model (Šimundić, 2009; Swift et al., 2020).

To create a ROC curve, we plot specificity on the x-axis and sensitivity on the y-axis. The strength of a predictive model's discriminative power is determined by examining the shape of a ROC curve and the area under the curve (Šimundić, 2009). The predictive model's ability to distinguish between diseased and non-diseased individuals is improved by the curve's proximity to the upper-left corner and the size of the area under the curve. Table 2 describes the relationship between the area under the ROC curve and predictive model accuracy (Šimundić, 2009).

Table 2: *Area under the ROC curve and predictive model accuracy (Šimundić, 2009)*

AUC value	**Predictive model accuracy**

$0.9 \leq \text{AUC} \leq 1.0$	Excellent
$0.8 \leq \text{AUC} < 0.9$	Very good
$0.7 \leq \text{AUC} < 0.8$	Good
$0.6 \leq \text{AUC} < 0.7$	Sufficient
$0.5 \leq \text{AUC} < 0.6$	Bad
< 0.5	Model not useful

Area under the ROC curve is a generic indicator of predictive model accuracy that is essential for overall evaluation and for comparing the results of two or more predictive models. The area under each of the two ROC curves can be compared to determine which test is more suited to separating the diseased individuals from the non-diseased (Hanley & McNeil, 1982; Šimundić, 2009; Swift et al., 2020).

2.4 Classification Tree

Techniques such as multiple linear regression can yield reliable predictive models when there is a linear relationship between a set of predictor variables and a response variable. On the other hand, non-linear approaches frequently result in more accurate models when there is a more complex relationship between a collection of predictor variables and a response. One such technique is classification and regression trees (CART), which constructs decision trees that predict the value of a response variable based on a collection of predictor variables. Regression trees are constructed when the response variable is continuous. On the other hand, classification trees are constructed when the

response variable is categorical (Flom, 2018; Loh, 2011). This study will construct a classification tree since the outcome variable is categorical.

Classification tree is obtained by iteratively partitioning the data space and fitting a basic prediction model within each partition, leading to a decision tree as a graphical representation of the partitioning (Flom, 2018; Loh, 2011). Classification trees are used when the dependent variable can have a finite number of unordered values and the prediction error is expressed in terms of the cost of misclassification. Graphically, we begin by grouping together all of the observations and then divide it into two groups by selecting the best predictor value for the split, which results in two nodes. Then, the process is repeated until a full tree is obtained. The full tree may, sometimes, overfit the data. Therefore, the best tree is obtained by pruning (Flom, 2018; Loh, 2011).

2.4.1 Classification tree algorithm

In a classification problem, there is a training sample of n observations on a response or class variable Y and p predictor variables, $X_1, X_2, \dots, X_p$. The goal is to find a model for predicting the values of Y from new X values (Loh, 2011). A simple tree structure is defined as

$$y(x_1, x_2) = \begin{cases} y_1 \; if \; x_1 \leq s_1 \\ y_2 \; if \; x_1 > s_1 \; and \; x_2 \leq s_2 \\ y_3 \; if \; x_1 > s_1 \; and \; x_2 > s_2. \end{cases}$$

The objective of a classification tree is to estimate a binary tree structure. This is achieved by performing three algorithms, namely; tree growing: step-optimal recursive partition, tree pruning and obtaining the honest tree. Tree pruning and obtaining the

honest tree are meant to minimise overfitting i.e., growing trees with no external validity (Mora, 2019).

Tree growing requires training or learning sample. At iteration i with tree structure T_i, consider all terminal nodes t^*T_i. In a classification tree, we let $i(T_i)$ to be an overall impurity measure using gini or entropy index. The best split at iteration i identifies the terminal node and split criterion that maximizes

$$i(T_i) - i(T_{i+1}).$$

Recursive partitioning ends with the largest possible tree, T_{max} where there are no nodes to split or the number of observations reach a lower limit also referred to as splitting rule. In this regard, T_{max} will usually be too complex (overfit) because it has no external validity and some terminal nodes should be aggregated (Mora, 2019). Besides, a more simplified structure will normally lead to more accurate estimates since the number of observations in each terminal node grows as aggregation takes place. In order to avoid overfitting, classification tree algorithm identifies a sequence of nested trees that results from recursive aggregation of nodes from T_{max} with a clustering procedure. For a given value α, let $R(\alpha, T) = R(T) + \alpha|T|$ where $|T|$ denotes the number of terminal nodes, or complexity, of tree T and $R(T)$ is the misclassification rate (Mora, 2019). The optimal tree for a given α, $T(\alpha)$, minimises $R(\alpha, T)$ within the set of subtrees of T_{max}. Pruning identifies a sequence of real positive numbers $\{\alpha_0, \alpha_1, \dots, \alpha_M\}$ such that $\alpha_j < \alpha_{j+1}$ and

$$T_{max} \equiv \mathrm{T}(\alpha_0) \to \mathrm{T}(\alpha_1) \to \mathrm{T}(\alpha_2) \to \cdots \to \{root\}.$$

Out of the sequence of optimal trees, $\{T(\alpha_j)\}_j$, T_{max} has lowest $R(T)$ in the learning sample by construction and $R(\cdot)$ increases with α (Mora, 2019).

The honest tree algorithm chooses the simplest tree that minimizes

$$R(T) + s \times SE\big(R(T)\big), \quad s \geq 0.$$

With partitioning into a learning and a test sample, on one hand, $R(T)$ and $SE\big(R(T)\big)$ are obtained using the test sample. On the other hand, with V-fold cross validation, the sample is randomly partitioned V times into a learning and a test sample. The measures α_i, $R(T)$ and $SE\big(R(T)\big)$ are obtained through averaging of results in the V partitions (Mora, 2019).

2.5 Review of studies that applied binary logistic regression model

Marsh et al. (1995) studied 1844 children (mean age, 26.4 months) with a primary diagnosis of malaria who were admitted in the paediatric ward of Kilifi District Hospital in Kenya. The primary goal of the study was to determine indicators of life-threatening malaria in African children. It was found that the mortality rate was 3.5% (95% CI of 2.7 - 4.3%), and 84% of the deaths occurred within 24 hours of admission. The study employed a binary logistic regression model in order to determine key prognostic indicators of death (outcome variable) from malaria. Four indicators were established, namely; respiratory distress (relative risk, 3.9; 95% CI, 2.0-7.7), impaired consciousness (relative risk, 3.3; 95% CI, 1.6-7.0), hypoglycaemia (relative risk, 3.3; 95% CI, 1.6-6.7), and jaundice (relative risk, 2.6; 95% CI, 1.1-6.3). The 54 out of 64 children who died were those with respiratory distress ($n = 251$; case fatality rate, 13.9%) or impaired consciousness ($n = 336$; case fatality rate, 11.9%), or both.

Mzumara et al. (2021) used binary univariable and multivariable logistic regression models to identify prognostic factors for severe metabolic acidosis and uraemia in 5425 children from nine African countries who had severe falciparum malaria. The results indicated the prognostic features of severe metabolic acidosis were deep breathing (OR: 3.94, CI 2.51-6.2), hypoglycaemia (OR:5.16, CI 2.74-9.75), coma (OR: 1.72 CI 1.17-2.51), respiratory distress (OR: 1.46, CI 1.02-2.1) and prostration (OR: 1.88, CI 1.35-2.59). Prognostic features associated with uraemia were coma (3.18, CI 2.36-4.27), prostration (OR: 1.78 CI 1.37-2.30), decompensated shock (OR: 1.89, CI 1.31-2.74), black water fever (CI 1.58, CI 1.09-2.27), jaundice (OR: 3.46 CI 2.21-5.43), severe anaemia (OR: 1.77, CI 1.36-2.29) and hypoglycaemia (OR: 2.77, CI 2.22-3.46). Results indicated that the strongest predictors of severe metabolic acidosis were hypoglycaemia and deep breathing. On the other hand, the strongest predictors of uraemia were jaundice, coma and hypoglycaemia.

Xu et al. (2023) developed a prediction model for predicting the risk of acute respiratory distress syndrome in sepsis patients. This retrospective cohort study recruited a total of 16,523 sepsis patients who were randomly divided into the training and testing sets. The outcome of interest was the occurrence of ARDS for ICU patients with sepsis. Univariate and multivariate logistic regression analyses were used in the training set to identify the factors that were associated with ARDS risk, which were adopted to establish the nomogram. The receiver operating characteristic and calibration curves were used to assess the predictive performance of nomogram. Results showed that a total of 2422 (20.66%) sepsis patients resulted in ARDS. It was found that that body mass index,

respiratory rate, urine output, partial pressure of carbon dioxide, blood urea nitrogen, vasopressin, continuous renal replacement therapy, ventilation status, chronic pulmonary disease, malignant cancer, liver disease, septic shock and pancreatitis might be predictors. The area under the curve of developed model were 0.811 (95% CI 0.802–0.820) in the training set and 0.812 (95% CI 0.798–0.826) in the testing set.

There has been an increase in the number of human adenovirus (HAdV)-related pneumonia cases in immunocompetent adults and acute respiratory distress syndrome in these patients is the predominant cause of HAdV-associated fatality rates. Based on this background, Lin et al. (2023) developed "a prediction model for acute respiratory distress syndrome in immunocompetent adults with adenovirus-associated pneumonia". The study used data from immunocompetent adults with HAdV-pneumonia between June 2018 and May 2022 in ten tertiary general hospitals in central China which was analysed retrospectively. The prediction model of HAdV-related ARDS was developed using multivariate stepwise logistic regression and visualized using a nomogram. Out of 102 patients with adenovirus pneumonia, 41 (40.2%) developed ARDS. Overall, most patients were male (94.1%), the median age was 38.0 years. Results of a multivariate logistic regression model indicated that dyspnea, Sequential Organ Failure Assessment (SOFA) score, lactate dehydrogenase (LDH) and mechanical ventilation status were independent risk factors for the development of ARDS. Using these factors, a nomogram was established with an associated concordance statistic of 0.904 (95% CI 0.844–0.963). The nomogram was meant to help predict early HAdV-related ARDS.

Community-acquired pneumonia (CAP) is a global health concern due to its high rates of morbidity and mortality. Bacterial pathogens are common causes of CAP. It is one of the most common causes of acute respiratory distress syndrome (ARDS). In a quest to ensure early identification of the occurrence and effective prevention of ARDS in patients with bacterial pneumonia, Lv et al. (2024) did a study aimed at establishing a predictive model for ARDS in patients with bacterial pneumonia. The study used clinical data of hospitalized patients with bacterial pneumonia in Affiliated Huzhou Hospital of Zhejiang University School of Medicine from January 2022 to November 2022. The independent risk factors for ARDS in patients with bacterial pneumonia were determined by using univariate and multivariate binary logistic regression analyses. The nomogram was constructed to display the predictive model, and the receiver-operating characteristic curve was plotted to evaluate the predictive value of ARDS. This study included 254 patients with bacterial pneumonia, of which 114 developed ARDS. The multivariate logistic regression analysis revealed that age (OR = 1.041, p = 0.003], heart rate (OR = 1.020, p = 0.028), lymphocyte count (OR = 0.555, p = 0.033), white blood cell count (OR = 1.062, p = 0.033), bilateral lung lesions (OR = 7.352, p = 0.011) and pleural effusion (OR = 2.512, p = 0.002) were the independent risk factors for ARDS. The predictive model was constructed based on the six independent factors and it gave AUC value of 0.794. It was concluded that the predictive model was beneficial to evaluate the disease progression in patients with bacterial pneumonia and identify ARDS. Also, the nomogram would help doctors predict the incidence of ARDS and conduct treatment as early as possible.

Malaria is still a pressing public health concern in the African region. There were 257,950 paediatric deaths in Africa in 2019, accounting for 67.2% of all malaria-related deaths across the board (WHO, 2021). This equates to a daily death toll of almost 707 children under the age of five (WHO, 2021). The prevalence is very high in Africa due to inadequate health care services and limited resources in health facilities. Literature has shown that there is a close association between malaria and the insurgence of acute respiratory distress. Several studies have been conducted focusing on determining risk factors associated with the development of acute respiratory distress in children worldwide so as to be able to develop preventative and early intervention measures. Prediction models have also been developed to predict the risk of developing acute respiratory distress in pneumonia, sepsis and trauma patients. However, no study has so far focused on developing a prediction model of baseline acute respiratory distress in African children who have severe *falciparum malaria*. For that reason, this study will embark on predicting baseline acute respiratory distress in African children who are diagnosed with severe malaria using statistical prediction models.

CHAPTER 3

METHODOLOGY

3.1 Study Design

This is a secondary retrospective analysis of a multicenter, open-label 'African Quinine-Artesunate Malaria Trial' (AQUAMAT) that was conducted from October 3, 2005, to July 14, 2010, among children (<15 years) who had been hospitalized for severe malaria. Eleven centres from nine African nations participated including; Mozambique, The Gambia, Ghana, Kenya, Tanzania, Nigeria, Uganda, Rwanda, and the Democratic Republic of the Congo (Dondorp et al., 2010). Inclusion criteria for children under 15 years old were a positive rapid diagnostic test for *Plasmodium falciparum* lactate dehydrogenase, clinical judgment of the admitting physician that the patient had severe malaria, and fully informed written consent from the patient or a guardian. Patients who had a positive malaria test and at least one of the WHO symptoms were considered to have severe malaria (Dondorp et al., 2010). Patients who had a compelling history of receiving parenteral quinine or an artemisinin derivative for more than 24 hours prior to admission were excluded from the study.

The AQUAMAT recruited 5426 children with 2713 patients in the artesunate arm and 2713 patients in the quinine arm (Dondorp et al., 2010; Mzumara et al., 2021). The participating countries and their corresponding number of subjects are as follows:

Mozambique (332, 12%), The Gambia (252, 9%), Ghana (218, 8%), Kenya (223, 8%), Tanzania (732, 27%), Nigeria (224, 8%), Uganda (330, 12%), Rwanda (192, 7%), and the Democratic Republic of the Congo (210, 8%). The main outcome measure of interest was to compare in-hospital mortality between treatments using intention-to-treat. Incidence of severe neurological problems and a combined outcome measure of mortality and severe persistent neurological sequelae were used as secondary outcome measures. The trial supported the use of parenteral artesunate in the treatment of *Plasmodium falciparum* in children worldwide (Dondorp et al., 2010).

3.2 Variables

3.2.1 Outcome variable

This study is aimed at predicting baseline acute respiratory distress in African children who have severe malaria. As such, the outcome variable of interest is baseline acute respiratory distress. This variable has binary outcomes (having baseline acute respiratory distress or not).

3.2.2 Predictor variables

Studies of children who were admitted to hospitals with a primary diagnosis of malaria revealed that fatalities were mainly associated with respiratory distress and impaired consciousness among other minor factors (Marsh et al., 1995; Oduro et al., 2007; Shah et al., 2021). According to Dondorp et al. (2010), WHO signs of severe falciparum malaria are plasma base excess less than –3.3 mmol/L, Glasgow coma scale less than 11 of 15 or

Blantyre coma scale less than 3 of 5 in preverbal children, haemoglobin less than 50 g/L and parasitaemia greater than 100,000 parasites per μL, blood urea greater than 10 mmol/L, compensated shock (capillary refill $\geq$ 3 or temperature gradient on legs, but no hypotension), decompensated shock, systolic blood pressure less than 70 mm Hg and cool peripheries, asexual parasitaemia more than 10%, visible jaundice and more than 100,000 parasites per μL, plasma glucose less than 3 mmol/L and respiratory distress, defined as costal indrawing, use of accessory muscles, nasal alar flaring, deep breathing, or severe tachypnoea.

Based on literature findings on risk factors associated with acute respiratory distress, the Paediatric Acute Lung Injury Consensus Conference (PALICC) definition of Paediatric Acute Respiratory Distress Syndrome (PARDS) and WHO signs of severe falciparum malaria, the candidate demographic and clinical predictor variables for the outcome of interest (baseline acute respiratory distress), in this study, are patient age (years), sex, weight (kg), respiratory rate (per minute), systolic blood pressure (mmHg), diastolic blood pressure (mmHg), pneumonia, sepsis, symptomatic severe anaemia (severe pallor combined with respiratory distress), coma at admission (GCS $\leq$ 10, BCS $\leq$ 2), convulsions > 30 minutes, compensated shock (capillary refill $\geq$ 3 sec temperature gradient), decompensated shock (adults: systolic BP< 80 mmHg, children: systolic BP < 70 mmHg), hyperparasitaemia (> 500 parasites per high powered field), severe acidosis: deep breathing, blood transfusion, mechanical ventilation, patient is currently treated for chronic illness, renal failure and severe prostration (not able to breastfeed < 6m, or able to sit > 6m).

3.3 Statistical Analysis

Data exploratory analysis and visualization was done using approaches like tables, histograms and bar graphs in order to gain insight into the data. Since the response variable of interest, in this study, is categorical with two possible outcomes, a binary logistic regression model is a reasonable predictive model in predicting baseline acute respiratory distress based on demographic and clinical predictor variables.

3.3.1 Univariable binary logistic regression model

To determine potential predictors of baseline acute respiratory distress, first a univariable binary logistic regression model was used. The model is given by

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_{i1} \qquad i = 1, 2, \dots, N$$

$$p(Y_i = outcome\ of\ interest | x_{i1}) = \frac{\exp(\beta_0 + \beta_1 X_{i1})}{1 + \exp(\beta_0 + \beta_1 X_{i1})} = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 X_{i1}))}$$

where p is the probability of observing the outcome variable of interest (a patient having acute respiratory distress), β_0 is the intercept, β_1 is the coefficient which relates a response and a predictor variable and X_{i1} is a predictor variable.

Predictor variables and some patient demographics which were statistically significant, at 5% level of significance, in the univariable model were included in the multivariable binary logistic regression model (predictive model) using forward stepwise approach. The multivariable binary logistic regression model is discussed below:

3.3.2 Multivariable binary logistic regression model

Let Y be a column vector of length N where each element Y_i is a random variable representing the number of successes for population i. Let the column vector y contain elements y_i representing the observed counts of the number of successes for each population. Let p be a column vector of length N with elements $= P(Y_i = 1)$, i.e., the probability of "success" for any given observation in the i^{th} population. Suppose the outcome of interest

$$Y_i \sim \text{Binomial}(n, p)$$

for a particular observation i, with a probability mass function

$$f(y|\beta) = \binom{n}{y} p^y (1-p)^{n-y}, \quad \text{for } y = 0, 1, 2, \dots, n.$$

where n is the number of trials. Then, a plausible link function is the logit given by

$$g(p) = \log\left(\frac{p}{1-p}\right).$$

The linear component of the model contains the design matrix and the vector of parameters to be estimated (Hardin & Hilbe, 2018; Kutner et al., 2004; McCullagh & Nelder, 1989). The design matrix of predictor variables $\boldsymbol{X}$ is composed of N rows and $K + 1$ columns, where K is the number of predictor variables specified in the model. The parameter vector $\boldsymbol{\beta}$ is a column vector of length $K + 1$. There is one parameter corresponding to each of the K columns of predictor variable settings in $\boldsymbol{X}$, plus one β_0, for the intercept. The logistic regression model equates the logit transform i.e., the log-odds of the probability of a success, to the linear component as follows

$$\log\left(\frac{p}{1-p}\right) = \boldsymbol{X}^T\boldsymbol{\beta} = \sum_{j=0}^{K} x_{ij}\,\beta_j \qquad i = 1, 2, \dots, N \qquad (23)$$

$$p(Y_i = outcome\ of\ interest|x_{ij}) = \frac{\exp(\sum_{j=0}^{K} x_{ij}\,\beta_j)}{1 + \exp(\sum_{j=0}^{K} x_{ij}\,\beta_j)} = \frac{1}{1 + \exp(-\sum_{j=0}^{K} x_{ij}\,\beta_j)}.$$

where p is the probability of the "outcome of interest" and the ratio $\frac{p}{1-p}$ is called the odds (Hardin & Hilbe, 2018; Kutner et al., 2004; McCullagh & Nelder, 1989).

3.3.2.1 Maximum likelihood estimation of parameters

The goal of logistic regression is to estimate $K + 1$ unknown parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ in Equation (23). This is achieved using maximum likelihood estimation which entails finding the set of parameters for which the probability of the observed data is greatest. Since each y_i represents a binomial count of i^{th}population, then, the joint probability mass function (likelihood function) of the outcome variable $\boldsymbol{Y}$ is

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{N} \binom{n_i}{y_i} p^{y_i}(1-p)^{n_i - y_i} \qquad (24)$$

where N is the sample size. For each population, there are $\binom{n_i}{y_i}$ different ways to arrange y_i successes from among n_i trials. Since the probability of a success for any one of the n_i trials is p, then, the probability of y_i successes is p^{y_i}. Likewise, the probability of $n_i - y_i$ failures is $(1-p)^{n_i - y_i}$.

The maximum likelihood estimates are the values for $\boldsymbol{\beta}$ that maximize the likelihood function in Equation (24). The critical points of a function (maxima and minima) occur when the first derivative equals 0. If the second derivative evaluated at that point is less than zero, then the critical point is a maximum. However, attempting to take the derivative of Equation (24) with respect to $\boldsymbol{\beta}$ is a difficult task due to the complexity of multiplicative terms. So, a log-likelihood function is used (Czepiel, 2016; Kutner et al., 2004; Kutner et al., 2005; McCullagh & Nelder, 1989).

The term $\binom{n_i}{y_i}$ in Equation (24) does not include p, so it is a constant that can be ignored. After rearranging terms, Equation (24) becomes

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{N} \left(\frac{p}{1-p}\right)^{y_i} (1-p)^{n_i}. \qquad (25)$$

Substituting the relation

$$\frac{p}{1-p} = \exp\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right)$$

and

$$p = \frac{\exp(\sum_{j=0}^{K} x_{ij}\,\beta_j)}{1+\exp(\sum_{j=0}^{K} x_{ij}\,\beta_j)}$$

in the first and second term, respectively, of Equation (25) yields

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{N} \left(\exp\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right)\right)^{y_i} \left(1 - \frac{\exp(\sum_{j=0}^{K} x_{ij}\,\beta_j)}{1+\exp(\sum_{j=0}^{K} x_{ij}\,\beta_j)}\right)^{n_i}. \qquad (26)$$

Replacing 1 in the second bracket of Equation (26) by $\frac{1+\exp(\sum_{j=0}^{K} x_{ij}\beta_j)}{1+\exp(\sum_{j=0}^{K} x_{ij}\beta_j)}$ and simplifying yields

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{N} \exp\left(y_i \sum_{j=0}^{K} x_{ij}\,\beta_j\right)\left(1+\exp\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right)\right)^{-n_i}. \qquad (27)$$

Equation (27) is the kernel of the likelihood function to maximize. However, it is still difficult to differentiate (Czepiel, 2016; Kutner et al., 2005; McCullagh & Nelder, 1989). Since the logarithm is a monotonic function, any maximum of the likelihood function will also be a maximum of the log-likelihood function and vice versa. Thus, taking the natural logarithm of Equation (27) yields the log-likelihood function as

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N} y_i \left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right) - n_i \cdot \log\left(1+\exp\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right)\right). \qquad (28)$$

The first-order partial derivative of Equation (28) with respect to each β_j is found as

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^{N} y_i x_{ij} - n_i \cdot \frac{1}{1+\exp\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right)} \cdot \frac{\partial}{\partial \beta_j}\left(1+\exp\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right)\right) \\
&= \sum_{i=1}^{N} y_i x_{ij} - n_i \cdot \frac{1}{1+\exp\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right)} \cdot \exp\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right) \cdot \frac{\partial}{\partial \beta_j}\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right) \\
&= \sum_{i=1}^{N} y_i x_{ij} - n_i \cdot \frac{\exp\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right)}{1+\exp\left(\sum_{j=0}^{K} x_{ij}\,\beta_j\right)} \cdot x_{ij} \\
&= \sum_{i=1}^{N} y_i x_{ij} - n_i p x_{ij}. \qquad (29)
\end{aligned}$$

Using Newton-Raphson method to determine the critical values of the derivative of the log-likelihood function, the values of the estimates for $\boldsymbol{\beta}$ are obtained by setting each of

the $K + 1$ equations in the derivative of the log-likelihood function in Equation (29) to zero and solving for each β_j (Blei, 2015; Czepiel, 2016; Kutner et al., 2005). Each such solution specifies a critical point. The critical point will be a maximum if the matrix (Hessian matrix) of second-order partial derivatives is negative definite. That is, if every element on the diagonal of the matrix is less than zero. It is formed by differentiating each of the $K + 1$ equations in Equation (29) a second time with respect to each element of $\boldsymbol{\beta}$ denoted by $\beta_{j'}$. The general form of the matrix of second-order partial derivatives is

$$\begin{aligned}
\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_{j'}} &= \frac{\partial}{\partial \beta_{j'}} \left(\sum_{i=1}^{N} y_i x_{ij} - n_i p x_{ij} \right) \\
&= \frac{\partial}{\partial \beta_{j'}} \left(\sum_{i=1}^{N} -n_i p x_{ij} \right) \\
&= -\sum_{i=1}^{N} n_i x_{ij} \frac{\partial}{\partial \beta_{j'}} \left(\frac{\exp(\sum_{j=0}^{K} x_{ij}\, \beta_j)}{1 + \exp(\sum_{j=0}^{K} x_{ij}\, \beta_j)} \right) \\
&= -\sum_{i=1}^{N} n_i x_{ij} p(1-p)\, x_{ij'}.
\end{aligned}$$

which is negative definite (Blei, 2015; Czepiel, 2016; Kutner et al., 2005). Therefore, the estimates of $\boldsymbol{\beta}$ obtained by setting Equation (29) to zero maximize the log-likelihood function in Equation (28) and hence maximize the likelihood function in Equation (24).

3.3.2.2 Odds ratio and interpretation

The most common interpretable measure of effect from logistic regression model is the odds ratio. For example, considering a binary logistic regression model given in Equation (23), the odds of having acute respiratory distress given a particular predictor variable are

$$\frac{P(Y_i = 1|X_i)}{1 - P(Y_i = 1|X_i)}.$$

In order to obtain the effect of a one-unit increase in the predictor variable on the outcome of interest, a measure known as odds ratio is used and it is calculated as follows:

$$Odds\ ratio = \frac{P(Y_i = 1|X_i + 1)\Big/1 - P(Y_i = 1|X_i + 1)}{P(Y_i = 1|X_i)\Big/1 - P(Y_i = 1|X_i)} = e^{\beta_j} \qquad j = 0, 1, \dots, K.$$

That is, for a one-unit increase in the predictor variable X_i, we expect e^{β_j} times odds of obtaining the outcome of interest (baseline acute respiratory distress).

In order to visualise the prediction model, a nomogram was plotted. A nomogram ranks the importance of a predictor variable in predicting the outcome (baseline acute respiratory distress) in the context of the other predictor variables in the model. Each of the predictor variables included in the predictive model were arranged one by one on a horizontal plane with its scoring system, ranging from 0 to 10, at the bottom. The total score ranged from 0 to 27. The most important predictors in predicting the outcome of interest have higher scores.

The goodness-of-fit for the predictive model given by Equation (23) was assessed using Hosmer-Lemeshow goodness-of-fit test. This statistical test measures the correspondence

of the observed and predicted values of the outcome variable. A better model fit is characterized by insignificant differences between the observed and predicted values. It tests the hypothesis H_0: there is no difference between the predicted and observed values against H_1: there is a difference between the predicted and observed values. To assess performance of the predictive model, measures of diagnostic accuracy such as sensitivity, specificity, positive predictive values, negative predictive values and area under the ROC curve (AUC) were computed. The predictive model with AUC value closer to 1 has a high discriminating power. That is, it has a high ability to correctly distinguish between a patient with baseline acute respiratory distress and a patient without the condition. On the other hand, an AUC value of 0.5 shows that the predictive model makes random choices whereas AUC value below 0.5 indicates that the predictive model is not useful. Classification and regression tree (CART) was also used to predict the outcome of interest based on the presented predictor variables. Classification and regression tree methodology is one of the oldest and most fundamental algorithms. It is used to predict outcomes based on certain predictor variables. The classification and regression tree is also used in machine learning to create predictive models that can be used to make predictions about data.

All the analyses were implemented in Stata Software Package version 17.0 and R Software Package version 4.2.1.

3.4 Ethical Considerations

AQUAMAT study ethical approval registered under ISRCTN50258054, was obtained from each participating institutional or national ethics committee in addition to the Oxford Tropical Research Ethics committee (Dondorp et al., 2010). The use of data for this study was approved by the Oxford-Mahidol research Unit Data Access Committee through an application for ‘Datasets under the Custodianship of Mahidol Oxford Tropical Medicine Research Unit (MORU) Tropical Network’.

CHAPTER 4

RESULTS

4.1 Exploratory Data Analysis

This study is aimed at predicting baseline acute respiratory distress in African children who have severe malaria. The study used AQUAMAT dataset comprising of 5426 children (<15 years) who had been hospitalized for severe malaria from eleven centres from nine African participating countries, namely; Mozambique, The Gambia, Ghana, Kenya, Tanzania, Nigeria, Uganda, Rwanda, and the Democratic Republic of the Congo (Dondorp et al., 2010). These patients were randomly assigned to two treatment arms, namely; the artesunate arm and the quinine arm. The baseline characteristics of the subjects were collected at admission as presented in Table 3.

Table 3 shows that there were no significant differences between patients assigned to the artesunate arm and quinine arm. This shows that recruitment of study participants into the treatment arms was balanced.

Table 3: *Baseline characteristics of patients recruited in the AQUAMAT trial*

Variable	**Total**	**Artesunate**	**Quinine**
Sample size	N=5,426	N=2,713	N=2,713
Country			
Congo	422 (8%)	212 (8%)	210 (8%)
Gambia	502 (9%)	250 (9%)	252 (9%)

Ghana	436 (8%)	218 (8%)	218 (8%)
Kenya	442 (8%)	219 (8%)	223 (8%)
Mozambique	664 (12%)	332 (12%)	332 (12%)
Nigeria	450 (8%)	226 (8%)	224 (8%)
Rwanda	386 (7%)	194 (7%)	192 (7%)
Tanzania	1,461 (27%)	729 (27%)	732 (27%)
Uganda	663 (12%)	333 (12%)	330 (12%)
Patient age in years (Median, IQR)	2 (1-4)	2 (1-4)	2 (1-4)
Weight (kg) (Mean, SD)	12 (5)	12 (5)	13 (5)
Sex			
female	2,611 (48%)	1,316 (49%)	1,295 (48%)
male	2,815 (52%)	1,397 (51%)	1,418 (52%)
Respiratory rate (per minute) (Mean, SD)	47 (14)	47 (14)	47 (14)
Systolic blood pressure (mmHg) (Mean, SD)	95 (14)	95 (14)	95 (14)
Diastolic blood pressure (mmHg) (Mean, SD)	56 (13)	56 (13)	56 (13)
Pneumonia	447 (8%)	225 (8%)	222 (8%)
Sepsis	653 (12%)	300 (11%)	353 (13%)
Symptomatic severe anaemia (severe pallor combined with respiratory distress)	2,213 (41%)	1,131 (42%)	1,082 (40%)
Respiratory distress: Costal indrawing/recession, respiratory insufficiency	867 (16%)	439 (16%)	428 (16%)
Coma at admission (GCS <= 10, BCS <= 2)	1,823 (34%)	881 (32%)	942 (35%)
Convulsions > 30 minutes	1,692 (31%)	812 (30%)	880 (32%)
Hyperparasitaemia (>500 parasites per high powered field)	100 (2%)	44 (2%)	56 (2%)
Compensated shock (capillary refill >= 3 sec temperature gradient)	485 (9%)	233 (9%)	252 (9%)
Decompensated shock (Adults: systolic BP< 80mmHg, Children: systolic BP < 70 mmHg)	178 (3%)	90 (3%)	88 (3%)

Suspected severe acidosis: Deep breathing	938 (17%)	443 (16%)	495 (18%)
Blood transfusion	2,982 (55%)	1,487 (55%)	1,495 (55%)
Mechanical ventilation	55 (1%)	23 (1%)	32 (1%)
Currently treated for chronic illness	39 (1%)	16 (1%)	23 (1%)
Renal failure	16 (0%)	8 (0%)	8 (0%)
Severe prostration (Not able to breastfeed < 6m, or able to sit > 6m)	2,974 (55%)	1,505 (55%)	1,469 (54%)

Data are presented as mean (SD) or median (IQR) for continuous measures, and n (%) for categorical measures.

Table 4 indicates that, of the 5426 patients recruited in the trial, 867 (15.98%) had respiratory distress whereas 4559 (84.02%) did not have respiratory distress. Table 3 shows that there was no significant difference between children with respiratory distress in the artesunate arm and those in the quinine arm. That is, 439 (16%) children in the artesunate arm and 428 (16%) in the quinine arm.

Table 4: *Children with respiratory distress in the AQUAMAT trial*

Respiratory distress	**Frequency**	**Percent**	**Cumulative**
No	4,559	84.02	84.02
Yes	867	15.98	100.00
Total	5,426	100.00	

Exploring patients' age distribution by gender shows that females have the mean age of 2.85 years with a standard deviation of 2.32 years while males have the mean age of 2.92 years with a standard deviation of 2.38 years. The maximum registered age in either sex was 14 years. These age statistics reflect no significant difference between female and male patients recruited in the trial as presented in Table 5.

Table 5: *Age statistics of female and male patients in the AQUAMAT trial*

Sex	Mean	SD	Max	Min
female	2.85	2.32	14	0
male	2.92	2.38	14	0
Total	2.89	2.35	14	0

Figure 1 is a bar graph that depicts proportion of patients with baseline acute respiratory distress who died or survived. The graph shows that a larger proportion of patients diagnosed with acute respiratory distress died as compared to a smaller proportion that survived.

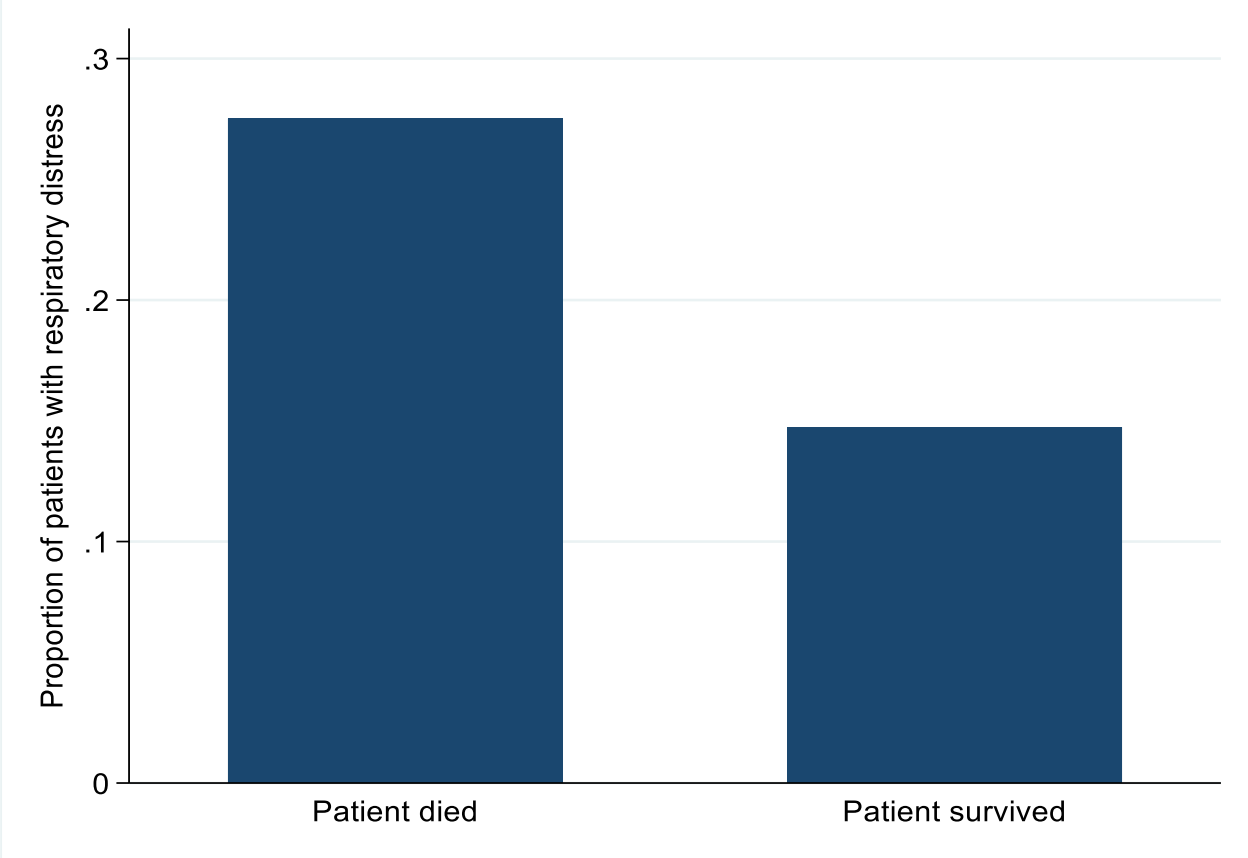

Figure 1: Proportion of patients with acute respiratory distress who died or survived

It is evident from Figure 1 that acute respiratory distress is predictive of high mortality in severe malaria children. Therefore, there is great need to determine risk factors associated with baseline acute respiratory distress and generate a predictive model so as to better manage the condition and be able to develop preventative and early intervention measures.

The histogram in Figure 2 shows that age of patients recruited in the AQUAMAT trial is skewed to the right. This implies that a lot of patients had their ages concentrated around 0 to 5 years with about 23% of them being around 2 years old. Less than 3% of the

patients were aged between 10 and 15 years in the trial. Presenting median (IQR) in the table of baseline characteristics is reasonable for this non-normal kind of distribution.

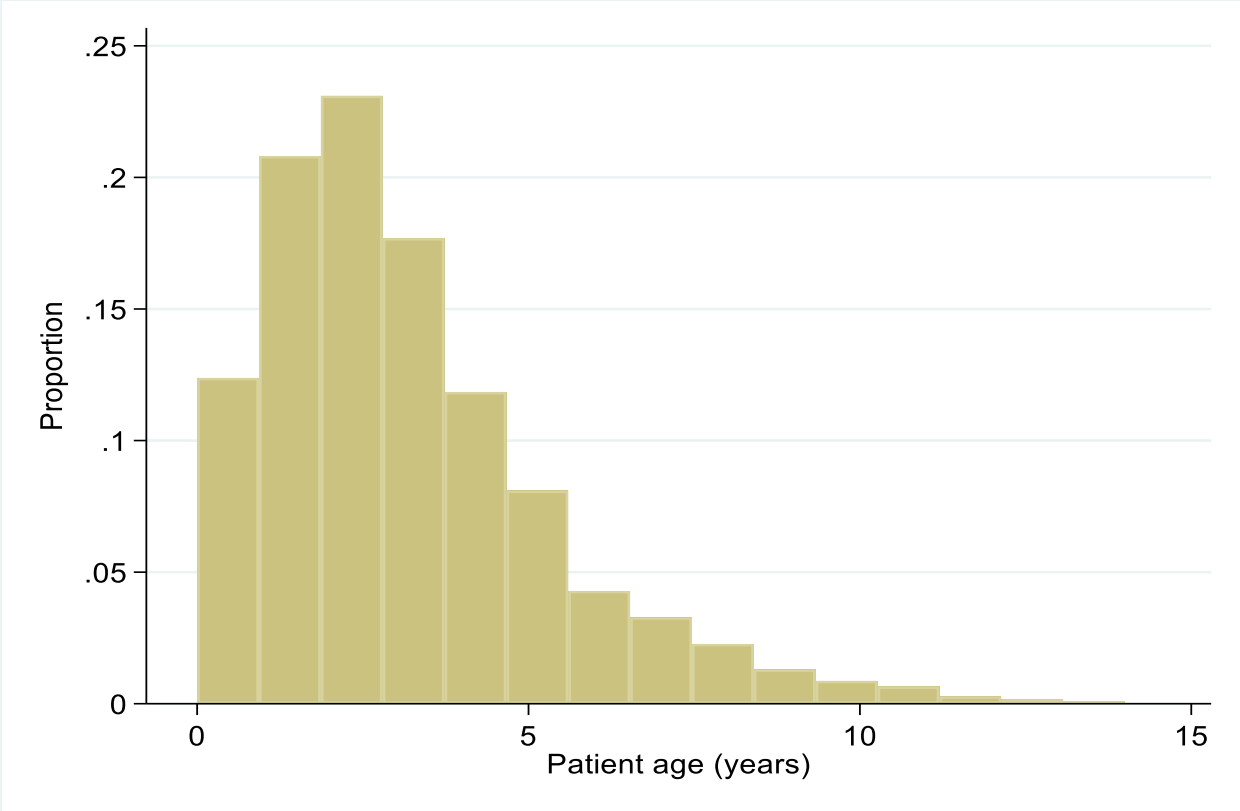

Figure 2: Histogram of patients' age (in years) distribution

Results in figure 2 imply that this study focuses on baseline acute respiratory distress in children who are different from adults, in terms of lung maturation with age, developmental stages, epidemiology, comorbidities, and prognosis.

Figure 3 shows that weight (kg) of the patients recruited in the trial was slightly skewed to the right. Most of the patients had their weight between 5 kg and 20 kg with the largest proportion around 10 kg to 12 kg. A very small proportion had their weights between 25 kg and 35 kg. This is typical for this study which involves children.

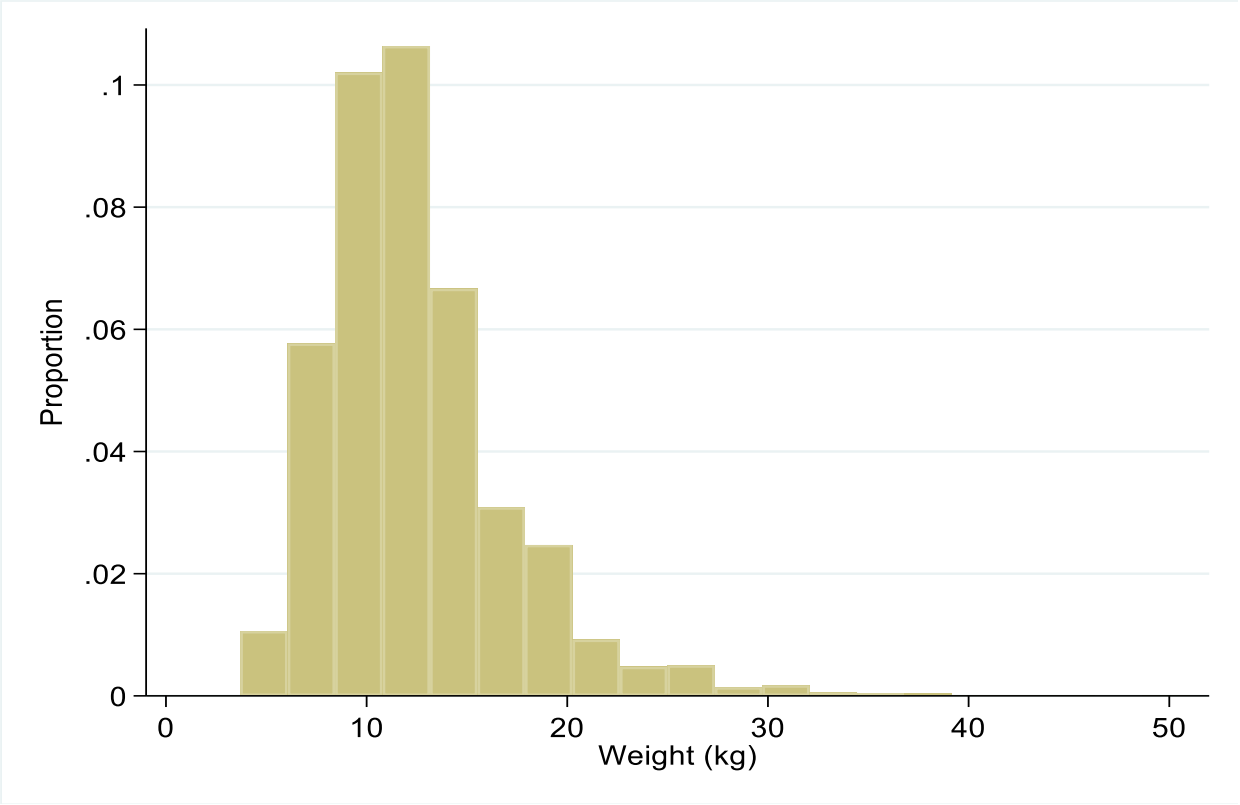

Figure 3: Weight (kg) of patients in AQUAMAT trial

Figure 4 looks into respiratory rate (per minute) of patients recruited in the AQUAMAT trial. The histogram shows that most of the patients in the study had their respiratory rate between 30 and 60 per minute with the largest proportion having respiratory rate of 42 per minute. The distribution seems to be approximately mound-shaped.

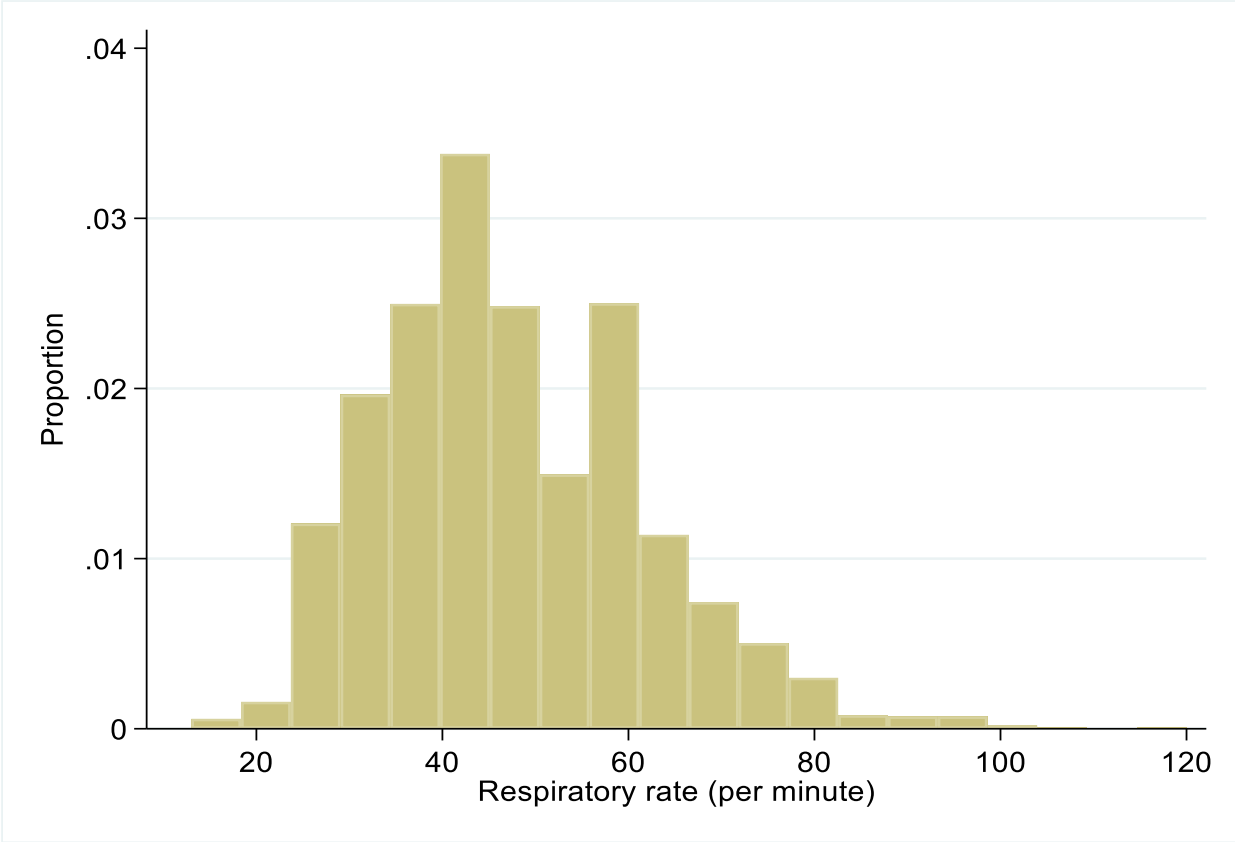

Figure 4: Respiratory rate (per minute) of patients recruited in the AQUAMAT trial

Figure 5 is a histogram presenting systolic blood pressure (mmHg) of patients recruited in the trial. The distribution is approximately normally distributed. Patients have their systolic blood pressure between 50 mmHg and 150 mmHg. The largest proportion of the patients have their systolic blood pressure around 90 mmHg.

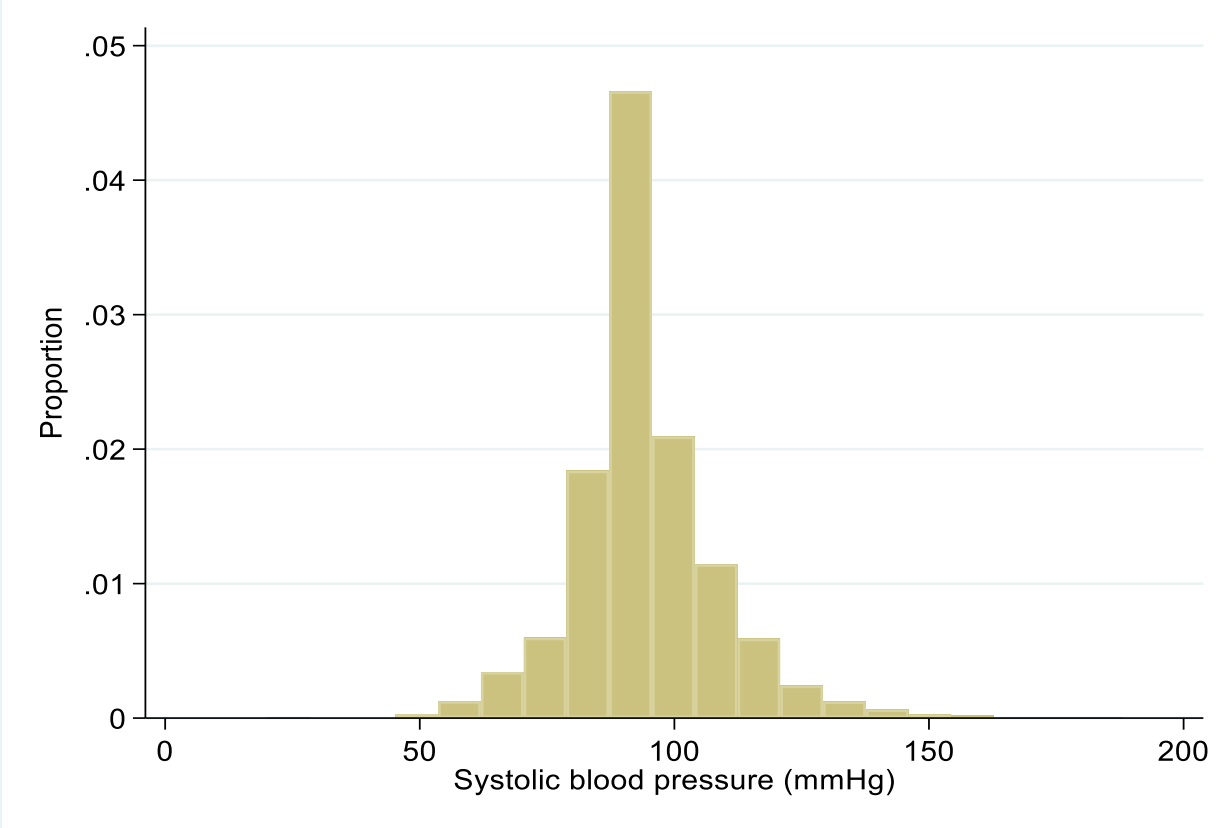

Figure 5: Systolic blood pressure (mmHg) of patients recruited in the AQUAMAT trial

Figure 6 presents diastolic blood pressure (mmHg) of patients recruited in the AQUAMAT trial. The distribution is approximately normally distributed. Patients have their diastolic blood pressure between 20 mmHg and 100 mmHg. The largest proportion of the patients have their diastolic blood pressure around 55 mmHg.

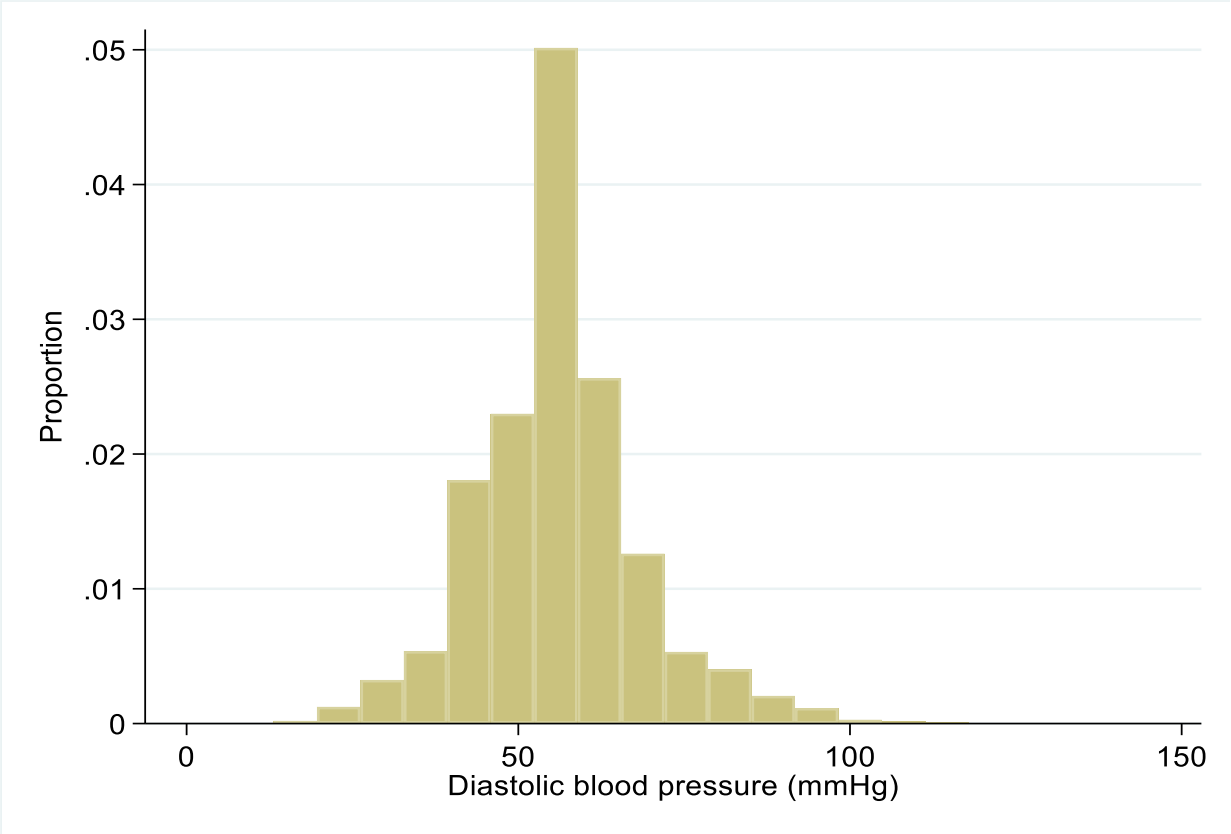

Figure 6: Diastolic blood pressure (mmHg) of patients recruited in the AQUAMAT trial

A bar graph in figure 7 considers the proportion of females and males who presented acute respiratory distress in the trial. The figure indicates that there were slightly more females who presented respiratory distress: costal indrawing/recession/respiratory insufficiency as compared to their male counterparts.

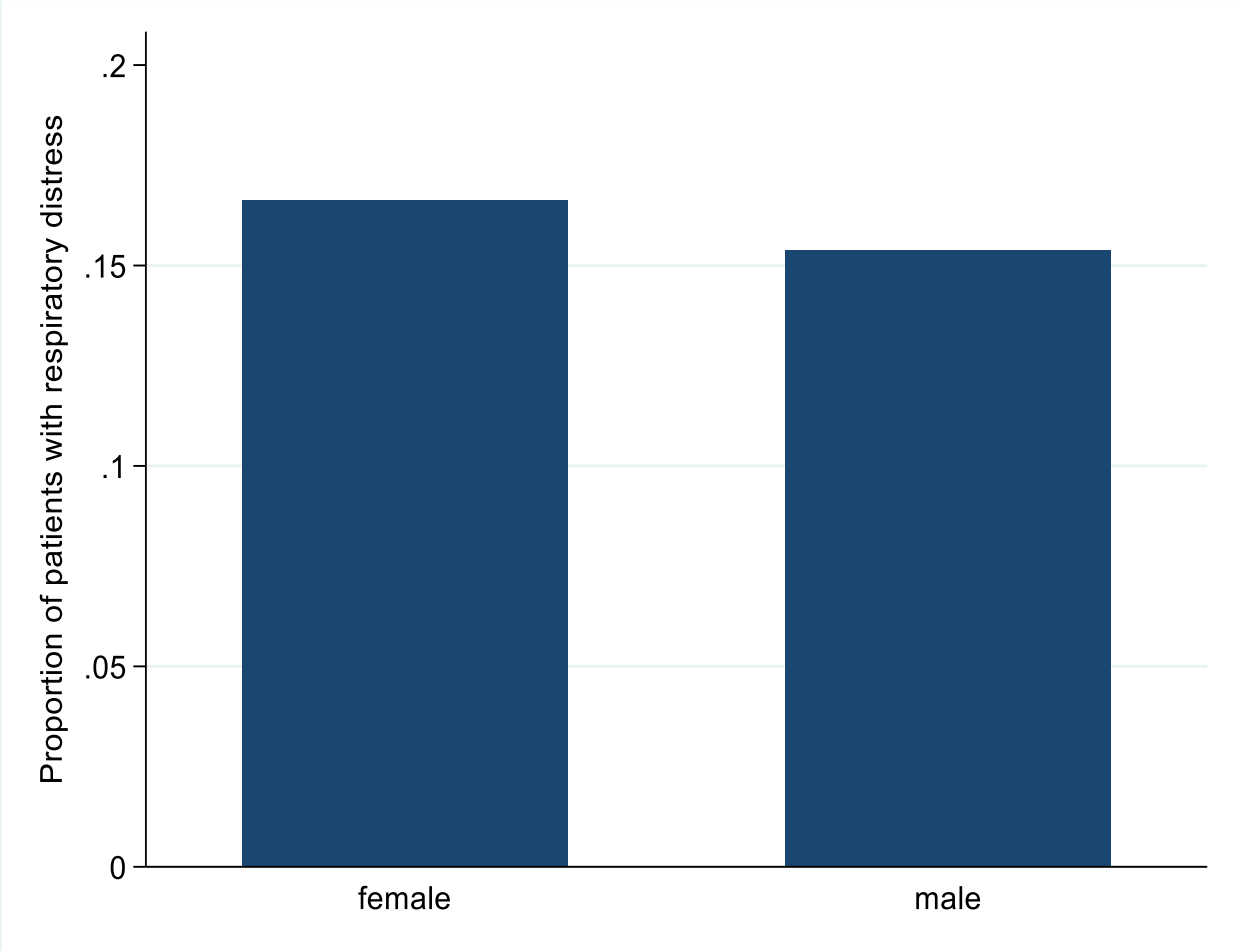

Figure 7: Proportion of patients with respiratory distress against sex

Figure 7 may suggest that acute respiratory distress has a slightly disproportionate effect on male and female patients with more female patients being affected as compared to male patients. However, the difference is marginal.

4.2 Analysis of Predictors of Baseline Acute Respiratory Distress

The association between baseline acute respiratory distress and each of the predictor variables is assessed by fitting a univariable binary logistic regression model. Table 6 shows the results obtained by regressing baseline acute respiratory distress and each of the predictor variables.

Results in table 6 show that the predictor variables; patient age (years), weight (kg), respiratory rate (per minute), diastolic blood pressure (mmHg), pneumonia, sepsis, symptomatic severe anaemia (severe pallor combined with respiratory distress), coma at admission (GCS $\leq$ 10, BCS $\leq$ 2), convulsions > 30 minutes, decompensated shock (adults: systolic BP< 80 mmHg, children: systolic BP < 70 mmHg), hyperparasitaemia (> 500 parasites per high powered field), severe acidosis: deep breathing, blood transfusion, mechanical ventilation, if a patient is currently treated for chronic illness and severe prostration (not able to breastfeed < 6 m, or able to sit > 6 m) are individually significant and are associated with the prediction of baseline acute respiratory distress in severe malaria children. They have p-values less than 0.05. On the other hand, sex, systolic blood pressure (mmHg), compensated shock (capillary refill $\geq$ 3 sec temperature gradient), and renal failure are individually not significant. Their p-values are greater than 0.05.

Table 6: *Univariable model analysis for the relationship between baseline acute respiratory distress and individual predictor variable*

Variable	Odds ratio (95% CI)	P-Value
Patient age (years)	0.84 (0.80, 0.87)	<0.001
Sex		
male	0.91 (0.79, 1.05)	0.213
Weight (kg)	0.91 (0.89, 0.93)	<0.001
Respiratory rate (per minute)	1.04 (1.04, 1.05)	<0.001
Systolic blood pressure (mmHg)	1.00 (0.99, 1.00)	0.209
Diastolic blood pressure (mmHg)	0.99 (0.99, 1.00)	0.022
Pneumonia	3.56 (2.90, 4.39)	<0.001
Sepsis	1.58 (1.30, 1.94)	<0.001
Severe anaemia	1.67 (1.45, 1.94)	<0.001
Coma at admission	1.32 (1.14, 1.54)	<0.001
Convulsions > 30 minutes	0.74 (0.63, 0.88)	<0.001

Compensated shock	1.06 (0.83, 1.36)	0.649
Decompensated shock	1.94 (1.38, 2.73)	<0.001
Hyperparasitaemia	2.19 (1.41, 3.39)	<0.001
Severe acidosis	3.32 (2.82, 3.91)	<0.001
Blood transfusion	1.47 (1.26, 1.70)	<0.001
Mechanical ventilation	2.18 (1.21, 3.92)	0.009
Currently treated for chronic illness	2.08 (1.03, 4.19)	0.041
Renal failure	1.21 (0.35, 4.27)	0.762
Severe prostration	0.65 (0.56, 0.76)	<0.001

Predictor variables which are statistically significant at 5% level of significance, from a univariable binary logistic regression model are used to generate a predictive model of baseline acute respiratory distress. This predictive model is a multivariable binary logistic regression model. Table 7 presents results of a multivariable analysis of admission features and their effect in predicting baseline acute respiratory distress in children with severe malaria.

From multivariable analysis in Table 7, respiratory rate (per minute), pneumonia, sepsis, convulsions > 30 minutes, hyperparasitaemia (>500 parasites per high powered field), severe acidosis: deep breathing, if a patient is currently treated for chronic illness and severe prostration (not able to breastfeed < 6 m, or able to sit > 6 m) are significant predictors of baseline acute respiratory distress in severe malaria children.

A one-unit increase in respiratory rate (per minute) of a patient with severe malaria leads to a 1.03 times increased risk of developing baseline acute respiratory distress (OR: 1.03, CI: 1.03 - 1.04, p-value < 0.001). A patient who presents pneumonia on admission has 2.49 times increased risk of developing baseline acute respiratory distress as compared to

a patient without pneumonia (OR: 2.49, CI: 1.99 - 3.13, p-value < 0.001). A severe malaria patient with sepsis on admission has 1.46 times increased risk of developing baseline acute respiratory distress as compared to a patient without sepsis on admission (OR: 1.46, CI: 1.18 - 1.82, p-value $= 0.001$). A patient with hyperparasitaemia (>500 parasites per high powered field) on admission has 1.96 times increased risk of developing baseline acute respiratory distress as compared to a patient without hyperparasitaemia on admission (OR: 1.96, CI: 1.21 - 3.16, p-value $= 0.006$). A severe malaria patient with convulsions > 30 minutes on admission has a 23% reduced risk of developing baseline acute respiratory distress as compared to a patient not presenting convulsions > 30 minutes on admission (OR: 0.77, CI: 0.63 - 0.93, p-value $= 0.007$). A patient presenting severe acidosis: deep breathing on admission has 2.49 times increased risk of developing baseline acute respiratory distress as compared to a patient not presenting severe acidosis (OR: 2.49, CI: 2.09 - 2.97, p-value < 0.001). A patient who is currently being treated for chronic illness has 2.32 increased risk of developing baseline acute respiratory distress as compared to a patient not currently treated for chronic illness (OR: 2.32, CI: 1.05 - 5.14, p-value $= 0.038$). A patient with severe prostration (not able to breastfeed < 6 m, or able to sit > 6 m) has 31% reduced risk of developing baseline acute respiratory distress as compared to a patient without severe prostration on admission (OR: 0.69, CI: 0.55 - 0.88, p-value $= 0.003$).

The predictive model shows that the greatest predictors of baseline acute respiratory distress in severe malaria African children are pneumonia, severe acidosis, if a patient is currently treated for chronic illness, hyperparasitaemia, sepsis, respiratory rate (per minute), convulsions > 30 minutes and severe prostration, in that order.

On the other hand, patient age (in years), weight (kg), diastolic blood pressure (mmHg), symptomatic severe anaemia (severe pallor combined with respiratory distress), coma at admission (GCS $\leq$ 10, BCS $\leq$ 2), decompensated shock (Adults: systolic BP< 80 mmHg, Children: systolic BP < 70 mmHg), blood transfusion and mechanical ventilation do not significantly predict the development of baseline acute respiratory distress in severe malaria African children.

To validate the constructed prediction model, Hosmer-Lemeshow goodness-of-fit test was carried out. This statistical test measures the correspondence of the observed and predicted values of the outcome variable (baseline acute respiratory distress). A better model fit is characterized by insignificant differences between the observed and predicted values. It tests the hypothesis H_0: there is no difference between the predicted and observed values against H_1: there is a difference between the predicted and observed values. With the p-value of 0.9935 in Table 7, we fail to reject the null hypothesis and conclude that there is no significant difference between the observed and predicted values of the predictive model, suggesting that the model fitted the data well.

Table 7: *Multivariable analysis (predictive model analysis) of admission features and their effect in predicting baseline acute respiratory distress in children with severe malaria.*

Variable	Odds ratio (95% CI)	P-Value
Patient age (years)	0.98 (0.91, 1.06)	0.629
Weight (kg)	0.97 (0.94, 1.01)	0.178
Respiratory rate (per minute)	1.03 (1.03, 1.04)	<0.001
Diastolic blood pressure (mmHg)	1.00 (1.00, 1.01)	0.338
Pneumonia	2.49 (1.99, 3.13)	<0.001
Sepsis	1.46 (1.18, 1.82)	0.001

Severe anaemia	1.12 (0.91, 1.37)	0.287
Coma at admission	1.26 (0.97, 1.64)	0.088
Hyperparasitaemia	1.96 (1.21, 3.16)	0.006
Convulsions > 30 minutes	0.77 (0.63, 0.93)	0.007
Decompensated shock	1.36 (0.93, 2.00)	0.116
Severe acidosis	2.49 (2.09, 2.97)	<0.001
Blood transfusion	0.98 (0.80, 1.19)	0.807
Mechanical ventilation	1.43 (0.76, 2.69)	0.269
Currently treated for chronic illness	2.32 (1.05, 5.14)	0.038
Severe prostration	0.69 (0.55, 0.88)	0.003
Constant	0.04 (0.02, 0.07)	<0.001
	χ^2	**P-Value**
Goodness-of-fit test		
Hosmer-Lemeshow	30.17	0.9935

In order to visualise the predictive model, a nomogram was plotted in Figure 8. A nomogram ranks the importance of a predictor variable in predicting the outcome (baseline acute respiratory distress) in the context of the other predictor variables in the model. Each of the sixteen predictor variables included in the predictive model were arranged one by one on a horizontal plane with its scoring system, ranging from 0 to 10, at the bottom. The total score ranged from 0 to 27.

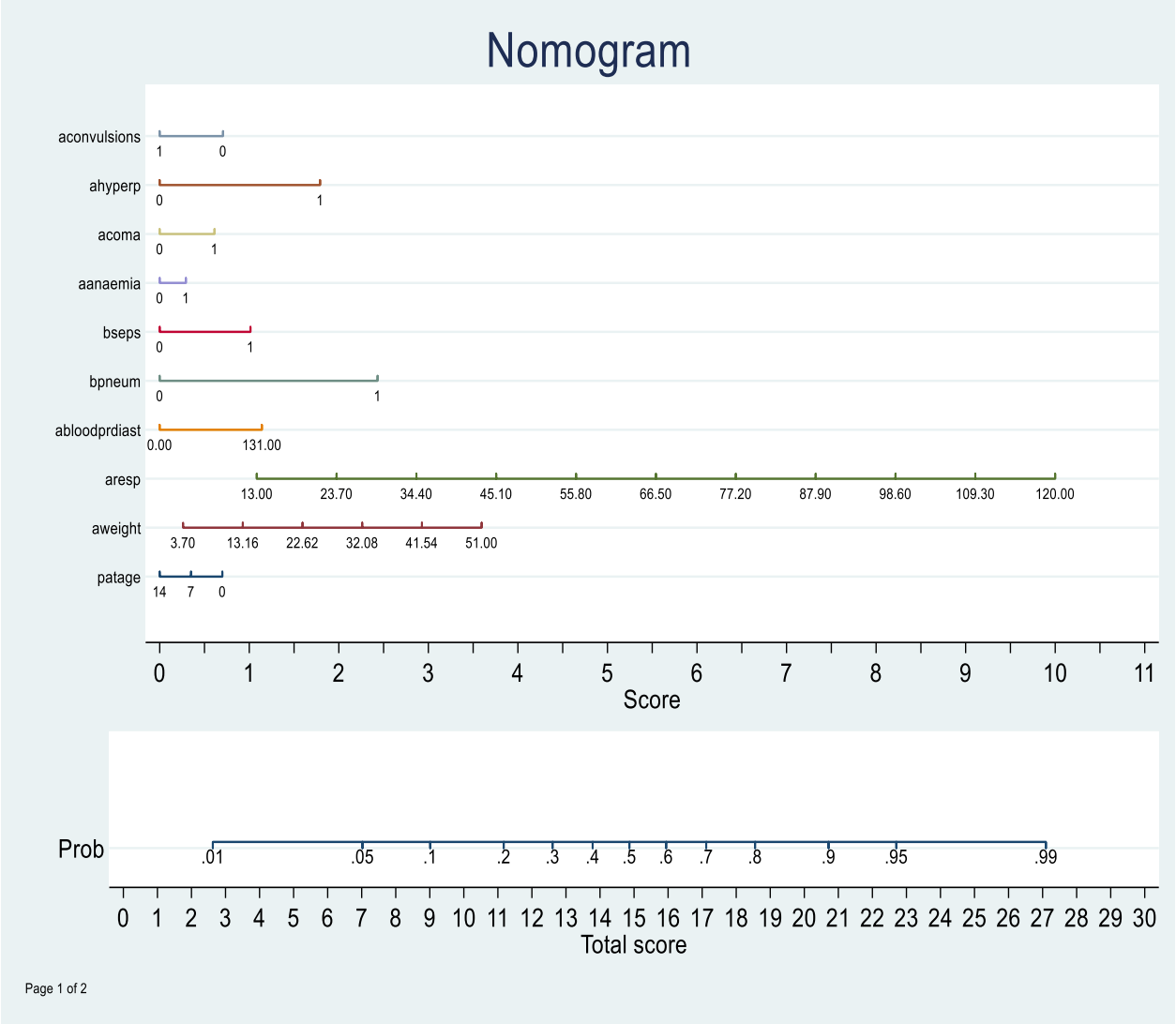

Figure 8: Nomogram for prediction of baseline acute respiratory distress

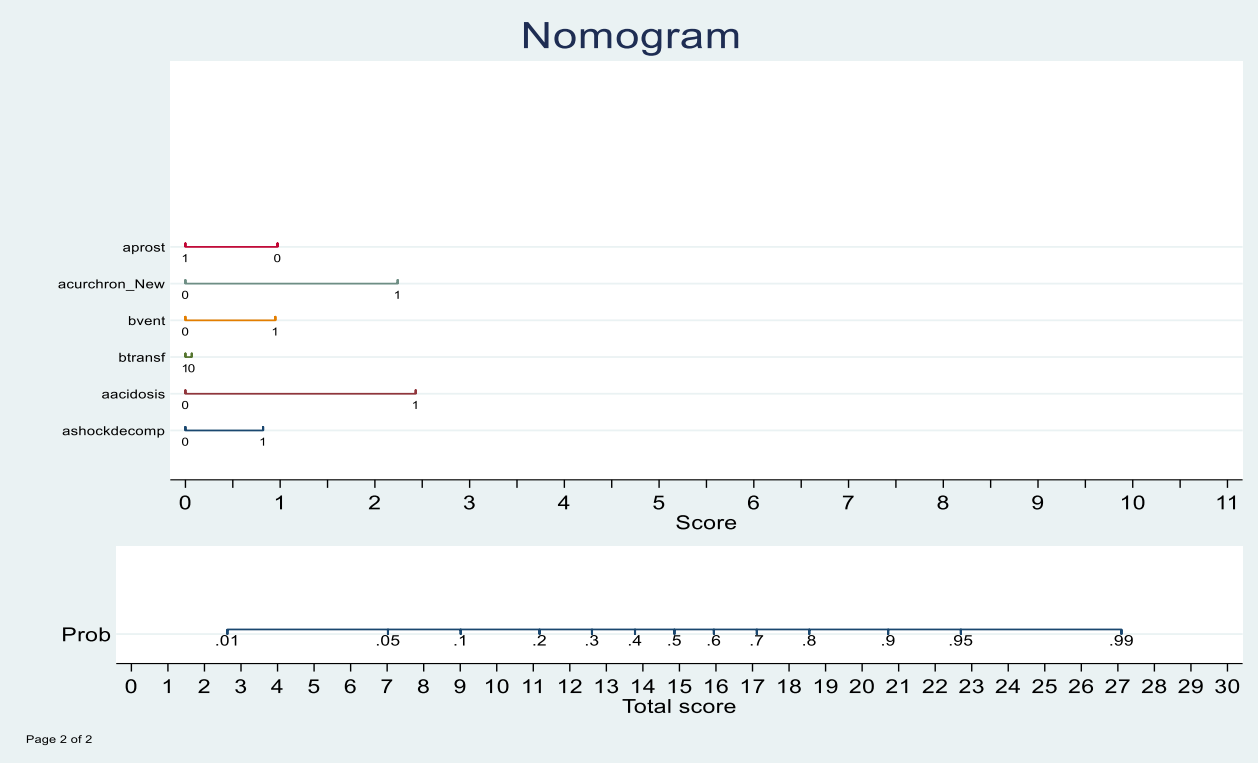

Figure 9: *Continued*

Results in Figure 8 show predictors in the predictive model ranked with their corresponding scores. Table 8 lists these predictors in order of importance (from highest to lowest) in predicting baseline acute respiratory distress.

Table 8: *Rank of predictors of baseline acute respiratory distress*

Predictor	**Score**
Respiratory rate (per minute)	10
Weight (Kgs)	3.6
Severe acidosis	2.5
Pneumonia	2.5
Currently treated for chronic illness	2.4
Hyperparasitaemia	2
Diastolic blood pressure	1.2
Sepsis	1
Mechanical ventilation	1

Severe prostration	1
Decompensated shock	0.8
Convulsions > 30 minutes	0.75
Patient age (Years)	0.75
Coma	0.6
Severe anaemia	0.4
Blood transfusion	0.1

4.3 Analysis of Sensitivity, Specificity, Positive and Negative Predictive Values

To assess the predictive ability of a model, diagnostic accuracy measures such as sensitivity, specificity, positive predictive values and negative predictive values are used. First, an optimal probability cutoff point is generated by plotting graphs of sensitivity and specificity on the same axes as in Figure 9.

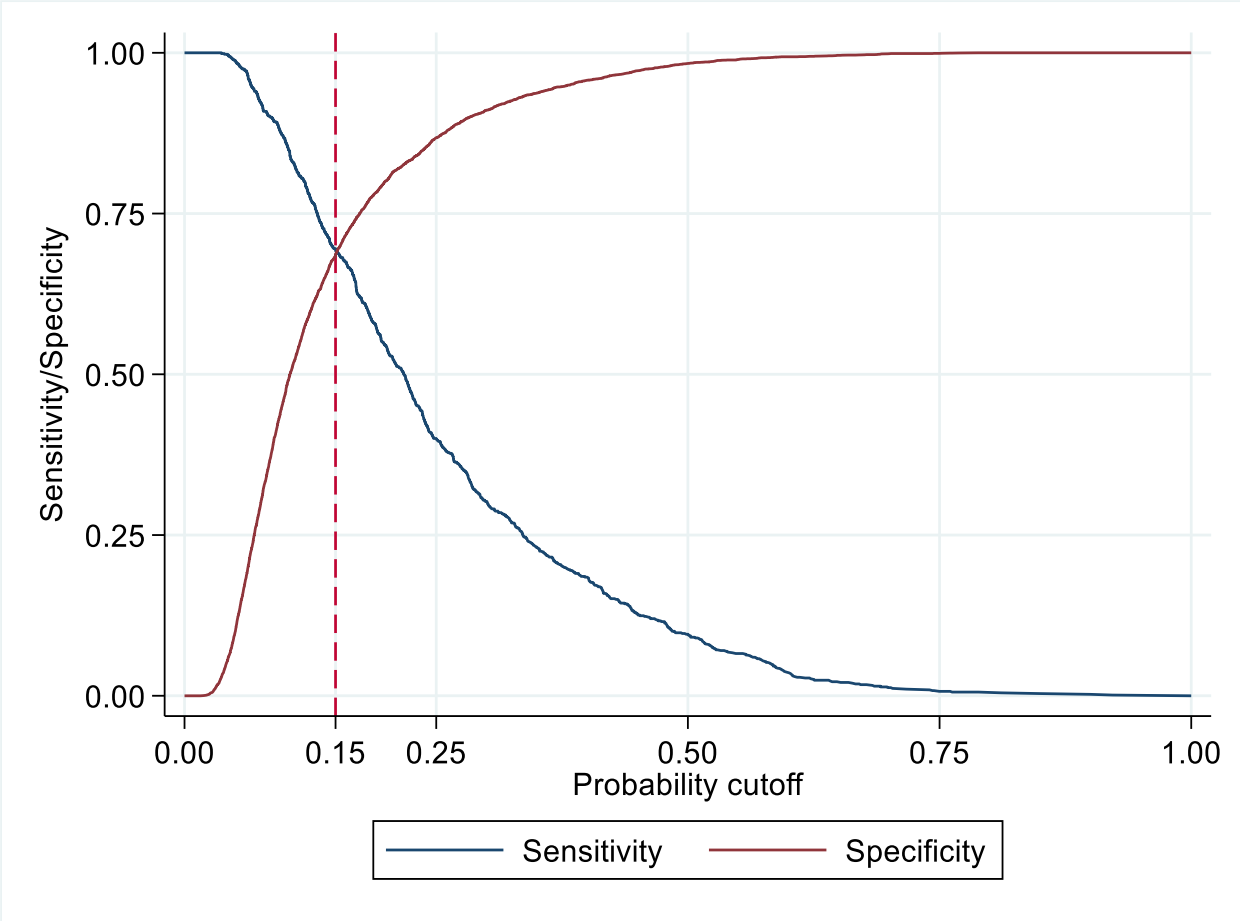

Figure 10: Optimal probability cutoff point

In Figure 9, optimal probability cutoff is found at the point where the graph of sensitivity and specificity intersect. In this case, optimal probability cutoff is estimated to be 0.15. This value is used as a cutoff point when calculating sensitivity, specificity, positive predictive values and negative predictive values. These metrics are presented in Table 9.

Table 9 shows that there are 603 true positives, 1433 false positives, 264 false negatives and 3126 true negatives. In the 5426 patients recruited in the study, a total of 867 had baseline acute respiratory distress while 4559 did not have baseline acute respiratory distress. This is the gold standard test. The predictive model classified a total 2036

patients as positives and 3390 as negatives. The overall rate of correct classification for the predictive model is estimated to be 68.72%, with 69.55% of the patients with baseline acute respiratory distress correctly classified positive for the disease (sensitivity) and 68.57% of the patients without baseline acute respiratory distress correctly classified negative for the condition (specificity). Table 9 also indicates that 29.62% of the patients are classified as having baseline acute respiratory distress given that the predictive model result is positive (positive predictive value) and 92.21% of the patients are classified as not having baseline acute respiratory distress given that the predictive model result is negative (negative predictive value).

Table 9: *Sensitivity, specificity, positive and negative predictive values*

Classified	**True**		**Total**
	Disease (D)	No disease ($\bar{D}$)	
Positive (T)	603	1433	2036
Negative ($\bar{T}$)	264	3126	3390
Total	867	4559	5426

Classified positive if predicted $\Pr(D) \geq 0.15$

True D defined as baseline acute respiratory distress $\neq 0$

Sensitivity	$\Pr(T\|D)$	69.55%
Specificity	$\Pr(\bar{T}\|\bar{D})$	68.57%
Positive predictive value	$\Pr(D\|T)$	29.62%
Negative predictive value	$\Pr(\bar{D}\|\bar{T})$	92.21%
False + rate for true $\bar{D}$	$\Pr(T\|\bar{D})$	31.43%
False − rate for true D	$\Pr(\bar{T}\|D)$	30.45%
False + rate for classified +	$\Pr(\bar{D}\|T)$	70.38%
False − rate for classified −	$\Pr(D\|\bar{T})$	7.79%
Correctly classified		68.72%

4.4 Analysis of the Area Under ROC Curve (AUC)

In order to assess the predictive model's discriminative power, sensitivity against specificity is plotted. That is, to determine the predictive model's ability to distinguish between children with baseline acute respiratory distress and those without baseline acute respiratory distress. This is achieved by examining the shape of a ROC curve and the AUC value. AUC measure is used to assess the diagnostic accuracy or the performance of a predictive model. Figure 10 is the ROC curve with the associated AUC value presented in Table 10.

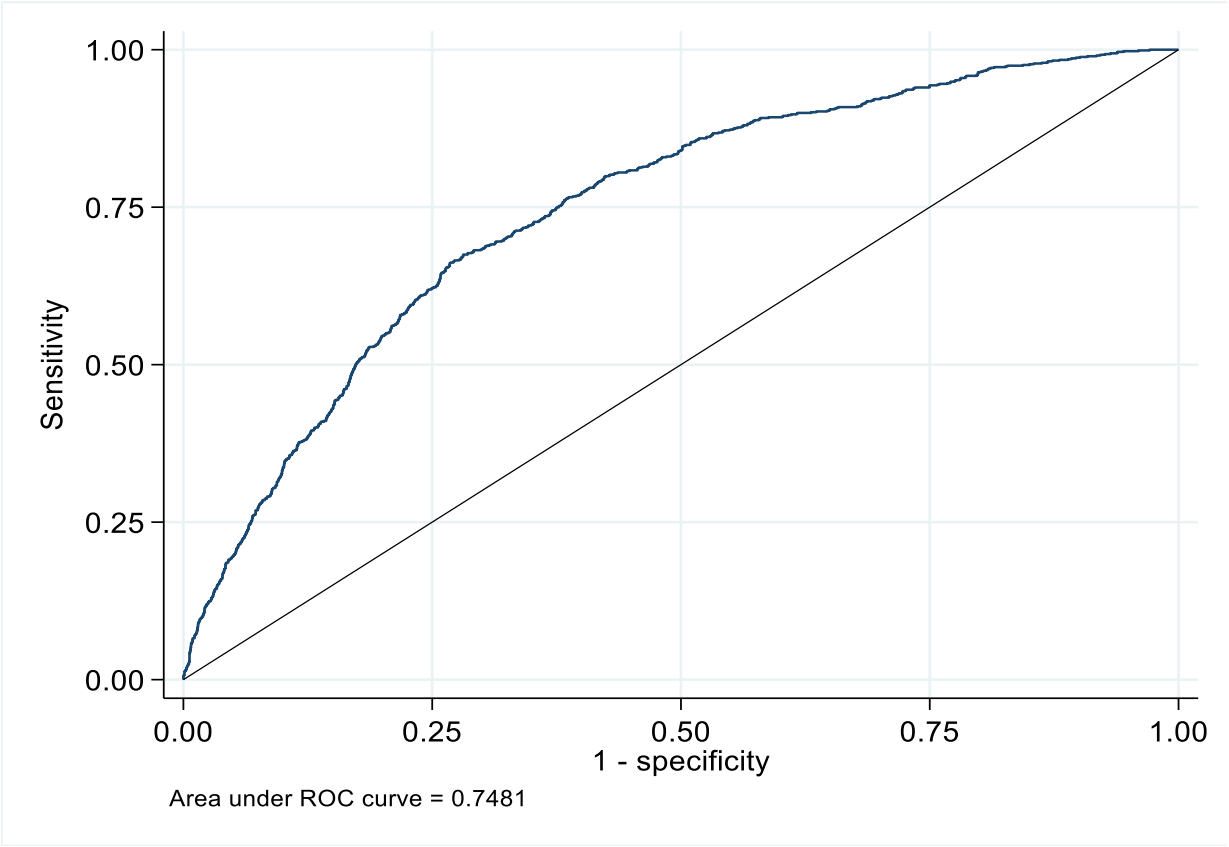

Figure 11: Receiver Operating Characteristic (ROC) curve

Figure 10 shows that the predictive model under consideration has a good ability to distinguish between children with baseline acute respiratory distress and those without baseline acute respiratory distress. This is evidenced by the curve's proximity to the upper-left corner.

The curve is Figure 10 has an area of 0.75 (95% CI: 0.73 - 0.77) under it as presented in Table 10. This indicates that the predictive model is good at classifying severe malaria patients as having baseline acute respiratory distress or not, i.e., a 0.75 rate indicates that the predictions are not by random choice.

Table 10: *Area under the ROC curve*

Observations	Area under ROC curve (95% CI)	Std. error
5,426	0.75 (0.73, 0.77)	0.009

4.5 Classification Tree Analysis

Classification trees are prediction models constructed by recursively partitioning a data set and fitting a simple model to each partition. The goal, in this study, is to find a model for predicting if a patient is at higher or lower risk of developing baseline acute respiratory distress depending on the value(s) of the clinical factors he/she is presenting. A classification tree is presented in Figure 11.

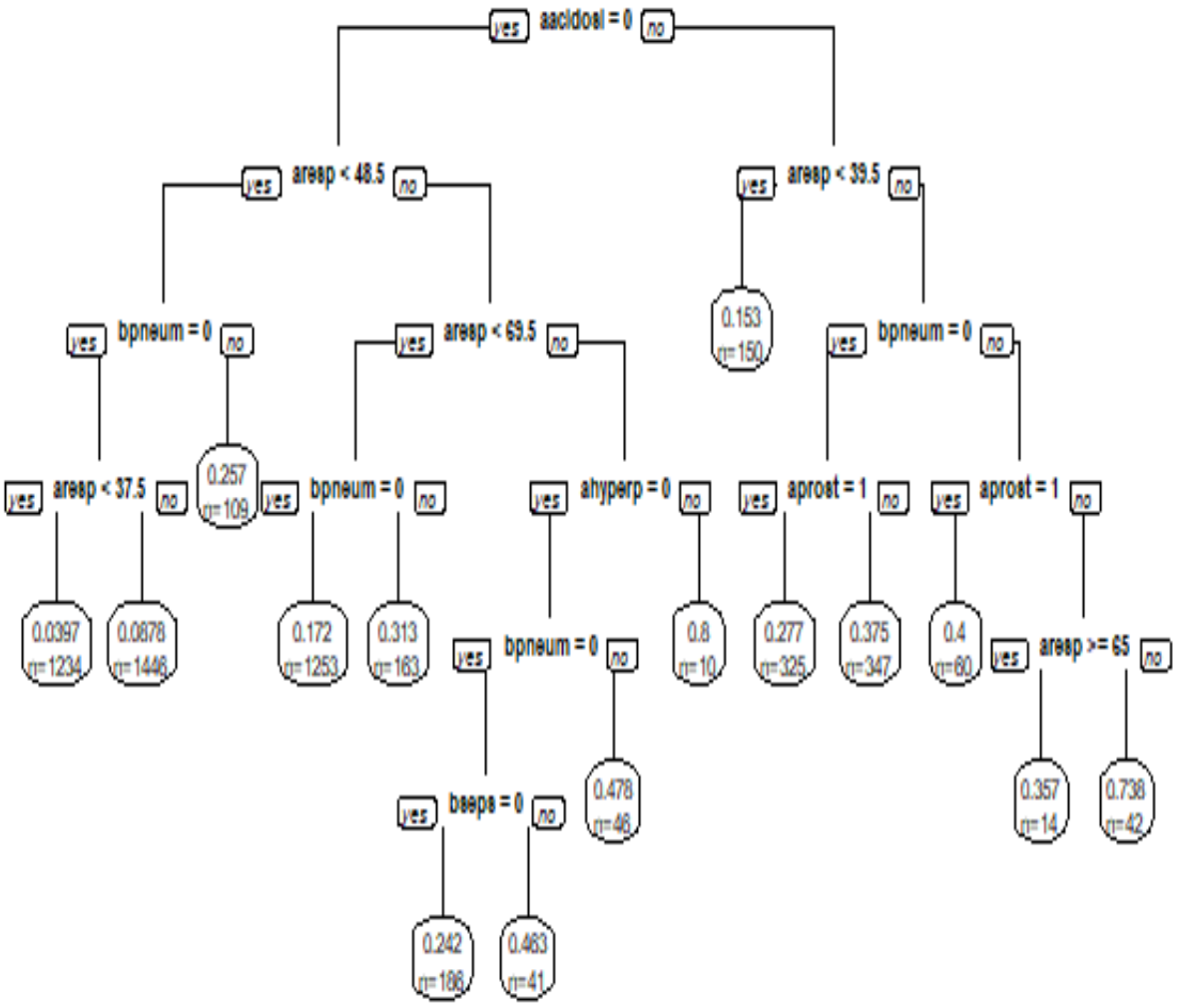

Figure 12: Classification tree for predicting baseline acute respiratory distress

Figure 11 is a classification tree created from a sample of 5426 observations partitioned into different branches depending on conditions presented by patients. The classification tree classifies patients as having a higher risk or lower risk of developing baseline acute respiratory distress. There is a splitting-criteria at each node of the tree. The value of n at each terminal node indicates the number of patients who fall in that category based on the

conditions they present on admission. The proportion at each terminal node represents the proportion of patients who are classified as having baseline acute respiratory distress out of n. Thus, higher and lower proportions are indicative of higher and lower risks, respectively, of developing baseline acute respiratory distress.

Results show that 10 patients are classified as having respiratory rate of at least 69.5 per minute, having hyperparasitaemia (>500 parasites per high powered field) and without severe acidosis: deep breathing. These patients have 80% increased risk of developing baseline acute respiratory distress (proportion $= 0.8, n = 10$). This is followed by 42 patients who are classified as having respiratory rate of 39.5 - 65 per minute, presenting severe acidosis: deep breathing, presenting pneumonia and without severe prostration (not able to breastfeed < 6 m, or able to sit > 6 m). These patients have 73.8% increased risk of developing baseline acute respiratory distress (proportion $= 0.738, n = 42$). Results also show that 46 patients are classified as having respiratory rate of at least 69.5 per minute, having pneumonia, without severe acidosis and without hyperparasitaemia (>500 parasites per high powered field). These patients have 47.8% higher risk of developing baseline acute respiratory distress (proportion $= 0.478, n = 46$).

Figure 11 indicates that 41 patients are classified as having respiratory rate of at least 69.5 per minute, having sepsis, without severe acidosis: deep breathing, without pneumonia and without hyperparasitaemia (>500 parasites per high powered field). These patients have 46.3% higher risk of developing baseline acute respiratory distress (proportion $= 0.463, n = 41$). 60 patients are classified as having respiratory rate of at least 39.5 per minute, having severe acidosis: deep breathing, having pneumonia and having severe prostration (not able to breastfeed < 6 m, or able to sit > 6 m). These

patients have 40% higher risk of developing baseline acute respiratory distress (proportion $= 0.4, n = 60$). 347 patients are classified as having respiratory rate of at least 39.5 per minute, having severe acidosis: deep breathing, without pneumonia and without severe prostration (not able to breastfeed < 6 m, or able to sit > 6 m). These patients have 37.5% higher risk of developing baseline acute respiratory distress (proportion $= 0.375, n = 347$).

Results in figure 11 also shows that 14 patients are classified as having respiratory rate of at least 65 per minute, having severe acidosis: deep breathing, having pneumonia and without severe prostration (not able to breastfeed < 6 m, or able to sit > 6 m). These patients have 35.7% higher risk of developing baseline acute respiratory distress (proportion $= 0.357, n = 14$). 163 patients are classified as having respiratory rate of 48.5 – 69.5 per minute, having pneumonia and without severe acidosis: deep breathing. These patients have 31.3% higher risk of developing baseline acute respiratory distress (proportion $= 0.313, n = 163$). 325 patients are classified as having respiratory rate of at least 39.5 per minute, severe acidosis: deep breathing, severe prostration (not able to breastfeed < 6 m, or able to sit > 6 m) and without pneumonia. These patients have 27.7% higher risk of developing baseline acute respiratory distress (proportion $= 0.277, n = 325$). 109 patients are classified as having respiratory rate less than 48.5 per minute, having pneumonia and without severe acidosis: deep breathing. These patients have 25.7% higher risk of developing baseline acute respiratory distress (proportion $= 0.257, n = 109$).

It is also observed that 186 patients are classified as having respiratory rate of at least 69.5 per minute, without severe acidosis: deep breathing, without pneumonia, without sepsis and without hyperparasitaemia (>500 parasites per high powered field). These patients have 24.2% higher risk of developing baseline acute respiratory distress (proportion $= 0.242, n = 186$). 1253 patients are classified as having respiratory rate of 48.5 – 69.5 per minute, without severe acidosis: deep breathing and without pneumonia. These patients have 17.2% higher risk of developing baseline acute respiratory distress (proportion $= 0.172, n = 1253$). 150 patients are classified as having respiratory rate less than 39.5 per minute and having severe acidosis: deep breathing. These patients have 15.3% increased risk of developing baseline acute respiratory distress (proportion $= 0.153, n = 150$). 1446 patients are classified as having respiratory rate of 37.5 – 48.5 per minute, without severe acidosis: deep breathing and without pneumonia. These patients have 8.8% increased risk of developing baseline acute respiratory distress (proportion $= 0.0878, n = 1446$).

Lastly, 1234 patients are classified as having respiratory rate less than 37.5 per minute, without severe acidosis: deep breathing and without pneumonia. These patients have 4.0% increased risk of developing baseline acute respiratory distress (proportion $= 0.0397, n = 1234$).

CHAPTER 5

DISCUSSION

The wide range of disease states linked to the development of acute respiratory distress in children and the fact that the diagnosis of acute respiratory distress in children is a syndrome rather than a distinct entity with a validated diagnostic confirmatory test add to the inherent difficulties of investigating the development of acute respiratory distress in children. This challenge is exacerbated by limited resources in most African settings and high prevalence of malaria in Sub-Saharan Africa.

This study used data from AQUAMAT trial which was conducted from October 3, 2005, to July 14, 2010, among children (<15 years) who had been hospitalized for severe malaria from eleven centres from nine participating African countries including; Mozambique, The Gambia, Ghana, Kenya, Tanzania, Nigeria, Uganda, Rwanda, and the Democratic Republic of the Congo (Dondorp et al., 2010). This study aimed at predicting baseline acute respiratory distress in African children who have severe malaria. In particular, the study intended to generate a predictive model based on demographic and clinical factors as well as develop a classification tool, based on the conditions presented by the patient, which can be used to predict children at high risk of developing baseline acute respiratory distress so as to timely escalate these cases for further laboratory tests. Identifying risk factors and understanding which patients are at risk of developing

baseline acute respiratory distress is significantly important in order to be able to develop preventative and early intervention mechanisms.

This study found that clinical features associated with baseline acute respiratory distress are pneumonia, severe acidosis: deep breathing, if a patient is currently treated for chronic illness, hyperparasitaemia (>500 parasites per high powered field), sepsis, respiratory rate (per minute), convulsions > 30 minutes and severe prostration (not able to breastfeed < 6 m, or able to sit > 6 m).

Both univariable and multivariable binary logistic regression models show that having pneumonia poses a highest risk of developing baseline acute respiratory distress in severe malaria children. This is so because pneumonia, in itself, is a form of acute respiratory illness and it directly affects the lungs by filling up the alveoli with pus and fluid which limits oxygen intake and makes breathing painful (WHO, 2021). Similar findings were reported by Bellani et al. (2016) from the Large Observational Study to Understand the Global Impact of Severe Respiratory Failure (LUNG SAFE) which recruited a sample of 29,144 patients from 459 ICUs and identified 3022 patients with acute respiratory distress. Of those patients, 59.4% had pneumonia as a risk factor for acute respiratory distress. This study has also shown that severe acidosis: deep breathing is the second ranked risk factor associated with baseline acute respiratory distress in severe malaria children. Lungs and kidneys are the organs which help in maintaining pH balance in the body. Excess of acids has a potential of damaging these organs hence resulting into acute respiratory distress. These findings are similar to what was reported by Mzumara et al.,

(2021) that the signs of acute respiratory distress are commonly associated with severe acidosis, as is the findings of similar studies in The Gambia and Kenya (English et al., 2002; Mzumara et al., 2021).

This study also suggests that patients who are currently treated for chronic illness are at a higher risk of developing baseline acute respiratory distress. Paediatric patients with preexisting chronic illness, such as human immunodeficiency virus (HIV) and cancer, and currently treated for such illnesses, are at an increased risk of developing baseline acute respiratory distress because these patients have worse outcomes, such as increased hospital mortality, and have proportionately more infections as the cause of baseline acute respiratory distress (Cortegiani et al., 2018; Erickson et al., 2007). Hyperparasitaemia (>500 parasites per high powered field), sepsis, respiratory rate (per minute), convulsions > 30 minutes and severe prostration (not able to breastfeed < 6 m, or able to sit > 6 m) also increase the risk of and are associated with development of baseline acute respiratory distress in severe malaria children. For instance, as reported by Khemani et al. (2018), sepsis is the most common cause of acute respiratory distress. Diffuse alveolar damage may arise as a result of endothelial activation, cytokine-mediated inflammatory disorders and reactive oxygen species that are present in individuals with severe sepsis (Truwit et al., 2014).

Though the univariable model, in this study, show that age is statistically significant, the multivariable model suggest that age is not associated with development of baseline acute respiratory distress in severe malaria children. This is in resonance with some literature

reports that indicate that the immune system develops and assumes more complexity with age which suggests reduced risk of developing baseline acute respiratory distress as age increases (Hartel et al., 2005). Thus, younger children are more vulnerable than older people. However, epidemiologic studies to date have not consistently supported distinct paediatric acute respiratory distress outcomes based on age, with the majority of research finding no association between age and acute respiratory distress in children (Flori et al, 2005).

Both univariable and multivariable models indicate no association between sex and baseline acute respiratory distress in severe malaria children. Similar findings are reported by Flori et al. (2005) indicating that there is no difference in the likelihood of worse clinical outcomes, emanating for paediatric acute respiratory distress, between male and female genders. The univariable model, in this study, show that weight (kg) is associated with baseline acute respiratory distress in severe malaria children. That is, increase in weight (or body mass index) results in increased risk of developing baseline acute respiratory distress. While underweight children with acute respiratory distress have increased rates of mortality, obese individuals require longer hospital stays and ICU, but display the lowest risk of in-hospital mortality when compared to other weight categories (Gong et al., 2010). This is what Zhi et al. (2016) called ‘obesity paradox’. On the other hand, the multivariable model indicates that weight (kg) is not associated with any risk of developing baseline acute respiratory distress in severe malaria children.

The univariable model, in this study, show that blood transfusion and mechanical ventilation are significant predictors of baseline acute respiratory distress in severe malaria children. Mechanical ventilation, for instance, causes direct lung injury to a patient affecting the alveoli epithelial and endothelial cells. This may result in development of baseline acute respiratory distress and the worse outcomes associated with it. Transfusion of blood products is a less common endeavour; however, it is a significant cause of acute lung injury and acute respiratory distress. Researchers have verified that transfusions of various blood products, especially those high in protein such as fresh frozen plasma and platelets, are linked to the development of acute respiratory distress in children as well as negative consequences like increased mortality (Church et al., 2009; Khan et al., 2007). On the contrary, however, the multivariable model suggests that there is no association between blood transfusion as well as mechanical ventilation with the development of baseline acute respiratory distress in severe malaria children.

A nomogram was constructed in order to visualise results of a predictive model. It showed that the most important (in descending order) predictors of baseline acute respiratory distress are respiratory rate (per minute), weight, severe acidosis, pneumonia, if a patient is currently treated for chronic illness, hyperparasiteamia, diastolic blood pressure, sepsis, mechanical ventilation, severe prostration, decompensated shock, convulsions, patient age, coma, severe anaemia and blood transfusion. However, cognizant of the fact that other predictors in the predictive model are not statistically significant, it implies that the major predictors of baseline acute respiratory distress are

respiratory rate (per minute), severe acidosis, pneumonia, if a patient is currently treated for chronic illness, hyperparasiteamia, sepsis, severe prostration and convulsions.

The study conducted a Hosmer-Lemeshow goodness-of-fit test to validate the model. The test gave a p-value of 0.9935 which indicted that there is no significant difference between the observed and predicted values of the predictive model, suggesting that the model fitted the data well.

With an optimal probability cutoff estimated to be 0.15, the calculated measures of diagnostic accuracy indicated that the predictive model, in this study, has 68.72% overall rate of correctly classifying patients as having baseline acute respiratory distress or not. The predictive model also show that 69.55% of the patients with baseline acute respiratory distress are correctly classified positive for the disease and 68.57% of the patients without baseline acute respiratory distress are correctly classified negative for the condition. Findings also indicate that 29.62% of the patients are classified as having baseline acute respiratory distress given that the predictive model result is positive and 92.21% of the patients are classified as not having baseline acute respiratory distress given that the predictive model result is negative. The predictive model has an AUC value of 0.75. These results demonstrate that the predictive model developed has a strong discriminative power. That is, it has a good ability to distinguish between severe malaria children with baseline acute respiratory distress and those without baseline acute respiratory distress. The AUC value shows that the prediction is not a random choice.

This study also developed a classification tree as a decision tool to help identify severe malaria patients who are at higher or lower risk of developing baseline acute respiratory distress. This has a potential of helping health practitioners in early identification of baseline acute respiratory distress by considering the conditions presented by a patient. This will also guide proper management and timely interventions provided to such patients in order to minimise worse outcomes associated with baseline acute respiratory distress in children. The classification tree has ranked the presence of pneumonia, severe acidosis: deep breathing, hyperparasitaemia (>500 parasites per high powered field), sepsis as well as increased respiratory rate (per minute) as major conditions classifying a patient of being at high risk of developing baseline acute respiratory distress. These results are consistent with what was reported by Kohne and Flori (2020).

CHAPTER 6

CONCLUSION, RECOMMENDATIONS, LIMITATIONS AND AREAS FOR FURTHER RESEARCH

6.1 Conclusion

Acute respiratory distress (ARD) is a global health concern due to its high rates of morbidity and mortality in children. Severe malaria is one of the most common conditions that accelerates development of ARD in African children. This study aimed at establishing a predictive model for predicting baseline ARD in African children with severe malaria as well as classify the predictors in order of importance of how they influence development of baseline ARD. This retrospective cohort study was a secondary analysis of AQUAMAT data collected from nine African participating countries including; Mozambique, The Gambia, Ghana, Kenya, Tanzania, Nigeria, Uganda, Rwanda, and the Democratic Republic of the Congo.

To determine demographic and clinical predictors of baseline ARD in African children with severe malaria, univariable and multivariable binary logistic regression models were used. These predictors were visualised and ranked using a nomogram. Several approaches were used to validate the predictive model such as Hosmer-Lemeshow goodness-of-fit test, sensitivity, specificity, positive predictive values, negative predictive values and area under the Receiver Operating Characteristic (ROC) curve.

The study revealed that several factors are associated with development of baseline acute respiratory distress in African children with severe malaria. The multivariable binary logistic regression model revealed that the major predictors of baseline ARD were pneumonia, severe acidosis, if a patient is currently treated for chronic illness, hyperparasitaemia, sepsis, respiratory rate, convulsions and severe prostration. The nomogram ranked important (in descending order) predictors of baseline acute respiratory distress as respiratory rate (per minute), severe acidosis, pneumonia, if a patient is currently treated for chronic illness, hyperparasiteamia, sepsis, severe prostration and convulsions.

The Hosmer-Lemeshow goodness-of-fit test indicated that the predictive model fitted well in the data. The predictive model was valuable in predicting baseline ARD with overall correct classification of 68.72%. It also had high discriminative power with area under the ROC curve of 0.75. That is, the predictive model was able to distinguish between a patient with baseline acute respiratory distress and a patient without the condition. Classification tree ranked pneumonia, severe acidosis, hyperparasitaemia, sepsis, increased respiratory rate as well as severe prostration as major conditions classifying a patient of being at high risk of developing baseline ARD. These findings will help medical practitioners in early identification of severe malaria children who are at high risk of developing baseline ARD. This will necessitate improved management and timely interventions provided to such patients in order to prevent development of baseline ARD. These findings will also save medical practitioners' time in identifying and treating children with baseline ARD. The findings will also help WHO and/or Ministries of

Health in different countries to come up with health policies and guidelines that guide diagnosis and management of acute respiratory distress in severe malaria children.

6.2 Recommendations

1. It is important to train medical practitioners on strategies that would help to prevent ARD in children with severe malaria by evaluating the risk of ARD during hospitalization.
2. It is important to formulate guidelines on early identification, interventions, management and treatment of children with severe malaria before developing into ARD.

6.3 Study Limitations

1. The data used in the analysis of this study was collected between 2005 and 2010. Due to numerous interventions, the results may not reflect the current situation on the ground.
2. Being an observational study, it is essential to highlight that the identified predictors may not imply causality, and further prospective studies are needed to establish a causal relationship.
3. The choice of variables included in the analysis may influence the results, and there could be other relevant predictors of baseline ARD that were not included in the study.

6.4 Areas for further research

1. The study proposes the same area of research but focusing on prospective study method to establish causal relationship between predictor variables and baseline ARD.
2. The study proposes the investigation of health system factors such as access to healthcare facilities, availability of resources, and quality of healthcare delivery that may influence the risk of baseline ARD in severe malaria children, thereby informing policy and resource allocation decisions.

REFERENCES

Abreu, M. N. S., Siqueira, A. L., Cardoso, C. S., & Caiaffa, W. T. (2008). Ordinal logistic regression models: Application in quality of life studies. *Cadernos de Saúde Pública*, *24*(4), 581-591. https://doi.org/10.1590/S0102-311X2008001600010

Ashbaugh, D., Boyd Bigelow, D., Petty, T., & Levine, B. (1967). Acute respiratory distress in adults. *The Lancet, 290*(7511), 319-323. https://doi.org/10.1016/s0140-6736(67)90168-7

Azevedo, L. C., Park, M., Salluh, J. I., Rea-Neto, A., Souza-Dantas, V. C., Varaschin, P., Oliveira, M. C., Tierno, P. F. G. M. M., dal-Pizzol, F., Silva, U. V., Knibel, M., Nassar Jr, A. P., Alves, R. A., Ferreira, J. C., Teixeira, C., Rezende, V., Martinez, A., Paula M Luciano, P. M., Schettino, G., & Soares, M. (2013). Clinical outcomes of patients requiring ventilatory support in Brazilian intensive care units: A multicenter, prospective, cohort study. *Critical Care, 17*(2), R63. https://doi.org/10.1186/cc12594

Bellani, G., Laffey, J. G., Pham, T., Fan, E., Brochard, L., Esteban, A., Gattinoni, L., Haren, F. V., Larsson, A., McAuley, D. F., Ranieri, M., Rubenfeld, G., Thompson, B. T., Wrigge, H., Slutsky, A. S., & Pesenti, A. (2016). Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. *JAMA, 315*(8), 788-800. https://doi.org/10.1001/jama.2016.0291

Bernard, G. R., Artigas, A., Brigham, K. L., Carlet, J., Falke, K., Hudson, L., Lamy, M., Legall, J. R., Morris, A., & Spragg, R. (1994). The American-European consensus conference on ARDS. definitions, mechanisms, relevant outcomes, and clinical trial coordination. *American Journal of Respiratory and Critical Care Medicine, 149*(3), 818-824. https://doi.org/10.1164/ajrccm.149.3.7509706

Blei, D. M. (2015). *Linear regression, logistic regression, and generalized linear models*. Columbia University.

Blumberg, L., Lee, R. P., Lipman, J., & Beards, S. (1996). Predictors of mortality in severe malaria: A two-year experience in a non-endemic area. *Anaesthesia and Intensive Care, 24*(2), 217-223. https://doi.org/10.1177/0310057x9602400213

Bos, L. D. J., & Ware, L. B. (2022). Acute respiratory distress syndrome: Causes, pathophysiology, and phenotypes. *The Lancet, 400*(10358), 1145-1156. https://doi.org/10.1016/S0140-6736(22)01485-4

Breman, J. G., Alilio, M. S., & Mills, A. (2004). Conquering the intolerable burden of malaria: what's new, what's needed: a summary. *American Journal of Tropical Medicine and Hygiene, 71*(2), 1-15.

Brun-Buisson, C., Minelli, C., Bertolini, G., Brazzi, L., Pimentel, J., Lewandowski, K., Bion, J., Romand, J. A., Villar, J., Thorsteinsson, A., Damas, P., Armaganidis, A., & Lemaire, F. (2004). Epidemiology and outcome of acute lung injury in European intensive care units. *Intensive Care Medicine, 30*(1), 51–61. https://doi.org/10.1007/s00134-003-2022-6

Bryce, J., Boschi-Pinto, C., Shibuya, K., & Black, R. E. (2005). WHO estimates of the causes of death in children. *The Lancet, 365*(9465), 1147-1152. https://doi.org/10.1016/s0140-6736(05)71877-8

Calfee, C. S., Matthay, M. A., Kangelaris, K. N., Siew, E. D., Janz, D. R., Bernard, G. R., May, A. K., Jacob, P., Havel, C., Benowitz, N. L., & Ware, L. B. (2015). Cigarette smoke exposure and the acute respiratory distress syndrome. *Critical Care Medicine, 43*(9), 1790–1797. https://doi.org/10.1097/ccm.0000000000001089

Caser, E. B., Zandonade, E., Pereira, E., Gama, A. M. C., & Barbas, C. S. V. (2014). Impact of distinct definitions of acute lung injury on its incidence and outcomes in brazilian ICUs. *Critical Care Medicine, 42*(3), 574-582. https://doi.org/10.1097/01.ccm.0000435676.68435.56

Church, G. D., Matthay, M. A., Liu, K., Milet, M., & Flori, H. R. (2009). Blood product transfusions and clinical outcomes in pediatric patients with acute lung injury. *Pediatric Critical Care Medicine, 10*(3), 297-302. https://doi.org/10.1097/pcc.0b013e3181988952

Confalonieri, M., Salton, F., & Fabiano, F. (2017). Acute respiratory distress syndrome. *European Respiratory Review, 26*(144), 160116. https://doi.org/10.1183/16000617.0116-2016

Cortegiani, A., Madotto, F., Gregoretti, C., Bellani, G., Laffey, J. G., Pham, T., Van Haren, F., Giarratano, A., Antonelli, M., Pesenti, A., & Grasselli, G. (2018). Immunocompromised patients with acute respiratory distress syndrome: secondary analysis of the LUNG SAFE database. *Critical Care, 22*(1), 157. https://doi.org/10.1186/s13054-018-2079-9

Czepiel, S. A. (2016). *Maximum likelihood estimation of logistic regression models: Theory and implementation*. https://czep.net/contact.html

Dondorp, A. M., Fanello, C. I., Hendriksen, I. C., Gomes, E., Seni, A., Chhaganlal, K. D., Bojang, K., Olaosebikan, R., Anunobi, N., Maitland, K., Kivaya, E., Agbenyega, T., Nguah, S. B., Evans, J., Gesase, S., Kahabuka, C., Mtove, G., Nadjm, B., Deen, J, Mwanga-Amumpaire, J., Nansumba, M., Karema, C., Umulisa, N., Uwimana, A., Mokuolu, O. A., Adedoyin, O. T., Johnson, W. B. R., Tshefu, A. K., Onyamboko, M. A., Sakulthaew, T., Ngum, W. P., Silamut, K., Stepniewska, K., Woodrow, C. J., Bethell, D., Wills, B.,

Oneko, M., Peto, T. E., von Seidlein, L., Day, N. P. J., & White, N. J. (2010). Artesunate versus quinine in the treatment of severe falciparum malaria in African children (AQUAMAT): An open-label, randomised trial. *The Lancet, 376*(9753), 1647-1657. https://doi.org/10.1016/s0140-6736(10)61924-1

English, M., Sauerwein, R., Waruiru, C., Mosobo, M., Obiero, J., Lowe, B., & Marsh, K. (1997). Acidosis in severe childhood malaria. *QJM, 90*(4), 263–270. https://doi.org/10.1093/qjmed/90.4.263

Erickson, S., Schibler, A., Numa, A., Nuthall, G., Yung, M., Pascoe, E., & Wilkins, B. (2007). Acute lung injury in pediatric intensive care in Australia and New Zealand–A prospective, multicenter, observational study. *Pediatric Critical Care Medicine, 8*(4), 317-23. https://doi.org/10.1097/01.pcc.0000269408.64179.ff

Farghali, R. A., Qasim, M., Kibria, B. M. G., & Abonazel, M. R. (2021). Generalized two- parameter estimators in the multinomial logit regression model: methods, simulation and application. *Communications in Statistics - Simulation and Computation*, *52*(7), 3327-3342. https://doi.org/10.1080/03610918.2021.1934023

Feliciano, H. P., & Mahapatra, S. (2017). *Acute respiratory distress syndrome (ARDS).* Stat Pearls Publishing

Flom, P. L. (2018). An introduction to classification and regression trees with PROC HPSPLIT. *Peter Flom Consulting, LLC,* Paper AA-42. https://www.mwsug.org/proceedings/2018/AA/MWSUG-2018-AA-42.pdf

Flori, H. R., Glidden, D. V., Rutherford, G. W., & Matthay, M. A. (2005). Pediatric acute lung injury: Prospective evaluation of risk factors associated with mortality. *American Journal of Respiratory and Critical Care Medicine, 171*(9), 995-1001. https://doi.org/10.1164/rccm.200404-544oc

Golub, G. H., & Van Loan, C. F. (1996). *Matrix Computations* (3rd ed.). Johns Hopkins.

Gong, M. N., Bajwa, E. K., Thompson, B. T., & Christiani, D. C. (2009). Body mass index is associated with the development of acute respiratory distress syndrome. *Thorax, 65*(1), 44-50. https://doi.org/10.1136/thx.2009.117572

Gong, M. N., Thompson, B. T., Williams, P., Pothier, L., Boyce, P. D., & Christiani, D. C. (2005). Clinical predictors of and mortality in acute respiratory distress syndrome: Potential role of red cell transfusion. *Critical Care Medicine, 33*(6), 1191-1198. https://doi.org/10.1097/01.ccm.0000165566.82925.14

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36. https://doi.org/10.1148/radiology.143.1.7063747

Hardin, J. W., & Hilbe, J. M. (2018). *Generalised linear models and extensions* (4th ed.). StataCorp LLC.

Hartel, C., Adam, N., Strunk, T., Temming, P., Muller-Steinhardt, M., & Schultz, C. (2005). Cytokine responses correlate differentially with age in infancy and early childhood. *Clinical and Experimental Immunology, 142*(3), 446-453. https://doi.org/10.1111/j.1365-2249.2005.02928.x

Heidemann, S. M., Nair, A., Bulut, Y., & Sapru, A. (2017). Pathophysiology and management of acute respiratory distress syndrome in children. *Pediatric Clinics of North America, 64*(5), 1017-1037. https://doi.org/10.1016/j.pcl.2017.06.004

Helegbe, G. K., Goka, B. Q., Kurtzhals, J. A., Addae, M. M., Ollaga, E., Tetteh, J. K., Dodoo, D., Ofori, M. F., Obeng-Adjei, G., Hirayama, K., Awandare, G. A., & Akanmori, B. D. (2007). Complement activation in Ghanaian children with severe Plasmodium falciparum malaria. *Malaria Journal, 6*(1). https://doi.org/10.1186/1475-2875-6-165

Hviid, L., & Jensen, A. T. R. (2015). PfEMP1 - *A* parasite protein family of key importance in plasmodium falciparum malaria immunity and pathogenesis. *Advances in Parasitology, 88(1),* 51-84. https://doi.org/10.1016/bs.apar.2015.02.004

Jeena, P. (2008). An approach to the child in respiratory distress. *South African Family Practice, 50*(3), 32-37. https://doi.org/10.1080/20786204.2008.10873713

Kallet, R. H., Zhuo, H., Ho, K., Lipnick, M. S., Gomez, A., & Matthay, M. A. (2017). Lung injury etiology and other factors influencing the relationship between dead-space fraction and mortality in ARDS. *Respiratory Care, 62*(10), 1241-1248. https://doi.org/10.4187/respcare.05589

Khan, H., Belsher, J., Yilmaz, M., Afessa, B., Winters, J. L., Moore, S. B., Hubmayr, R. D., Gajic, O. (2007). Fresh-frozen plasma and platelet transfusions are associated with development of acute lung injury in critically ill medical patients. *Chest, 131*(5), 1308- 1314. https://doi.org/10.1378/chest.06-3048

Khemani, R. G., Smith, L., Lopez-Fernandez, Y. M., Kwok, J., Morzov, R., Klein, M. J., Yehya, N., Willson, D., Kneyber, M. C. J., Lillie, J., Fernandez, A., Newth, C. J. L., Jouvet, P., & Thomas, N. J. (2018). Paediatric acute respiratory distress syndrome incidence and epidemiology (PARDIE): An international, observational study. *The Lancet Respiratory Medicine, 7*(2), 115-128. https://doi.org/10.1016/s2213-2600(18)30344-8

Kohne, J. G., & Flori, H. R. (2020). *Risk factors and etiologies of pediatric acute respiratory distress syndrome.* Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-21840-9_4

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). *Applied linear regression models* (4th ed.). McGraw-Hill/Irwin.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (4th ed.). McGraw-Hill/Irwin.

Lin, F., Zhou, Q., Li, W., Xiao, W., Li, S., Liu, B., Li, H., Cui, Y., Lu, R., Li, Y., Zhang, Y., & Pan, P. (2023). A prediction model for acute respiratory distress syndrome in immunocompetent adults with adenovirus-associated pneumonia: A multicenter retrospective analysis. *BMC Pulmonary Medicine, 23*(1).https://doi.org/10.1186/s12890-023-02742-8

Lin, H., Tao, J., Kan, H., Qian, Z., Chen, A., Du, Y., Liu, T., Zhang, Y., Qi, Y., Ye, J., Li, S., Li, W., Xiao, J., Zeng, W., Li, X., Stamatakis, K. A., Chen, X., & Ma, W. (2018). Ambient particulate matter air pollution associated with acute respiratory distress syndrome in Guangzhou, China. *Journal of Exposure Science & Environmental Epidemiology*, *28*(4), 392-399. https://doi.org/10.1038/s41370-018-0034-0

Lindsey, J. K. (1997). *Applying generalised linear models*. Springer-Verlag Inc.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(1), 14–23. https://doi.org/10.1002/widm.8

Lv, L., Shao, X., & Cui, E. (2024). Establishment of a predictive model for acute respiratory distress syndrome in patients with bacterial pneumonia. *Journal of Inflammation Research, 17*, 2825-2834. https://doi.org/10.2147/JIR.S458690

Marsh, K., Forster, D., Waruiru, C., Mwangi, I., Winstanley, M., Marsh, V., Newton, C., Winstanley, P., Warn, P., Peshu, N., Pasvol, G., & Snow, R. (1995). Indicators of

life-threatening malaria in African children. *New England Journal of Medicine, 332*(21), 1399–1404. https://doi.org/10.1056/nejm199505253322102

Marsh, K., & Kinyanjui, S. (2006). Immune effector mechanisms in malaria. *Parasite Immunology*, *28*(1-2), 51-60. https://doi.org/10.1111/j.1365-3024.2006.00808.x

McCullagh, P., & Nelder FRS, J. A. (1989). *Generalised linear models* (2nd ed.). Chapman and Hall.

Meyer, N. J., Gattinoni, L., & Calfee, C. S. (2021). Acute respiratory distress syndrome. *The Lancet, 398*(10300), 622-637. https://doi.org/10.1016/S0140-6736(21)00439-6

Mitran, C., Opoka, R. O., Conroy, A. L., Namasopo, S., Kain, K. C., & Hawkes, M. T. (2023). Pediatric malaria with respiratory distress: Prognostic significance of point-of-care lactate. *Microorganisms, 11*(4), 1-20. https://doi.org/10.3390/microorganisms11040923

Moazed, F., Hendrickson, C., Jauregui, A., Gotts, J., Conroy, A., Delucchi, K., Zhuo, H., Arambulo, M., Vessel, K., Ke, S., Deiss, T., Ni, A., Leligdowicz, A., Abbott, J., Cohen, M. J., Sinha, P., Gomez, A., Kangelaris, K., Kornblith, L., Matthay, M., Benowitz, N., Liu, K., & Calfee, C. S. (2022). Cigarette smoke exposure and acute respiratory distress syndrome in sepsis: Epidemiology, clinical features, and biologic markers. *American Journal of Respiratory and Critical Care Medicine, 205*(8), 927-935. https://doi.org/10.1164/rccm.202105-1098OC

Mora, R. (2019, October 20). *Crtrees: An implementation of classification and regression trees (CART) and random forests in stata* [PowerPoint slides]. https://www.stata.com/meeting/spain19/slides/Spain19_Mora.pdf

Mzumara, G., Leopold, S., Marsh, K., Dondorp, A., Ohuma, E. O., & Mukaka, M. (2021). Identifying prognostic factors of severe metabolic acidosis and uraemia in African children with severe falciparum malaria: a secondary analysis of a randomized trial. *Malaria Journal, 20*(1), 282-9. https://doi.org/10.1186/s12936-021-03785-0

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General), 135*(3), 370-384. https://doi.org/10.2307/2344614

Oduro, A. R., Koram, K. A., Rogers, W., Atuguba, F., Ansah, P., Anyorigiya, T., Ansah, A., Anto, F., Mensah, N., Hodgson, A., & Nkrumah, F. (2007). Severe falciparum malaria in young children of the Kassena-Nankana district of northern Ghana. *Malaria Journal, 6*(1). https://doi.org/10.1186/1475-2875-6-96

Pediatric Acute Lung Injury Consensus Conference Group. (2015). Pediatric acute respiratory distress syndrome: Consensus recommendations from the pediatric acute lung injury consensus conference. *Pediatric Critical Care Medicine, 16*(5), 428-439. https://doi.org/10.1097/pcc.0000000000000350

Quasney, M. W., López-Fernández, Y. M., Santschi, M., & Watson, R. S. (2015). The outcomes of children with pediatric acute respiratory distress syndrome. *Pediatric Critical Care Medicine, 16*(5),118-131. https://doi.org/10.1097/pcc.0000000000000438

Ranieri, V. M., Rubenfeld, G. D., Thompson, B. T., Ferguson, N. D., Caldwell, E., Fan, E., Camporota, L., & Slutsky, A. S. (2012). Acute respiratory distress syndrome: The Berlin definition. *JAMA, 307*(23), 2526-2533. https://doi.org/10.1001/jama.2012.5669

Reilly, J. P., Christie, J. D., & Meyer, N. J. (2017). Fifty years of research in ARDS. Genomic contributions and opportunities. *American Journal of Respiratory*

and Critical Care Medicine, 196(9), 1113-1121. https://doi.org/10.1164/rccm.201702-0405cp

Reilly, J. P., Zhao, Z., Shashaty, M. G. S., Koyama, T., Christie, J. D., Lanken, P. N., Wang, C., Balmes, J., Matthay, M., Calfee, C. S., & Ware, L. B. (2018). Low to moderate air pollutant exposure and acute respiratory distress syndrome after severe trauma. *American Journal of Respiratory and Critical Care Medicine. 199*(1), 62-70 https://doi.org/10.1164/rccm.201803-0435oc

Rubenfeld, G. D., Caldwell, E., Peabody, E., Weaver, J., Martin, D. P., Neff, M., Stern, E. J., & Hudson, L. D. (2005). Incidence and outcomes of acute lung injury. *New England Journal of Medicine, 353*(16), 1685–1693. https://doi.org/10.1056/nejmoa050333

Schouten, L. R. A., Veltkamp, F., Bos, A. P., van Woensel, J. B. M., Serpa Neto, A., Schultz, M. J., & Wösten-van Asperen, R. M. (2015). Incidence and mortality of acute respiratory distress syndrome in children. *Critical Care Medicine, 44*(4), 819-29. https://doi.org/10.1097/ccm.0000000000001388

Simou, E., Leonardi-Bee, J., & Britton, J. (2018). The effect of alcohol consumption on the risk of ARDS. *Chest, 154*(1), 58–68. https://doi.org/10.1016/j.chest.2017.11.041

Šimundić, A. M. (2009). Measures of diagnostic accuracy: basic definitions. *Electronic Journal of the International Federation of Clinical Chemistry and Laboratory Medicine,19*(4), 203-211

Shah, S. S., Fidock, D. A., & Prince, A. S. (2021). Hemozoin promotes lung inflammation via host epithelial activation. *Microbe Biology, 12*(1). https://doi.org/10.1128/mBio.02399-20

Shakhawat, H. S., Ejaz, A., & Hatem, A. H. (2014). Model selection and parameter estimation of a multinomial logistic regression model. *Journal of Statistical Computation and Simulation*, *84*(7), 1412-1426. https://doi.org/10.1080/00949655.2012.746347

Shalizi, C. R. (2015, December 4). *Classification and Regression Trees*. https://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/27/lecture-27.pdf

Sweeney, R. M., & McAuley, D. F. (2016). Acute respiratory distress syndrome. *The Lancet*, *388*(10058), 2416–2430. http://dx.doi.org/10.1016/S0140-6736(16)00578-X

Swift, A., Heale, R., & Twycross, A. (2019). What are sensitivity and specificity? *Evidence Based Nursing*, *23*(1), 2-4. https://doi.org/10.1136/ebnurs-2019-103225

Thomas, N. J., Jouvet, P., & Willson, D. (2013). Acute lung injury in children—kids really aren't just "little adults." *Pediatric Critical Care Medicine*, *14*(4), 429-432. https://doi.org/10.1097/pcc.0b013e31827456aa

Toy, P., Looney, M. R., Popovsky, M., Palfi, M., Berlin, G., Chapman, C. E., Bolton-Maggs, P., & Matthay, M. A. (2022). Transfusion-related acute lung injury: 36 years of progress (1985-2021). *Annals of the American Thoracic Society*, *19*(5), 705-712. https://doi.org/10.1513/annalsats.202108-963cme

Truwit, J. D., Bernard, G. R., Steingrub, J., Matthay, M. A., Liu, K. D., Albertson, T. E., Brower, R. G., Shanholtz, C., Rock, P., Douglas, I. S., deBoisblanc, B. P., Hough, C. L., Hite, R. D., & Thompson, B. T. (2014). Rosuvastatin for Sepsis-Associated Acute Respiratory Distress Syndrome. *New England Journal of Medicine*, *370*(23), 2191-2200. https://doi.org/10.1056/nejmoa1401520

Ware, L. B., Zhao, Z., Koyama, T., May, A. K., Matthay, M. A., Lurmann, F. W., Balmes, J. R., & Calfee, C. S. (2016). Long-term ozone exposure increases the risk of developing the acute respiratory distress syndrome. *American Journal of Respiratory and Critical Care Medicine, 193*(10), 1143-1150. https://doi.org/10.1164/rccm.201507-1418oc

Watkins, T. R., Nathens, A. B., Cooke, C. R., Psaty, B. M., Maier, R. V., Cuschieri, J., & Rubenfeld, G. D. (2012). Acute respiratory distress syndrome after trauma. *Critical Care Medicine, 40*(8), 2295-2303. https://doi.org/10.1097/ccm.0b013e3182544f6a

Williams, B. G., Gouws, E., Boschi-Pinto, C., Bryce, J., & Dye, C. (2002). Estimates of world- wide distribution of child deaths from acute respiratory infections. *The Lancet Infectious Diseases, 2*(1), 25-32. https://doi.org/10.1016/s1473-3099(01)00170-0

World Health Organisation. (1990). Severe and complicated malaria. *Transactions of the Royal Society of Tropical Medicine and Hygiene,* 84(2), 1-65.

World Health Organisation. (2021, August 31). *Pneumonia.* Author

World Health Organisation, (2022). *World malaria report*. Geneva, Switzerland. https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2022

Xu, C., Zheng, L., Jiang, Y., & Jin, L. (2023). A prediction model for predicting the risk of acute respiratory distress syndrome in sepsis patients: A retrospective cohort study. *BMC Pulmonary Medicine, 23*(1), 78. https://doi.org/10.1186/s12890-023-02365-z

Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagnosis, with special reference to x-ray techniques. *Public Health Reports (1896-1970), 62*(40), 1432-1449. https://doi.org/10.2307/4586294

Zhi, G., Xin, W., Ying, W., Guohong, X., & Shuying, L. (2016). "Obesity paradox" in acute respiratory distress syndrome: Asystematic review and meta-analysis. *PLOS ONE, 11*(9), e0163677. https://doi.org/10.1371/journal.pone.0163677

APPENDICES

APPENDIX A: STATA Codes

```
***DO FILE FOR MSC BIOSTATISTICS THESIS-INNOCENT GONDWE***
use "H:\INNOCENT GONDWE\SEMESTER 3-4_Research\DATA\AQUAMAT_child.dta"

**viewing dataset**
browse

**standardising variable sex to be lower cases only**
replace sex = lower(sex)
tabulate sex

**encoding, recoding and labeling variable sex**
encode sex, generate(sex_numeric)
recode sex_numeric (1=0) (2=1), generate(sex_numeric_recoded)
rename sex_numeric_recoded sex1
label define gender 0 "female" 1 "male"
label values sex1 gender

**overview of the variables of interest in the data**
tabulate arespins
tabulate sex1
tabulate patage
tabulate aweight
tabulate aresp
tabulate abloodprsyst
tabulate abloodprdiast
tabulate bcmaln
tabulate btransf
tabulate bvent
tabulate aanaemia
```

```
tabulate ashockcomp
tabulate ashockdecomp
tabulate ahyperp
tabulate acoma
tabulate aconvulsions
tabulate aacidosis
tabulate aprost
tabulate acurchron
tabulate bcgast
tabulate bpneum
tabulate bseps
tabulate odead
tabulate odead2
tabulate brenal

**cleaning data by replacing missing values**
egen mean_aweight =mean(aweight)
replace aweight = mean_aweight if missing(aweight)
tabulate aweight

egen mean_aresp =mean(aresp)
replace aresp = mean_aresp if missing(aresp)
tabulate aresp

egen mean_abloodprsyst=mean(abloodprsyst)
replace abloodprsyst= mean_abloodprsyst if missing(abloodprsyst)
tabulate abloodprsyst

egen mean_abloodprdiast =mean(abloodprdiast)
replace abloodprdiast = mean_abloodprdiast if missing(abloodprdiast)
tabulate abloodprdiast

recode acurchron (2=0), generate(acurchron_New)
tabulate acurchron_New
```

renaming variable labels in the data

label var odrug "Which study-drug did the patient receive (value)"

label var odrug2 "Which study-drug did the patient receive (labels)"

label var arespins "Respiratory distress: Costal indrawing/recession, respiratory insufficiency"

label var country "Country"

label var sex "Sex"

label var sex1 "Sex1"

label var patage "Patient age in years"

label var aweight "Weight (kg)"

label var aresp "Respiratory rate (/min)"

label var abloodprsyst "Blood pressure - Systolic (mmHg)"

label var abloodprdiast "Blood pressure - Diastolic (mmHg)"

label var bcmaln "Severe malnutrition"

label var brenal "Renal failure (urine output <0.5 ml/kg/hour, for >24 hours)"

label var btransf "Blood transfusion"

label var bvent "Machanical ventilation"

label var acurchron_New "Currently treated for chronic illness (value)"

label var acurchron2 "Currently treated for chronic illness (label)"

label var bcgast "Gastro enteritis"

label var bpneum "Pneumonia (Y/N)"

label var bseps "Sepsis (Y/N)"

label var odead "Patient died/patient survived (value)"

label var odead2 "Patient died/patient survived (label)"

label var aanaemia "Symptomatic severe anaemia (severe pallor combined with respiratory distress)"

label var acoma "Coma at admission (GCS <= 10, BCS <= 2)"

label var ashockcomp "Compensated shock (Only for children: capillary refil >= 3 sec/temperature gradient)"

label var ahyperp "Hyperparasitaemia (>500 parasites per high powered field)"

label var aacidosis "Suspected severe acidosis: Deep breating"

label var ashockdecomp "Decompensated shock (Adults: systolic BP < 80 mmHg, Children: systolic BP < 70 mmHg)"

label var aconvulsions "Convulsions > 30 minutes"

label var aprost "Severe prostration (Not able to breastfeed < 6m, or able to sit > 6m)"

creating table of baseline characteristics (table 1)

```
ssc install table1_mc

table1_mc, by(odrug2) ///
vars( ///
country cat %4.0f \ ///
 patage conts %4.0f \ ///
 aweight contn %4.0f \ ///
 sex1 cat %4.0f \ ///
 aresp contn %4.0f \ ///
 abloodprsyst contn %4.0f \ ///
 abloodprdiast contn %4.0f \ ///
 bpneum bin %4.0f \ ///
 bseps bin %4.0f \ ///
 aanaemia bin %4.0f \ ///
 arespins bin %4.0f \ ///
 acoma bin %4.0f \ ///
 aconvulsions bin %4.0f \ ///
 ahyperp bin %4.0f \ ///
 ashockcomp bin %4.0f \ ///
 ashockdecomp bin %4.0f \ ///
 aacidosis bin %4.0f \ ///
 btransf bin %4.0f \ ///
 bvent bin %4.0f \ ///
 acurchron_New bin %4.0f \ ///
 brenal bin %4.0f \ ///
 aprost bin %4.0f \ ///
) ///
nospace onecol missing total(before) ///
saving("table 1.xlsx", replace)

**descriptive statistics**
**exploring data**
tabstat patage, statistics(mean sd max min)
tabstat patage, statistics(mean sd max min) by(sex1)
```

labeling values for respiratory distress

label define arespins_1 0 "No" 1 "Yes"

label values arespins arespins_1

tabulate arespins

Bar chart of respiratory distress by survival

graph bar arespins, over(odead2) ytitle(Proportion of patients with respiratory distress) ylabel(, angle(horizontal)) graphregion(fcolor(white))

Histogram for patient age (years)

histogram patage, bin(15) ytitle(Proportion) xtitle(Patient age (years)) ylabel(, angle(horizontal)) graphregion(fcolor(white))

Histogram for weight (kg)

histogram aweight, bin(20) ytitle(Proportion) xtitle(Weight (kg)) ylabel(, angle(horizontal)) graphregion(fcolor(white))

Histogram for respiratory rate (per minute)

histogram aresp, bin(20) ytitle(Proportion) xtitle(Respiratory rate (per minute)) ylabel(, angle(horizontal)) graphregion(fcolor(white))

Histogram for systolic blood pressure (mmHg)

histogram abloodprsyst, bin(20) ytitle(Proportion) xtitle(Systolic blood pressure (mmHg)) ylabel(, angle(horizontal)) graphregion(fcolor(white))

Histogram for diastolic blood pressure (mmHg)

histogram abloodprdiast, bin(20) ytitle(Proportion) xtitle(Diastolic blood pressure (mmHg)) ylabel(, angle(horizontal)) graphregion(fcolor(white))

Bar chart of respiratory distress by country

graph bar arespins, over(sex1) ytitle(Proportion of patients with respiratory distress) ylabel(, angle(horizontal)) graphregion(fcolor(white))

scatter plot of weight (kg) against patient age (years)

scatter aweight patage, xtitle(Patient age (years)) ytitle(Weight (kg)) ylabel(, angle(horizontal)) graphregion(fcolor(white))

univariable logistic regression model********
logit arespins patage, or
logit arespins i.sex1, or
logit arespins aweight, or
logit arespins aresp, or
logit arespins abloodprsyst, or
logit arespins abloodprdiast, or
logit arespins bpneum, or
logit arespins bseps, or
logit arespins aanaemia, or
logit arespins acoma, or
logit arespins ahyperp, or
logit arespins aconvulsions, or
logit arespins ashockcomp, or
logit arespins ashockdecomp, or
logit arespins aacidosis, or
logit arespins btransf, or
logit arespins bvent, or
logit arespins acurchron_New, or
logit arespins brenal, or
logit arespins aprost, or

multivariable logistic regression model********
logit arespins patage aweight aresp abloodprdiast bpneum bseps aanaemia acoma ahyperp aconvulsions ashockdecomp aacidosis btransf bvent acurchron_New aprost, or

generating optimal cutoff point by plotting sensitivity and specificity using lsens
lsens, genprob(cutoff) recast(line) ylabel(, angle(horizontal)) graphregion(fcolor(white)) xline(0.15, lpattern(dash)) xlab(0 0.15 0.25(0.25)1)

calculating sensitity, specificity, positive predictive value and negative predictive value
estat classification, cutoff(0.15)

predicted probabilities/values for respiratory distress

predict arespins_prdct

plot of ROC

roctab arespins arespins_prdct, graph recast(line) rlopts(lcolor(black) lwidth(vthin)) ylabel(, angle(horizontal)) graphregion(fcolor(white))

Area under the ROC curve

roctab arespins arespins_prdct, detail

// development of prognostic scoring system**

** installation of nomolog for this old version stata-journal**

net from http://www.stata-journal.com/software

net cd sj15-2

net describe st0391

net install st0391

window menu append item "stUserGraphics""&Npmogram post logistic regression""dbnomolog"

window menu refresh

nomogram code

logit arespins patage aweight aresp abloodprdiast bpneum bseps aanaemia acoma ahyperp aconvulsions ashockdecomp aacidosis btransf bvent acurchron_New aprost, or

nomolog

Hosmer-Lemeshow test

estat gof

APPENDIX B: R Codes

```
#CLASSIFICATION TREE R CODE – INNOCENT GONDWE
#Loading libraries
library(haven) #for importing and reading stata (.dta) dataset
library(dplyr) #for recoding variables
library(rpart) #for fitting decision trees
library(rpart.plot) #for plotting decision trees

#reading dataset from stata
AQUAMAT_child <- read_dta("H:/INNOCENT GONDWE/SEMESTER 3-4_Research/DATA/AQUAMAT_child.dta")
View(AQUAMAT_child)

#cleaning data by replacing missing values
AQUAMAT_child$aresp[is.na(AQUAMAT_child$aresp)]<-mean(AQUAMAT_child$aresp, na.rm=TRUE)

#recoding variable
AQUAMAT_child %>% mutate(acurchron=recode(acurchron, "2=0; 0=0; 1=1"))

#setting seed
set.seed(123)

#building the initial tree
tree <- rpart(arespins ~ bpneum+aacidosis+acurchron+ahyperp+bseps+aresp+aconvulsions+aprost,
data=AQUAMAT_child, control=rpart.control(cp=.0001))

#viewing results of the initial tree
printcp(tree)

#identifying best cp value to use
best <- tree$cptable[which.min(tree$cptable[,"xerror"]),"CP"]
```

```
#producing a pruned tree based on the best cp value
pruned_tree <- prune(tree, cp=best)

#plotting the pruned tree
prp(pruned_tree,
    faclen=0, #use full names for factor labels
    extra=1, #display the number of observations that fall in the node
    branch=1, #produce square shouldered branch lines
    yesno=2, #write 'yes' and 'no' at all splits
    roundint=T, #round values to integers at splitting nodes
    digits=3) #display 3 decimal places in terminal nodes
```