MODELLING COMMUNITY EFFECTS ON ANEMIA IN SCHOOL-AGED CHILDREN IN MALAWI USING A MULTILEVEL REGRESSION MODEL

MASTERS OF SCIENCE (BIOSTATISTICS) THESIS

GLORY GONDWE MSHALI

UNIVERSITY OF MALAWI

AUGUST, 2023

MODELLING COMMUNITY EFFECTS ON ANEMIA IN SCHOOL-AGED CHILDREN IN MALAWI USING A MULTILEVEL REGRESSION MODEL

MSc. (BIOSTATISTICS) THESIS

By

GLORY GONDWE MSHALI

Bachelor of Science-University of Malawi

Thesis submitted to the Department of Mathematical Sciences, Faculty of Science, in partial fulfilment of the requirement for degree of Master of Science (Biostatistics)

UNIVERSITY OF MALAWI

AUGUST, 2023

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.

Glory Gondwe Mshali

Full Legal Name

Signature

01 September, 2023

Date

CERTIFICATE OF APPROVAL

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

Signature: _____ **Date:** 01 September 2023

Jupiter Simbeye, PhD (Senior Lecturer)

Supervisor

DEDICATION

First and foremost, I am grateful to my Almighty God. Your eyes see where human beings cannot see and I thank you for the assurance that nothing is impossible with you. Thanks for giving me wisdom to know what to do.

To mum and dad, I say thank your for raising me up. I really appreciate your love and care. To my husband Rodney Mshali, my children Florence, Rodney Junior, and Ethel Kwangu and all my sisters, many thanks for your inspiration, motivation and encouragement.

ACKNOWLEDGEMENTS

Dr J. Simbeye, your supervision, guidance and continuous support will remain to be remembered. You kept encouraging me from the beginning up to the end.
Finally, I thank the National Statistical Office for part payment of my fees in order to study Master of Science in Biostatistics at the University of Malawi, Chancellor College, Zomba. I would like to thank all my lecturers, classmates and friends for your love and unity, willingness, commitment and encouragement rendered to me during the entire period of my study in Biostatistics.

ABSTRACT

Anemia is a serious health problem in Malawi that usually results from poor nutrition, infection, or chronic diseases. Anemia in school-aged children 5-14 years has been associated with poor cognitive performance, impaired immunity and decrease working capacity. Therefore the present study focused in modeling community effects on anemia in school-aged children 5-14 years old in Malawi using multilevel logistic regression analysis. The study used Cross-sectional data from the Malawi Demographic Health Survey (MDHS) 2015-16 and the Malawi Micronutrient Survey(MNS) (2015-16). The statistical models that suited the hierarchical data such as variance components model, random intercept model and random coefficients model were used in the analysis. The Log Likelihood approach was used to estimate the fixed effect and random effects in the multilevel analysis. The results of the descriptive statistics showed that effect on anemia in school-aged children 5-14 years was 19%. Performing the logistic regression analysis showed that: place of residence, age of child, education of child, source of drinking water, inflammation, wealth index and head of household sex had a significant effect on anemia in school-aged children. Multilevel logistic regression model better suited the hierarchical clustered data with higher values of log likelihood estimates of -348.65 verses -355.63 for logistic regression. The random intercept model, AIC of 730.53 and random coefficient model of 733.50 did not differ much in variations and were both treated as the better fit model as compared to variance component model with AIC of 769.69. Therefore, anemia in school-aged children 5-14 years still remains a challenge in Malawi and that Multilevel modeling identified considerable community variations in its distribution which requires stakeholders, policy makers and the public health to pay attention to these significant effects on anemia

TABLE OF CONTENTS

LIST OF TABLES

ABBREVIATIONS

CDC	Centre for Disease Control and Prevention
CI	Confidence Interval
CPR	Crude Prevalence Ratio
MDGs	Millennium Development Goals
MDHS	Malawi Demographic Health Survey
MLMs	Multilevel Linear Models
MNS	Malawi Micro nutrient Survey
MRDR	Modified Relative Dose Response
NSO	National Statistical Office
OR	Odds Ratio
PCD	Partnership for Child Development
PR	Prevalence Ratio
PSC	Preschool children
RR	Relative Risks
SAC	School Aged Children
SDGs	Sustainable Development Goal
SHSU	Community Health Sciences Unit
WHO	World Health Organization
WRA	Women of Reproductive Age

CHAPTER 1

INTRODUCTION

1.1. Background

Anemia is characterized by low levels of hemoglobin and protein in red blood cells responsible for carrying oxygen in the blood. Iron deficiency is estimated to contribute to approximately one-half of anemia cases worldwide (Ngnie-Teta, Receveur, & Kuate-Defo, 2007). Other micronutrient deficiencies includes vitamin A, folate, vitamin B12 and non-nutritional causes such as blood disorders, malaria, schistosomiasis, and heminthic infections are also causes of anemia. Anemia impairs children's physical and cognitive development, increases susceptibility to infections citepSyed2016 and (WHO., 2008). Anemia also increases risk of child and maternal mortality (Calis et al., 2016).

Anemia is a global public health problem which affects both the developing and the developed countries and globally, anemia affects 1.62 billion people, which corresponds to 24.8% of the population. Anemia remains difficult to manage in malaria endemic countries of Africa (WHO., 2008). Young children and pregnant women were the most affected with an estimated global prevalence of 43% and 51% respectively. According to The Malawi Demographic Health Survey (MDHS) 2015-16 and Malawi Micro-Nutrient Survey (MNS) 2015-16 anemia prevalence among school-aged children 5 to 14 years old was found to be 22 percent and 30 percent in pre-school children. In Malawi more than three in every five children 0 to 4 years were found to be anemic, and two in every five children aged 5 to 14 years old were found to be anemic (National Statistical Office, 2015).

Ngnie-Teta et al. (2007) described that there were many causes of anemia on both an individual and environmental factors that were related to both the impaired delivery

of oxygen and the compensatory mechanism that were developed in response to the situation. Anemia could also be caused due to lack of nutritional requirements for the body, the unavailability of iron, vitamin B12, and folic acid, inadequate iron supply, due to either true iron deficiency or inflammation-induced sequestration of iron in macrophages(Bates & Sarkinfada, 2007).

Many studies comprehensively examined the risk factors of anemia using descriptive statistics and simple logistic regression in Malawi and across countries in under-five and school-aged children. These studies concentrated on significance of risk factors of anemia at individual level (Syed et al., 2016). However, (Ntenda Morton. & chih Chuang, 2017) did a study on multilevel analysis of effects of individual and community (cluster) level factors on childhood anemia, severe anemia and hemoglobin concentration in Malawi in pre-school children aged between 6 to 59 months using the MDHS data of 2010. Generalized linear mixed models were constructed for anemia, and linear mixed-effect models were used in hemoglobin (Hb) concentration and binary modelling was used to measure factors associated with anemia. The study did not investigate the within and between community or district effects on anemia in school-aged children, 5 to 14 years old.

Kazembe and Ngwira (2015) aimed on fitting cumulative logistic threshold model on young children of 6-59 months permitting nonlinear effects for continuous variables and spatial effects on district of residence to investigate the risk factors affecting childhood anemia. Inference were based on empirical Bayes framework and spatial variations using ordered Bayesian and cumulative logistic regression was used to measure severity of anemia. Their findings revealed substantial spatial variation with increased risk of anemia observed in some of the districts.

Information on trends in the effects on anemia within and between communities among school-aged children in Malawi is lacking. The Malawi Micronutrient Survey (MNS) 2015-16 showed that 5 percent of the school-aged children (SAC) 5 to 14 years old had iron deficiency and 22% were anemic. The prevalence of anemia was higher in SAC aged 5 to 10 years with 20%, compared to those aged 11 to 14 years with 10% ($p < 0.05$) (National Statistical Office, 2015).

Given the difficulty in implementing effective measures for controlling anemia, it is important to investigate the scope and strength of individual and community risk factors on anemia in populations where anemia is common to design more effective intervention planning at district, region and national level.

Usage of multilevel regression in analyzing hierarchical data from MDHS and MNS 2015-16 (data collected in cluster form) that tends to investigate individual-level and community-level effects on anemia in school-aged children is rare in literature. In this thesis, a multilevel regression model was developed to determine the individual-level and community-level effects on anemia in school-aged children 5 to 14 years old in Malawi. The hierarchical approach was used by considering the nature of data.

1.2. **Problem statement**

Few studies have considered multilevel models that explicitly consider the nesting of data and checking for variations to determine the strength of the effects in factors associated with anemia in school-aged children 5 to 14 years old. Application of multilevel logistic regression model with variance component will help to investigate the individual and community effects on anemia in SAC in Malawi.

Overall, anemia is associated with greater morbidity and mortality. In Malawi, previous studies have linked anemia to factors such as sex of the child, if mother is anemic, ever breastfed, age, wealth index, if parents are alive, education of parents, nutritional status of child and non-nutritional causes such as blood disorders, malaria and schistosomiasis (Calis et al., 2016).

Most of anemia studies conducted in Malawi used single-level analysis technique with population groups localized in a specific study area (Ngwira & Kazembe, 2016). Likewise, for a Micronutrient 2015-16 study conducted in Malawi on determinants of anemia in school-aged children 5-14 years old, also used single-level technique in their analysis (National Statistical Office, 2015). From a study by (Barth et al., 2018), the single level analysis assumed that there is no community-level effects beyond the characteristics of individuals. This showed that the impact of community-level effects on

anemia among preschool children, school-aged children 5-14 years old remained understudied in Malawi. Moreover, analyzing hierarchical data like the MDHS and MNS under single-level analysis leads to incorrect estimation of parameters and standard errors (Gebremeskel et al., 2020).

By using techniques in multilevel analysis, the community-level effects can be identified from individual-level effects (Ntenda Morton. & chih Chuang, 2017). This approach has never been used in Malawi on effects on anemia in school-aged children 5-14 years old to identify community and district effects on anemia. In this study, multilevel logistic regression model was used to investigate correlations on anemia in school-aged children at individual and community-level.

1.3. **Main Objective**

The objective of this study was to model community effects on anemia in school-aged children 5-14 years in Malawi using data from the MDHS and MNS 2015-16.

1.3.1 Specific Objectives

- Develop a multilevel model for measuring community effects on anemia in school-aged children 5-14 years in Malawi
- To investigate the correlates of anemia in school-aged children 5-14 years in Malawi.
- Investigate the impacts of community on the estimation on anemia in school-aged children.

1.4. **Significance of the study**

There was need to investigate the community effect on anemia in school aged children in Malawi. This was important because prone areas with higher variability could be targeted for research and monitoring. The model developed in the study was used to analyze the effect on anemia in school-aged children basing on the nested sources of variability and the hierarchical structure of data. Units of level-1 (for instance individual school-aged children of the households) were nested in units of level-2 (com-

munity). This study will help key partners like; Stakeholders, Government, NGOs, Development partners and others) to reduce anemia in school-aged children by intensifying interventions basing on correlatable effects which were significant like, mild and severe inflammation and children coming from poorest families.

CHAPTER 2

LITERATURE REVIEW

2.1. Introduction

Logistic regression models have been widely used in modelling child anemia for 0 to 4 years in Malawi. However, little is known on whether communities or districts where children live have an impact on deepening our understanding on anemia in school-aged children (SAC) of 5–14 years. The Malawi Demographic and Health Surveys (MDHS) of 2015-16 and Micronutrient Survey of 2015-16 used descriptive analysis to check for determinants of anemia in school-aged children. The results did not take into account the clustering nature of data set where estimates of effects on anemia in school-aged children can be compared by performing logistic regression model or multilevel logistic regression model where households can be quantified by communities and districts as second level and third level of analysis.

2.2. Anemia Studies conducted in Malawi

Among studies conducted in Malawi on anemia using a cross sectional data include the Malawi Demographic Health Survey and Malawi Micronutrient Survey 2015-16. In this study, hemoglobin concentration and prevalence of anemia in preschool children, school-aged children and pregnant women were investigated. Descriptive statistics (mean and frequencies) and simple logistic regression were used based on the individual-level within the specified population. The result of the study showed that anemia was found in 30% of preschool children, 22% of school-aged children and 21% of non-pregnant women of reproductive age where iron deficiency was relatively uncommon in all groups except young children with 22% of preschool children, 5% of school-aged children, 15% of non-pregnant women of reproductive age and 1% of men. The study made it difficult to check for the effects within and between clusters or dis-

tricts. (National Statistical Office, 2015).

In a study conducted by (Austin & Merlo, 2017), examined individual and community-level factors associated with childhood anemia, severe anemia, and hemoglobin (Hb) concentration in Malawi. Data from the 2010 Malawi Demographic and Health Survey was used. The multilevel regression models were constructed to analyze 2,597 children aged 6–59 months living in 849 communities. From the study, the results of multilevel analysis showed that both childhood anemia and severe anemia were negatively associated with child's age, no fever in the past 2 weeks and height-for-age, and positively associated with residing in poor household. Childhood anemia was negatively associated with female education. Child's age, no fever in the past 2 weeks and maternal Hemoglobin (Hb) levels were positively associated with child Hb concentration, while residing in poorest households was negatively associated with children's Hb concentration. The results could not explain the within and between community effects of anemia among the poorest household.

Kazembe and Ngwira (2015) fitted a multinomial cumulative logistic regression model on young children of 6-59 months to investigate the risk factors affecting the severity of childhood anemia in Malawi. This author aimed at investigating factors of childhood anemia by using multinomial ordered outcome model that extended to permit nonlinear effects of some continuous variables and spatial effects of district of residence. The spatial effects had some unknown influences like climate and environmental factors, access to good transport system and access to good child health care services. Inferences were based on imperial bayes framework and spatial variations using ordered bayesian and cumulative logistic regression which was used to measure severity of anemia. Their findings revealed substantial spatial variation with increased risk of anemia observed in some of the districts like Nsanje, Chikwawa, Salima, Nkhotakota, Mangochi, Machinga and Balaka. In addition, determinants like wasting, fever, underweight and stunting increased childhood anemia. Furthermore, infant anemia decreased with child's age and wealth index. The study did not take into account for individual and community variations.

2.3. **Anemia Studies in other countries**

Children with anemia develop low resistance to diseases and increased susceptibility to infection, poor cognitive development, impaired physical development, poor school performance and reduced work capacity with impaired social and economic development of the country (Gutema, Adissu, Asress, & Gedefaw, 2014). Similarly, (Dey & Raheem, 2016) stated that poverty increases the risk of an individual to be affected by infectious diseases, which may cause malnutrition and anemia.

In a study conducted by Chinese National Nutrition and Health Survey (CNNHS) 2010-2012 on prevalence of anemia in children and adolescents (6-17 years) old and its associated factors using descriptive analysis found that 6.6 percent of the children were anemic and of these 7.4 percent were girls and 6.0 percent were boys. Multi-variable logistic regression analysis was used to analyze the relationship between anemia and possible predictors such as age, sex, region type and income. Odds ratio (OR) and 95 percent confidence interval (CI) were determined using a logistic regression model and a two-tailed p-value of < 0.05 which was considered as statistically significant. On children living in rural and urban, the results showed that higher prevalence was greater in rural areas having greater risk of anemia. The study also showed negative correlation between household wealth and the prevalence of anemia. These results were similar to results of 2015-16 Malawi Micronutrient Survey showing higher prevalence of anemia in rural areas as compared to urban areas (National Statistical Office, 2015).

A study conducted in Ethiopia, defining anemia in children of (6 to 59 months) of age, described health variables as birth weight, childhood wasting, underweight, stunting, symptoms of acute respiratory infection, child fever and diarrhea, maternal anemia. In the analysis the bi-variate multilevel logistic regression were considered as candidates for multi-variable analysis and the result showed that children who were born in families with more than six children had higher odds of anemia than the first-older children. These results were similar to findings from a prior study done in New Delhi, India (Gebremeskel et al., 2020).

In Mali and Benin, Demographic Health Survey (DHS 2001) the risk factors of anemia were considered at the individual, household, and community levels where multilevel

analysis was applied to indicate clustering-level of anemia in communities. Comparative analyses were carried out using simple logistic regression and multilevel logistic regression models. Despite the immediate causes of anemia among children were known (such as malnutrition and infections), the impact of household and community socioeconomic determinants were only being explored, and the interrelationship between such contextual and individual factors remained under-studied. From their study, the outcome variable was anemia and explanatory variables were potential risk factors for anemia, which reflected the rationale for the multilevel analysis to be carried out in their study. Model technique by (Ngnie-Teta et al., 2007) were performed by applying multilevel analysis that allowed incorporating of explanatory variables at different levels of hierarchy. Likewise, in a study carried out by (Ntenda Morton. & chih Chuang, 2017) also applied multilevel logistic regression approaches to separate individual and household factors from contextual factors associated with moderate and severe anemia in children using MDHS 2010. The analysis was repeated by using multilevel logistic regression model techniques (Ngnie-Teta et al., 2007). The multilevel logistic regression model allowed incorporation of explanatory variables at different levels of the hierarchy. However, the results did not separate the contribution of individual characteristics to the risk of anemia contribution of the community, and could not compare on anemia clustering (Prieto-Patron, der Horst Zsuzsa V. Hatton, & Detzel, 2018).

A cross-sectional study conducted to determine iron status at the two schools in the suburbs of Jakarta on children and adolescents aged (6 to 18 years) old applied chi-square test for testing relationships between categorical variables. Independent T-Test was performed to compare mean values between two groups, and ANOVA test was performed to compare mean values between three groups or more. Normality test was performed using Kolmogorov Simirnov. P-value of less than 0.05 was considered as statistically significant (Syed et al., 2016). The results showed that there was no association between the prevalence of iron status according to age group. The limitation in their study was that they did not explore other causes of anemia and failed to compare the prevalence of iron deficiency anemia (IDA) and and iron deficiency (ID) between children with low socioeconomic status and children with high socioeconomic status.

The results of a study by (Sanku, 2013) which identified predictors of childhood ane-

mia 0-4 years in North-East India through ordinary logistic regression analysis showed that rural children were at greater risk of severe anemia. In the same study conducted in North-East India on childhood anemia (Sanku, 2013) did not take into account the stratified nature of data where the children were naturally nested into mothers, and mothers nested into households and households into primary sampling units (PSUs), and PSUs into regions. Multilevel regression model avoids the possible under-estimation of the parameters from single-level model when clustered data is used in the analysis. (Kumar Chowdhury et al., 2019) and (Hossain, Kamruzzaman, & Wadood, 2018) emphasizes that using multilevel models helps to account for the correlation structure of the data that frequently occurs in social sciences and in multistage survey sampling.

An epidemiological study of anemia in children, adolescent girls, and women in the country of Bhutan in South of Asia used multivariate analysis to estimate anemia prevalence and explored risk factors in children and women using data from Bhutan's National Nutrition Survey 2015. Models of individual and household factors of anemia were developed for the demographic groups separately using modified Poison regression and also using the generalized linear model command with robust standard errors. The results of the study explained that risk of anemia was greater in children who were younger.

Evidence shows that social economic status is important in the health of a child. For example, (Ngnie-Teta et al., 2007) and (Custodio et al., 2008) found that the children with lower living standards and those with lower social educational levels were at greater risk of anemia. Furthermore, in estimating the overall prevalence of anemia in Arba, Minch Zuria District, Southern Ethiopia, a binary logistic regression model was used to assess the possible association of independent and outcome variables. All covariates that were significant in the bivariate analysis were considered for multivariate analysis to control for possible confounder. To measure the strength of association between dependent and independent variables, Adjusted Odds Ratio (AOR) with 95 percent Confidence Interval (CI) was calculated. The level of significance was declared at p-value of < 0.05 and used Hosmer-Lemeshow to test model goodness of fit.

Prieto-Patron et al. (2018) described that in less developed countries, maternal, house-

hold and community factors have been reported to increase the risk of being anemic in early childhood such that the prevalence of anemia in infancy remains high. Additionally, traditional logistic regression and multilevel logistic regression analysis were applied to study the association between hemoglobin concentrations in; household, child, maternal and socio-demographic variables. It was found that child anemia was strongly associated with maternal anemia, household wealth, maternal education and low birth weight (Hossain et al., 2018). Using traditional logistic regression, a chi-square test for categorical variables was performed in assessing factors relating to anemia, and the response variable was binary indicating whether a person had anemia or not. The independent covariates were comprised of individual variables that included sex, age, fever, education, family size, income and wealth index (Ntenda Morton. & chih Chuang, 2017). Odds ratio (OR) was also used to proximate the prevalence ratio (PR) despite it had weaknesses in exaggerating the true relative risks (RR). A binomial regression model was recommended for estimating (RRs) and (PRs) in multiple analysis, but it had convergence problems. A Poisson regression was used to calculate Crude Prevalence Ratios (CPR) between potential risk factors and outcomes while adjusted Poisson model assessed the association between socio-demographic and economic factors of anemia. Results of the study indicated that percentage of children without anemia was higher in the urban setting than in rural community. These results were in agreement with the result of the study conducted in 2010-2012 by Chinese National Nutrition and Health Survey (CNNHS) on prevalence of anemia in children and adolescents 6-17 years old and its associated factors that also showed that higher prevalence of anemia was greater in rural areas than in urban areas.

A National Health Survey conducted in Indonesia in 2013 showed that the prevalence of anemia in school-aged children and adolescents tripled from a National Survey that was conducted in 2007. Children and adolescents were particularly susceptible to iron deficiency anemia (IDA) and iron deficiency (ID) because of their rapid growth and puberty. Teenage girls were at risk because of their menstrual bleeding. Low socioeconomic status in children and adolescence was also a strong risk factor for experiencing iron deficiency.

From the studies conducted in Malawi and other countries on anemia, little is known

about specific effects that can influence anemia in school-aged children 5-14 years old at micro-level and macro-level in different geographical locations in Malawi. This situation triggers a need to generate more knowledge about modeling effects of anemia in school-aged children in Malawi. Since children within same locality may have similar characteristics, the thesis reviewed the literature on health, social, geographical and economic consequences associated with child anemia to understand how these effects varied at individual-level, household-level, community-level and district level. This thesis checked for within and between geographical locations on effects of anemia in SAC in Malawi and also reviewed the literature to understand these consequences that were associated with child anemia.

2.4. **Literature on statistical models applied to hierarchical correlated data**

Hierarchical data that has a binary outcome is analyzed differently based on the objectives of the research. The most common approach being non-linear mixed effects model which uses likelihood-based approaches like in a study conducted by (Liang & Carriere, 2013) to investigate best methodological approaches that frequently arise in the analysis of non-independent discrete hierarchical medical data, a logistic regression was applied that assumed all observations were independent e.g., to examine the effect of alcohol abuse in a person while adjusting for age and sex of the person. The odds ratio was performed looking at the binary outcome of interest as (Yes/No) taking into account relative independent variables.

However, (Bates & Sarkinfada, 2007) approach was by using the mixed effect model which was fitted into the model by maximizing an approximation to the likelihood over the random effects. The weakness for using this numerical issue for maximum likelihood estimation in nonlinear mixed-effects models was the evaluation of the log-likelihood function of data, because it involved the evaluation of a multiple integral that, in most cases, did not have a closed-form expression. As a remedy, (Liang & Carriere, 2013) opted to use a quasi-likelihood to fit a generalized linear model (GLM) in their study of comparing statistical methods for analyzing discrete hierarchical data (A Case Study of Family Data on Alcohol Abuse) which estimated β as long as the data were sampled from a population in which the specified mean and variance functions were

correct.

Liang and Carriere (2013) observed a weakness in using generalized linear models (GLM) and Non-Linear Mixed Effect Model (NLME) where the data was hierarchical because of the auto correction structure that developed. For example, GLM analyze outcomes when assumption in linear regression analysis is violated. GLM has three main components; the systematic component, random components and link function which gives f(Y) as a linear combination of the predictors; $f(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$. For instance, if f is a logarithm, it will constrain the output of the model to positive numbers only, thus imposing a lower bound of 0 to response. Note that this is still a linear model. Although the relationship between Y and the predictors X_i is not linear, but the relationship between $f(y)$ and X_i. For a sample of N the random component can define the response variable $Y = (y_i, y_2, \cdots, y_N)$, and gives a probability distribution for Y and a binomial distribution in this case was assumed to have the outcome y_i as binary. When the outcomes are discrete and non-negative, the Poisson distribution was assumed with the systematic component that gave a linear combination of the covariates in the model as, $\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$.

The link function on the other hand provides the function relationship between the mean of random component, $E(Y) = \mu$, and the systematic component. The GLM is expressed as $g(m) = \alpha + \beta_1 X_1 + \beta_2 X2 + \cdots + \beta_k X_k$. The other link functions that are used in GLM is a binomial logit link function to model a binary outcome, which gives a GLM called binomial logistic regression $\log\left(\frac{p(Y)}{1-p(Y)}\right) = X\beta$, where $p(Y)$ is the probability of Y happening.

2.4.1 Mixed effect models

Mixed effect models provide unifying framework for analyzing observations that assumes independence when applied to correlated data. Mixed effect models rely on assumptions that random effects induces correlation among repeated measures. Careful consideration must be given to the interpretation of parameters when using nonlinear link function like logit (Liang & Carriere, 2013). However, mixed effect regression are associated with change in individual covariates which may not address the problem

statement in this study.

2.4.2 Generalized Estimate Equation (GEE)

Generalized Estimate Equation (GEE) are methods which are popularly applied for longitudinal studies and other correlated response data, particularly if responses are binary or in form of counts (Lipi, Alam, & Hossain, 2021). It is an extension of generalized linear model (GLM) first introduced by (Liang & Carriere, 2013). The method takes into account the within-subjects correlation by introducing the (n x n) correlation matrix $R(\alpha)$ which is fully characterized by the correlation parameter α. Although the generalized estimating equations' methodology considers correlation among the repeated observations on the same subject, it ignores the between-subject correlation and assumes subjects are independent. GEE aims at producing reasonable estimates of model parameters, along with standard errors, without specifying a likelihood function in its entirety, which can be quite difficult with a multivariate categorical response. We have to account for the correlation among the multiple responses that arise from a single subject, but we can largely estimate these correlations from the data without having to accurately specify the form of the correlation structure.

In GEE, the data set is split into some clusters with correlation within clusters, while correlations between clusters are assumed to be zero. In GEE literature, there exist different kind of working correlation structures such as exchangeable, auto-regressive of order one, toeplitz, unstructured etc. Inappropriate choice of correlation structure in GEE will lead to inefficient parameter estimation (Lipi et al., 2021). The main benefit of GEE is the production of reasonably accurate standard errors, confidence intervals with the correct coverage rates and development of techniques to model and estimate the between-cluster variation and the residual variances. The challenges with using GEE is that the procedure is not available in the stata package that is used in this study of modelling different correlations at different levels and cannot accommodate both cluster-specific intercepts.

The GEE method in estimating for β is an extension of the independence estimating

equation to correlated data. The GEE is given by the score function of

$$S(\beta) = \Sigma_{i=1}^{k} \frac{\partial \pi_i^T}{\partial \beta} V_i^{-1}(y_i - \pi_i(\beta)) = 0 \tag{2.1}$$

Where K equal to number of school-aged children 5-14 years old. $R_i(a)$ as the true correlation matrix of Y_i, true covariance matrix of Y_i is given as $V_i = A_i^{1/2} R_i(a) A_i^{1/2}/\phi$, where A_i is an n_i x n_i diagonal matrix with var (π_{ij}) as the j^{th} diagonal element and ϕ as the scale parameter (Liang & Carriere, 2013).

To account for GEE in the marginal model, we can assume that the correlation among school-aged children be the same. In so doing we can analyze the anemia data using the unstructured correlation matrix. But the drawback of using the GEE model is that they are mostly used in hierarchical linear modelling hence, can not be applied in this study.

2.4.3 Multivariate Models

2.4.3.1 Binary Logistic Regression

Binary outcomes are very common in medical research, and logistic regression is a popular modeling approach for binary responses and when the dependent variable is dichotomous and ordinal or multinomial, from a set of predictor variables that are continuous, discrete, dichotomous, or a mixture of any of these. A logistic regression will model the chance of an outcome based on individual characteristics. The logarithm of the chances are given as:

$$log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m \tag{2.2}$$

where π indicates the probability of event of being anemic and β_i are the regression coefficients associated with the reference group and the X_i explanatory variables. In this case, $\log\left(\frac{\pi}{1-\pi}\right)$ is the link function and it models the log of the odds of observing the outcome.

Logistic regression helps to investigate the impact of various explanatory variables on the response variable. Logistic regression allows one to predict a discrete outcome, especially when the outcome variable is a measure on a binary scale like when the responses may be Yes/No, or Success/Failure, the two categories can be expressed as

"1" and "0," so that they are represented numerically. The model is also known as polymers or polychotomous logistic regression in the health science field and as discrete choice model in econometrics (Berhie & Gebresilassie, 2016). Logistic regression can also be preffered when the independent variables are categorical and response variable is dichotomous. Application of logistic regression have also been extended to cases where the dependent variable is of more than two cases known as multinomial logistic regression. In this case the logit transformation is used with probability of occurrence ϕ.

$$\phi = \frac{exp(\beta_0 + \beta_1 X_1 +, ..., +\beta_k X_k)}{1 + exp(\beta_0 + \beta_1 X_1 +, ..., +\beta_k X_k)} \tag{2.3}$$

Where β_j is the coefficient corresponding to the predictor variable X_j and $j = 1, \ldots, k$, In which the logistic regression can be given as,

$$\log\left[\frac{\phi}{1-\phi}\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{2.4}$$

By algebraic manipulation, the logistic regression equation can be written in terms of an odds ratio for success:

$$\left[\frac{p(Y=1|X_i)}{p(1-p(Y=1|X_i))}\right] = \left[\frac{\pi}{1-\pi}\right] = exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) \tag{2.5}$$

$$log\left[\frac{p(Y=1|X_i)}{p(1-p(Y=1|X_i))}\right] = \left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k = \sum_{j=0}^{k} \beta_j X_{ij} \tag{2.6}$$

where i = 1, 2, …., n; j= 0, 1, 2, …., k

The coefficient can be interpreted as for example; the change in the log-odds of having an anemic child before a given period of time per unit change of corresponding covariates. In case of categorical predictor variable, it is interpreted as the log-odds of having an anemic child before a given period of time with a given category compared to the reference category (Dayton, 1992).

2.4.3.2 Assumptions of Binary Logistic Regression

As indicated in the above sections, the advantage of the logistic regression is that it has flexible assumptions as compared with discriminant analysis. There are, however, other assumptions one should consider for the efficient use of logistic regression as detailed

in (Berhie & Gebresilassie, 2016).

- Linear relationship exists through logit transformation of the dependent variable.
- The dependent variable is categorical to have an outcome.
- The dependent variable may assume a distribution from an exponential family (e.g. binomial, Poisson, multinomial, normal); binary logistic regression assume binomial distribution of the response.
- The groups for the predictors must be mutually exclusive and exhaustive.
- Larger samples are needed than for linear regression because maximum likelihood coefficients are large sample estimates. A minimum of 50 cases per predictor is recommended.
- There should not be severe collinearity among predictor variables.

However, the structure of data in this study is hierarchical and sample from such a population can be viewed as a multistage sample. Therefore, multilevel model may be appropriate that assumes a hierarchical data set with one single dependent variable that can be measured at the lowest level and explanatory variables at all existing levels.

2.4.4 Odds Ratio

In logistic regression the relationship between the response variable and the set of explanatory variables is not linear. From (Berhie & Gebresilassie, 2016), the logistic probabilities from a model containing one dichotomous covariate coded 0 and 1 and the odds of the response being present among individuals with $x = 1$ and $x = 0$ is given as:

$$\text{Odds } (X = 1) = \frac{P(Y|X=1)}{1-P(Y|X=1)}$$
$$\text{Odds}(X = 0) = \frac{P(Y|X=0)}{1-P(Y|X=0)}$$

The odds ratio denoted as OR, is the ratio of the odds for $x = 1$ to the odds for $x = 0$ shown as: $OR = \frac{odds(X=1)}{odds(X=0)}$. In this regard, the odds of the response variable is multiplied by $OR = e^{\beta}$ for change from reference category to the estimated category of the given explanatory variable. This means that the odds at level $X + 1$ equal the odds at x

multiplied by e^{β} and when the odds is less than one, it indicates the occurrence is less likely than non-occurrence and if it is greater than one indicates the occurrence is more likely than non occurrence.

2.4.4.1 Clustered Data

Clustered data arise when subjects are physically grouped into different groups (or clusters), with some groups containing multiple subjects. This grouping can be due to things like geography or shared relationship like in DHS data in Malawi which is grouped into households, communities, districts and region. This grouping gives the data a multilevel structure in which subjects can be nested within their cluster groups. In this study we will restrict the analysis to multilevel structure where school-aged children will be clustered within communities.

2.4.4.2 Two-Level Model

For example, in educational studies hierarchical data are very common and a two-level model are applied. Like a study of factors that affect student performance might measure for each student and each group of exams whether the student passed. Students are nested within schools, and the model could incorporate variability among student as well as variability among schools. The model then could analyze effect of student's characteristics as X_1 = gender and X_2 = score on each achievement exam a year ago, and effects of characteristics of the school the student attends such as X_3 = the school budget, per student, and X_4 = average class size. Observation for the same student on different exams would probably tend to be more alike than observation for different students. Likewise, students in the same school might tend to have more-alike observations than students from different schools, because students within a school tend to be similar on characteristics such as socioeconomic status. Relevant models contain explanatory variables and random effect terms for the student and for the school (Hossain et al., 2018).

When the structure of data is obtained in the Demographic Health Survey (DHS) the clustering sampling scheme often introduces multilevel dependency or correlation among the observations that can have implications for model parameter estimates. For multistage clustered samples, the dependence among observations often comes from several

levels of the hierarchy. The problem of dependencies between individual observations also occurs in survey research, where the sample is not taken randomly but cluster sampling from geographical areas is used instead. In this case, the use of single-level statistical models is no longer valid and reasonable. Hence, in order to draw appropriate inferences and conclusions from multistage stratified clustered survey data may require tricky and complicated modelling techniques like multilevel modelling.
For Example: Let Y_{ij} = whether student i in school j passed exams (1=Yes, 0=No) then explanatory variable for level one model will be:

$$logit[P(Y_{ij}=1)] = U_{ij} + \alpha_j + \beta_1 X_{1ij} + \beta_2 X_{2ij} \tag{2.7}$$

In the equation the random effect U_{ij} for student i in school j accounts for variability among students. For the school-specific explanatory variables, the level-two model takes the school-specific term α_j from the level-one model and expresses it as:

$$\alpha_j = S_j + \alpha + \beta_3 X_{3j} + \beta_4 X_{4j} \tag{2.8}$$

where S_j is a random effect for school j. This random effect reflects heterogeneity among the schools due to school-specific explanatory variables not measured and substituting the level-two model into the level-one model it produced:

$$logit[P(Y_{ij}=1)] = U_{ij} + S_j + \alpha + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij} \tag{2.9}$$

This is a multilevel model with random intercepts U_{ij} at the student level and S_j at the school level.

2.4.5 Multilevel Logistic Regression

The aim of multilevel logistic regression is to estimate the odds that an event will occur the (yes/no) outcome while taking dependency of data into account, an example is that of pupils nested in classroom in the previous paragraph of level-2 model (Sommet & Morselli, 2017). Multilevel logistic regression allows to estimate odds as function of lower-level variables e.g., pupils age, higher level variables, classroom size and also the way they are interrelated (cross-level interactions). Another example where multilevel regression can be used is in social psychology such as when outcome variable describes the presence/absence of an event (Sommet & Morselli, 2017). Note, mul-

tilevel modelling is flexible enough to deal with this kind of unbalanced data, that is having unequal numbers of participants within clusters. With this data structure, you cannot run a standard logistic regression analysis. The reason is that this violates one of the most important assumptions in the linear model, namely the assumption of independence (or lack of correlation) of the residuals (Merlo, Wagner, Ghith, & Leckie, 2016).

Multilevel logistic regression modelling notably aims to disentangle the within-cluster effects, the extent to which some participant characteristics are associated with the odds from the between-cluster effects. The fact that clusters, districts have a variety of independent variables; environmental factors, health service provider, level of education of the people living in the community, level of educated family, access to safe drinking water, sanitation and different infrastructures to encourage the reduction of anemia in under-five children and school aged children. Indeed, not only community-level differentials but also there are individual-level factors attributed for under-five children and school-aged children anemia and demographic factors of children as well. This differential among individuals, communities, districts and also through continent-level indicates the facts that, the rate of child anemia in developed and developing country has different structure. But, so many studies done using single-level eliminate those variation across community-level or districts-level regarding school-aged children anemia, under-five children anemia and women of reproductive age anemia, in the world-wide and at national-level that invites errors. In fact, there is clear heterogeneity among the individual and community-level characteristics that leads to variations while clustered those factors at single-level. First, it is better to check whether there is heterogeneity proportion in data for school-aged children anemia between communities in Malawi before going to multilevel analysis.

In summary, a multilevel logistic regression model is used to account for lack of independence across levels of nested data (e.g., school children nested within communities). In this study, multilevel binary logistic regression model would be adopted to the variations on anemia status of school-aged children within communities or districts in Malawi and the basic data structure of the two-level logistic regression will be the collection of N groups (communities) and within-group j (j=1, 2, ..., N), while random sample of level-one units would be school-aged children.

2.4.5.1 Variance Component Model

The variance component two-level model for a dichotomous outcome variable refers to level-two units (community) that specifies the probability distribution for group dependent probabilities P_j in $Y_{ij} = P_j + E_{ij}$ without taking further explanatory variables into account. The variance component model focuses on models that transform probabilities to normal distribution by the formula:

$$log\left[\frac{P_{ij}}{1-P_{ij}}\right] = \beta_{0j} + U_{0j} \tag{2.10}$$

Where β_0 is the population average of the transformed probabilities and U_{0j} the random deviation from this average for group j. Intraclass correlation coefficient (ICC) represents the proportion of the total variance that is attributable to between-group differences and it provides an assessment of whether or not significant between-groups variation exists. Then the intraclass correlation coefficient (ICC) at community level is given by:

$$ICC = \rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \tag{2.11}$$

where σ_u^2 is the between-groups variance which can be estimated by U_{0j} and σ_u^2 (Janjua, Khan, and Clemens (2006)).

2.4.5.2 The Random Intercept Logistic Regression Model

In the random intercept logistic regression model, the intercept is the only random effect meaning that the groups differ with respect to the average value of a response variable. The random intercept model is used to model unobserved heterogeneity in the response by introducing random effects. The random intercept models have many applications like, estimating the community effects on anemia status, adjusting for individual women level factors, and within the model, evaluate the performance of the communities in anemia status. This can be done by obtaining the odds ratio for each community. This community effect can be a measure of the situation of anemia status due to the community relative to the average of all communities. If the odd of anemia status for community effects is sufficiently larger than one, the community is considered to have performed worse than average and if it is significantly smaller than one,

the community is considered to have better performance.

2.4.5.3 Estimation of Model Parameters

The Likelihood method: The likelihood method is a general estimation procedure which produces estimates for the population parameters that maximize the probability of observing the data that are observed. Assuming the conditional distributions of Y_{ij} is given the random effect U_j which are independent of each other, the conditional density of Y_{ij} is given by P_{ij}:

$$Y_{yij/uj} = \left(Y_{ij}/U_{ij}\right) \sim Bernoulli \tag{2.12}$$

For the two-level logistic Bernoulli response model, the random effects are assumed to be multivariate normal and independent across units, the marginal likelihood function is

$$l(\beta,\alpha) = \pi f \pi[(\pi_{ij})^{Y_{ij}}(1-\pi_{ij})^{1-Y_{ij}}] \tag{2.13}$$

where α is variance covariance matrix.

$$\pi_{ij} = [1+exp(-X_{ij}\beta_j)], \beta_j = \beta + U_j \tag{2.14}$$

where α_{ij}, α is typically assumed to be the multivariate normal density.

2.4.5.4 The Intraclass correlation Coefficient (ICC)

A key concept in multilevel analysis is the intraclass correlation coeffcient (ICC). The ICC quantifies the proportion of variation in the outcome that can be attributed to systematic differences in the outcome between clusters (Janjua et al., 2006) . The ICC tells the degree of similarity between individuals belonging to the same group. For example, from an epidemiologic perspective, this allows one to ascertain how much of variability in outcome can be attributed to the clustering unit. How much variability in patient outcomes can be attributed to the hospital to which they were admitted. If there is no systematic between-hospital variation in patient outcomes, then we say there is absence of hospital effect on patient outcomes (Snijders, 2001).

The ICC is calculated by dividing the between-cluster variation in the outcome by the total variation in the outcome. The ICC is equal to correlation between two individuals

drawn from the same group, and it ranges from 0 to 1. If it is 0, there is no evidence of clustering effects in the data. If ICC is 1, then the grouping accounts for all variations in the data (Thompson, Fernald, & Mold, 2012). This will mean all individuals within the same group have identical responses on the outcome variable. The ICC always varies according to measure and to the type of clustering that is present.

2.4.5.5 Model Selection

Model selection is important to find goodness of fit and complexity. The need to select a model is of great importance in statistics to ensure goodness of fit and adjust or penalize for model complexity. Therefore, several models are fitted to find the best fit model. Models with smaller values of AIC or BIC shows is the best fit model and those with larger values are considered not the best fit model.

Akaike Information Criterion (AIC): For maximum likelihood or empirical Bayesian, one can use the Akaike Information Criterion. The Akaike (1973, 1974) information criterion was developed as estimators of the expected Kullback-Lieber discrepancy between the model generating the data and a fitted candidate model. One of the statistics to select the best model fit is the AIC (Curini, Franzese, & Steenbergen, 2020).

$$AIC - 2K - 2ln(L) \tag{2.15}$$

Where K is the number of parameters in the statistical model, and L is the maximized value of the likelihood function for the estimated model. The AIC is calculated for each model under consideration using the same data and the model with the lowest/minimum AIC is chosen. The 2K is a penalty for discouraging fitting too many variables in the models which always leads to a smaller likelihood. This provides the tradeoff between over fitting and optimum model fit.

Bayesian Information Criterion (BIC): The BIC is another model selection in statistics and is based on the empirical log-likelihood, and it does not require the specifications of the priors. The BIC is favoured in situations where the priors are difficult to set. Both BIC and AIC penalize the model complexity. The best fitted model has less BIC values and is given as

$$2ln(L) + kln(n) \tag{2.16}$$

Where L is the maximized value of the likelihood function of the model and k is the number of free parameters to be estimated. N is the number of observations.

2.4.6 Literature Summary

Most of the literature used in this chapter was covering children from 6-59 months. Literature on children aged 5-14 years was lacking because only few studies have been conducted in other countries like Ethiopia and China. It is even difficult to find published papers relating to community effects of anemia in school-aged children 5-14 years in Malawi. Now, the study of modeling community effects on anemia in school-aged children using multilevel logistic regression analysis will be the first of its kind in Malawi.

In this chapter, we reviewed modeling of anemia using different statistical methods on how to understand the related effects of social-economic factors and demographic factors. The literature has stated that anemia is a widespread public health problem and severe anemia is a significant cause of childhood mortality. The World Health Organization (WHO) considers anemia as a major public health problem. From the papers studied, the response variable (anemia status) was dichotomous indicating whether one is anemic or not. The explanatory variables, which used to determine the status of anemia in children 6-59 months were: socioeconomic, demographic, health and environmental factors. From the source of data, the following variables were considered; region, place of residence, wealth index, marital status, child's age in months, sex of children, husband/partner's education level, given vitamin A, source of drinking water, mothers current working status, and child's size at birth.

From the literature, it has also been found that biological related causes of anemia, socioeconomic and demographic factors associated with anemia are well documented. Notably, checking variations of effects on anemia within community and between community have not been widely reported. It is evident that using hierarchical data will allow us to fit a multilevel regression model on effect of anemia in school-aged children 5-14 years old while accounting for systematic unexplained variation among the communities. Also to estimate the true effect on anemia will also allow the implementation of Government policies for future planning.

CHAPTER 3

MATERIALS AND METHODS

3.1. **Data description**

This study used the weighted data of 2015-16 Malawi Demographic Health Survey which had to run parallel with the Malawi Micronutrient Survey of 2015-16. These surveys were conducted by National Statistical Office from 19 October 2015 to 18 February 2016 in joint collaboration with the Ministry of Health (MoH), Centre for Disease Control and Prevention (CDC), the Community Health Services Unit (CHSU) and International Care Foundation (ICF). The survey was based on a nationally representative sample that provided estimates at the national and regional levels and for urban and rural areas with key indicator estimates at the district level. The survey included 26,361 households from MDHS and 2,114 households selected from the 26,361 households for MNS.

3.1.1 Variables included in the study

The variables considered in the study were at national level. As discussed in literature review they cover demographic characteristics, health characteristics, social, community characteristics and geographic characteristics. These variables were classified as dependent and explanatory variables (Table 1).

3.1.1.1 Dependent Variable

The dependent variable for the thesis was anemia. It was dichotomous coded as (1) if the child had anemia and (0) if the child had no anemia.

3.1.1.2 Explanatory Variables

In the study the explanatory variables like demographic, economic, health and socio were expected to have impact on anemia in school-aged children (SAC) and were classified as individual level variables and community level variables as given below:

- *Individual variables*: age, sex, mother alive, father alive, religion, mother age at marriage and number of children ever born to mother's.
- *Health variables*: episode of diarrhea, stunting, thinness, obesity, underweight, overweight, inflammation, if child took vitamin A, malaria in the past two weeks.
- *Social and community variables*: education, source of drinking water,
- *Economic variables*: Business, Employment and wealth.
- *Geographical variables*: rural, urban

Table 1: Description of variables used in the Analysis

Characteristics	**Description**
Anemia	For 5 - 11 years $<$ 11.5g/dL, 12 - 14 years $<$ 12.0g/dL
Education Attainment	No education = 1, incomplete primary = 2, complete primary = 3, and incomplete secondary = 4
Child Age	5 - 10 years and 11 - 14 years categories
Stunting	Length or height for age z-score < -2
Thinness	BMI for age z-score <-2 or BMI <18kg per metre square
Obesity	BMI for age z-score > 2
Underweight	WAZ $<$ -2 for 5n- 10 years WHO growth standards
Overweight	BAZ > 2 WHO growth standards for 5 - 14 years
Inflammation	High level of c-reactive protein (CRP) in blood if $>$ 5mg/L
Malaria in the last two weeks	If a child had malaria the past two weeks = 1 and 0 otherwise
Wealth Index	Poorest = 0, poorer = 1, middle = 3, richer = 4, richesr = 5, reference richest
Life status of parent	Mother alive = 1(yes), Father alive = 1(yes) and 2(No) otherwise
Vitamin A	$>$ 0.060 if child received vitamin A = 1 and 0 otherwise
Iron Deficiency	$<$ 15mg/L
Sex of Household Head	Female = 1, and Male = 2
Household size	1 - 5 = 1; 6 - 10 = 2; and 11 - 15 = 3
Residence	Urban = 1 and rural = 2
Child sex	Female = 1, and Male = 2

Data source Micronutrient survey 2015 -16

3.1.2 Sampling Design

The MNS 2015-16 was selected as a subsample of the MDHS 2015-16 to produce estimates of key indicators for the country as a whole, as well as results stratified by region (North, Central, South) and residence (urban, rural). A subsample of 105 clusters (35 clusters in each of the 3 regions) were randomly selected from the 850 MDHS clusters (Table 2). The sample consisted of 20 households that were selected in urban areas and

22 in rural areas. In each selected household, all eligible participants (defined as usual members of the household who spent the night in that household before the survey) were invited to participate. In the MNS study, school-aged children from 6 households were randomly selected in a cluster (National Statistical Office, 2015).

Table 2: Micronutrient sample allocation of clusters by region and residence

	Number clusters allocated			**Number households allocated**		
	Urban	Rural	Total	Urban	Rural	Total
North	8	27	35	160	594	754
Central	8	27	35	160	594	754
South	8	27	35	160	594	754
Malawi	**24**	**81**	**105**	**480**	**1782**	**2262**

From (Table 2), the sample allocations were derived using the information obtained from the Malawi Demographic Health Survey (MDHS) 2015-16 and the Malawi Micronutrient Survey (MNS) 2015-16. The MNS was conducted in 2, 262 residential households, including 480 households in urban areas and 1,782 households in rural areas. The average number of women age 15-49 per household was 1.09 in urban areas and 0.94 in rural areas. The average number of children age 5-14 years per household was 1.67 in rural areas meaning that a household had an average of at least two children per household in rural areas and 1.42 in urban areas, meaning one child per household in urban areas National Statistical Office (2015).

3.2. **Statistical Methods**

In this study multilevel logistic regression was used to model community effects on anemia in school-aged children 5-14 years old. Multilevel model for measuring community effects, investigating correlates of anemia and the impact of community variance components on the coefficients of the estimated model on anemia in school-aged children 5-14 years old was developed. The response variable of the study was anemia. Ordinary logistic regression was the obvious model of choice when one thinks of modeling a binary outcome. Considering the nature of data used in this study and considering cluster sampling, multilevel dependency among the observations that could have implications for model parameter estimates was introduced and used for the analysis. Since the data was from multistage-clustered samples, the dependence among observations came from several levels of the hierarchy. The use of single-level statistical models was no longer valid and reasonable in this study. Consequently, when using the single-level (logistic regression) most of the factors would appear significant which would result in giving wrong policy implications for Malawi. In order to draw appropriate inferences and conclusions from multistage stratified clustered survey data, application of the modeling techniques as multilevel modeling was applied. (Messelu & Trueha, 2016).

3.2.1 Binary logistic Regression Model

The binary logistic regression model was used to investigate effects of predictors on anemia before going in to multilevel modeling.

We focused on using two-level hierarchical data individuals and community levels in

estimating the effect on anemia (binary outcome) in children aged 5-14 years. The response variable was defined below whose link function as defined in section 3.2.11.

$$Y_{ij} = \begin{cases} 1 & \text{for children having anemia} \\ 0 & \text{for normal(not anemic) children} \end{cases} \tag{3.1}$$

with probability $P_{ij} = P(Y_{ij} = 1/X_{ij}, U_j)$ being the probability of children with any anemia for the i^{th} child in the j^{th} community and the probability $1 - P_{ij} = P(Y_{ij} = 0/X_{ij})$ being the probability of non-anemic (normal) i^{th} children, for the children in the j^{th} communities. Here, Y_{ij} follows a Bernoulli distribution.

Let π denote the proportion of a school-aged child having anemia.

$$P(Y_{ij} = 1) = \pi_{ij}, P(Y_{ij} = 0) = 1 - \pi_{ij} \tag{3.2}$$

And $Y_i \sim Bernoulli(\pi_{ij})$

The model in the binary logistic regression will be defined as :

Let $Y_{n\times 1}$ be a dichotomous outcome random variable with categories 1 (presence of anemia) and 0 (absence of anemia). Let $X_{n\times(k+1)}$ denote k-predictor variables of the response where i^{th} school-aged child has anemia given that the vector of the predictor variable X_i is denoted by $P_i = P(y_i = 1/X_i)$. This can be expressed as:

$$logit[P_i] = log\left(\frac{P_i}{1-P_i}\right) = \Sigma_{j=0}^{k}\beta_j X_{ij}, i = 1,2,...,n; j = 0,1,...,k \tag{3.3}$$

and the binary logistic regression data matrix of k predictor variable on anemia in school-aged children can be given as:

X =

$$\begin{bmatrix} 1 & X_{11} & X_{12} & . & . & . & X_{1k} \\ 1 & X_{21} & X_{22} & . & . & . & X_{2k} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & X_{n1} & X_{n2} & . & . & . & X_{nk} \end{bmatrix} \sim nx(k+1), \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_k \end{bmatrix} \sim (k+1)xl \qquad (3.4)$$

where X - is the design matrix

β - is the vector of unknown coefficients of the covariates and intercept.

3.2.2 Checking for Clustering in Data

3.2.2.1 Intraclass correlation coefficient (ICC)

Intraclass correlation coefficient (ICC) is a key note in multilevel analysis. The ICC quantifies the variations of the outcome that can be attributed to systematic differences in the outcome between communities. The ICC in the study was explained by the degree of similarity between individuals belonging to the same group and the variability attributed to the community. If no systematic between community variation in school-aged children, then we may conclude that community effect on school-aged children were not existing.

The ICC was calculated by dividing the between-community variation in the outcome by total variation in the outcome. This can be similar to the process of comparing the between and within group variances in the analysis of variance. The ICC was equal to correlation between two individuals drawn from the same group and it could range from 0 to 1. If zero, meaning there was no evidence in clustering effects in the data. If ICC was 1, then clustering existed in the data, showing that all school-aged children within the same community had identical responses on the outcome variable. Note that outcomes of ICC are not rarely 0 or 1.

The ICC can be estimated by mathematical formula:

$$ICC = \rho = \log \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \tag{3.5}$$

where σ_u^2 was the between-groups variance which was estimated by U_{0j} and σ_e^2 was the within-group variance.

3.2.3 Model diagnostics and adequacy checking

3.2.3.1 Multicollinearity Diagnostics

First, one has to check for multicollinearity before analyzing the data set using multilevel regression analysis. Multicollinearity refers to condition when two or more independent variables are correlated to each other. There are various statistical techniques that can be applied to measure correlation between two variables, and one of them was the variation inflation factor (VIF) which was appied in the analysis in this study as a technique that could detect the multicollinearity by computing the r-squared metric

$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

VIF computation formula for variable $'i'$

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

squared metric formulation

the R-squared metrics measures how well data fits between 0 and 1,whereby, values close 1 reflects good model meaning no multicollinearity exists for a given data sets. when VIF increases there exists multicollinearity;

VIF = 1: no multicollinearity

VIF = between 1 to 5: moderate multicollinearity

VIF = >5: High multicollinearity

Multicolliniarity was checked by using a variation inflation factor (VIF). Independent variables were checked in the model if multicolinearity were causing a problem or not. Interactions between variables was checked for the variables which were found to be significant. Also model selection was carried out by using Akaike information criteria (AIC) to check for a better explanatory model.

Similarly, in the correlation matrix it was not easy to spot the multicollinearity between variables and the combination of variables. We checked for magnitude of standard errors for each variable. If the standard error were very large it would apply that multicollinearity existed and the model would not be statistically stable. Solving this problem would involve omitting the variable with high colliniarity (De Leeuw & Meijer, 2008) .

3.2.3.2 Null Model

The null model in level-two model variable referred to a population of level-2 units and described the probability distribution for the dependent probabilities π_j that is the probability that i^{th} school-aged child in j^{th} community had anemia without including any explanatory variables. The model with $f\pi_j$ have a normal distribution and can be expressed for a general link function as: $f(\pi_j) = \beta_0 + U_{0j}$ where $f(\pi_j)$ is the population average of the transformed probabilities of β_0. U_{0j} is the random deviation from the average for community j. The log odds have a normal distribution with: Logit $(\pi_j) = \beta_0 + U_{0j}$

where U_{0j} is assumed that they are independent random variables with a normal distribution and with mean of zero and variance of σ_0^2.

3.2.4 Multilevel Models

In most cases, the data generated in public health have hierarchical structures. The researchers usually come across or generate data that has been grouped in some ways, be it natural grouping or statistical. For instance, in demographic and healthy surveys (DHS) and Micro-nutrient Survey (MNS) we often get data that was grouped either at community level, or villages which were nested within a traditional authorities that were nested within district and then regions which forms the national data. In the DHS/MNS data, individuals, households, communities (clusters), traditional authorities, districts, regions were levels of hierarchy that the collected data at individual-level (level-1) passes to national-level (level-4). In DHS/MNS for example, interventions are made at District-level but measurements are made at individual-levels because they are close to each other within the community. In a similar way, the expectation was that SAC within the community would have similar characteristics because they were in the same environment. This is common to data that has a hierarchical structure, sometimes called multilevel structures or clustered data.

Dependence of data in multilevel structures disqualifies use of routine data analysis techniques like the binary regression analysis. However, models found in (Guo & Zhao, 2000; Skrondal & Rabe-Hesketh, 2007) made the hierarchical analysis possible. The hierarchical models were categorised into two parts: random effect and fixed effect. Fixed effect models measured effects of group level like district effects and community effects e.t.c; while random effects measured averages of variable like the average prevalence of anemia in each community (cluster) in Malawi etc. The model part in the multilevel models were measured at once.

The importance of multilevel models could not be overemphasized. Amongst them includes: (a) Correct inferences; traditional multiple regression techniques that treat the units of analysis as independent observations. One consequence of failing to recognise hierarchical structures is that standard errors of regression coefficients are underestimated, leading to an overstatement of statistical significance. Standard errors for the coefficients of higher-level predictor variables are most affected by ignoring grouping (Reviews, 2016); (b) substantive interest in group effects; in many situations are a key research question concerns of the extent of grouping in individual outcomes and the identification of 'outlying' groups. In evaluations of community performance, for example, interest centres on obtaining 'value-added' community effects on children whether anemic or not. Such effects correspond to community-level residuals in a multilevel model which is adjusted for prior wellness (Hox, 1998; Sanagou, Wolfe, Forbes, & Reid, 2012).

Despite, multilevel data having challenges that was taken into consideration in the thesis. The first one was that variance quantification at higher level needed to account for the variance in the lower level which most researchers miss. The second was that dependence of data would result in narrow confidence levels which were desirable but not true reflection of variations that existed. Therefore, inferences made using these confidence intervals were statistically incorrect. Thirdly, too many levels would need larger samples to make inferences. (Skrondal & Rabe-Hesketh, 2007).

3.2.5 Multilevel logistic Regression Model

Before going to multilevel modeling we looked into binary logistic regression model which was used to investigate the effect of predictors on the probability of having anemia in equation 3.2 and 3.3 where a dependent variable was stated. The decisions were made at community level while outcome was measured at individual level in households.

Despite that, communities in Malawi have a variety of environmental effects like social, health, economic, geographical that would contribute to effect on anemia in SAC. We would experience the community effects and individual effects on anemia in SAC despite that many studies concentrate more at individual differences, eliminating those variations across the communities or district levels. Infact, there was a clear heterogeneity among the individual and community level characteristics that led to variations while factors were clustered at single level.

In order to build multilevel logistic regression models we used level-2 model to check anemia variations between communities in SAC. In the study, the basic data structure of the level-2 logistic regression was a collection of N groups (Communities) and within-group j for $(j = 1,2,...,N)$ a random sample of level N_j level- 1 units (school-aged children). The outcome variable anemia was a dichotomous by $Y_{ij} = 0$ (meaning not having anemia) and that $Y_{ij} = 1$ (meaning school-aged child i was found anemic) in community $j(i = 1,2,....,n_j$, and j = 1, 2, ..., N).

In order to check if multilevel analysis could be applied we tested for heterogeneity of the proportions between communities. A chi-square based non parametric test and parametric test was used. A non parametric test was used to test if systematic differences between communities existed. The test statistics:

$$\chi^2 = \Sigma_{j=1}^{n} n_j \frac{(P_j - P)^2}{p(1-P)} \tag{3.6}$$

Where, $p_j = \frac{1}{n_j}\Sigma_{i-1}^{n_j} Y_{ij}$, the proportion of SAC who are anemic in the community j for:

$$P = \frac{1}{K}\Sigma_{j=1}^{N}\Sigma_{i=1}^{n_j}Y_{ij} \tag{3.7}$$

To find the proportion of SAC being anemic we applied $K = \Sigma_{j-1}^{N} n_j$ where the χ^2 followed approximately chi-square distribution with N - 1 degrees of freedom. Note that p_j was an estimate for the group-dependent probability p_j and an estimator for variance of P_j which could be obtained by using:

$$\tau^2 = S_{between}^2 - \frac{S_{within}^2}{n} \tag{3.8}$$

and

$$S_{within}^2 = \frac{1}{K-1}\Sigma n_j p_j (1-p_j)$$

3.2.6 Heterogeneity Proportions

The basic data structure of the level-2 logistic regression in the study was the collection of N (communities or clusters) and the within-community j (j = 1,2,.....,N). Also a random sample N_j of level-1 units were the (school-aged children in households). By letting the response variable $Y_{ij} = 1$ and if i^{th} school-aged child in the j^{th} community had anemia and $Y_{ij} = 0$ otherwise.

Let the outcome variable in equation 3.3, $Y_{ij}(i = 1,2,...,nj; j = 1,2,...,N)$ denoted for level-1 where unit i is nested in level-2 community j. The total sample size being $M = \Sigma_{j=1}^{N} n_j$. If explanatory variables were not taken into account the probability of SAC having anemia would be assumed to be constant in all the communities (número 442, 2012).

Let the probability of SAC having anemia in the community j be denoted as π_j. The dichotomous outcome variable for the SAC i in community j where Y_{ij} be expressed as the sum of the probability in community j, π_j was the average of i levels in community j, $E(Y_{ij}) = \pi_j$) and the individual dependent residual being:

$$y_{ij} = \pi_j + (1 - e_{ij}) \tag{3.9}$$

the residual term assumes a mean of zero and variance.

$$var(\varepsilon_{ij}) = \pi_j(1-\pi_j) \tag{3.10}$$

The community sample average was the proportion of SAC having anemia in community j written as: $\hat{\pi}_j = \frac{1}{n_j}\Sigma_{i-1}^{n_j} Y_{ij}$
where: $\hat{\pi}_j$ was an estimate for community dependent probability π_j.

3.2.7 Test for Heterogeneity

For a proper application of multilevel analysis we need to test the heterogeneity of between communities (número 442, 2012). To test if systematic differences between communities exist, the test statistics commonly used was the chi-square for contingency table (Franke, Ho, & Christie, 2012).
It was written as:

$$\chi^2 = \Sigma_{j=1}^{n} n_j \frac{\left(\hat{\pi}_j - \hat{\pi}\right)^2}{\hat{\pi}\left(1-\hat{\pi}\right)} \sim \chi^2(N-1) \tag{3.11}$$

3.2.8 Random Intercept Binary Logistic Regression Model

When data is from different communities, a varying - intercept model can be interpreted as a model with different intercept within each community (Howie, 2008). In this case the intercept model considered only the random effect of SAC meaning that the communities differed with respect to the SAC who were anemic, but not explaining differences between communities. The random intercept model expresses the log odds and the logit of P_{ij} as a sum of linear functions of the explanatory variables as:

$$logit(P_{ij}) = log\left[\frac{P_{ij}}{1-P_{ij}}\right] = \beta_{0j} + \sum_{h=1}^{k} \beta_h X_{hij} \tag{3.12}$$

for $i = 1,2,3,\cdot,n_j; j = 1,2,\cdots,k$

Where, logit(P_{ij}) does not include level-1 of SAC having anemia. β_{0j} is assumed to vary randomly and is given by the sum of average intercept of β_0 and community dependent deviations of U_{0j}. Replacing $\beta_{0j} = \beta_0 + U_{0j}$ in equation (3.12) will get:

$$logit(P_{ij}) = \beta_0 + \Sigma_{h-1}^{k} \beta_h X_{hij} + U_{0j} \tag{3.13}$$

now solving for P_{ij}

$$p_{ij} = \frac{e^{\beta_0 + \sum_{h=1}^{k} \beta_h X_{hij} + U_{0j}}}{1 + e^{\beta_0 + \sum_{h=1}^{k} \beta_h X_{hij} + U_{0j}}} \tag{3.14}$$

from equation 3.13 it did not include a level 1 residual just because it is an equation for the probability P_{ij} Where, β_h was the unit difference between X_h values of individuals in the same community which was associated with the log-odds with the difference of β_h. U_{0j} was the random part of the model and was assumed that they were mutually independent and normally distributed with mean zero and variance of σ_0^2.

3.2.9 The Random coefficients Logistic Regression Model

In random coefficient model, both the intercept and slopes differ across the communities. Consider k, the explanatory variables $X_1, \cdots, X_k$ and values of $X_h (h = 1, \cdots, k$ which can be indicated as X_{hij} for $h = 1, \cdots, k; i = 1, \cdots, n$ and $j = 1, \cdots, N$. Some of these variables could be level-1 variables where the effects on anemia probability may not be the same for all the school-aged children in a given locality. The effects on anemia probability may depend on the individuality of individual school-aged children but on the same time on their localities and was donated as P_{ij}. the outcome variable was expressed as the sum of effects on anemia probability which was the (expected value of the outcome variable) and the residual term e_{ij}. where,

$$y_{ij} = p_{ij} + e_{ij} \tag{3.15}$$

The residual e_{ij} are assumed to have mean zero and variance σ_e^2. The logistic regression models with random coefficients expresses the log-odds of logit P_{ij}, sum of a linear function of the explanatory variables with randomly varying coefficients. That is:

$$logit\left(P_{ij}\right) = log\left(\frac{P_{ij}}{1 - P_{ij}}\right) = \beta_{0j} + \beta_{1j} X_{1ij} +, \cdot, + \beta_{kj} X_{kj} \tag{3.16}$$

let $\beta_{0j} = \beta_0 + U_{0j}$ and $\beta_{hj} = \beta_h + U_{hj}$ for $h = 1, \cdots, k$

$$logit\left(P_{ij}\right) = log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \beta_0 + \sum_{h=1}^{k} \beta_h X_{hij} + U_{0j} + \sum_{h=1}^{k} U_{hj} X_{hij} \tag{3.17}$$

P_{ij} can be solved as:

$$\frac{e^{\beta_0 + \sum_{h=1}^{k} \beta_h X_{hij} + U_{0j} + \sum_{h=1}^{k} U_{hi} X_{hij}}}{1 + e^{\beta_0 + \sum_{h=1}^{k} \beta_h X_{hij} + U_{0j} + \sum_{h=1}^{k} U_{hj} X_{hij}}} \tag{3.18}$$

Therefore, a unit difference between the X_h values of two school-aged children 5-14 years in the same community is associated with a difference of β_h in their log-odds, or equivalently, a ratio of exp (β_h) in their odds. The solved P_{ij} do not include level one residual because it is the equation for probability of P_{ij} rather than outcome Y_{ij}. The fixed part of the model $\beta_0 + \sum_{h=1}^{k} \beta_h X_{hij}$.

3.2.10 The Random Slope Binary Logistic Regression Model

The multilevel modeling accommodates the hierarchical nature of data and corrects the estimated standard errors to allow clustering of observations within units. The random effect model was used to estimate the degree of correlation in the outcome that existed at the community level (Li, Lingsma, Steyerberg, & Lesaffre, 2011). The intercepts β_{0j} and the slope β_{1j} were community dependant. The community coefficients were split into average coefficient and community dependent deviation where:

$$\beta_{0j} = \beta_0 + U_{0j}$$
$$\beta_{1j} = \beta_1 + U_{1j}$$

Let a single level-1 explanatory variable denoted by P_{ij} for school aged children where:

$$logit\left(P_{ij}\right) = log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \beta_{0j} + \beta_{1j} X_{1ij} \tag{3.19}$$

Now, substituting in equation 3.19 the β_{0j} and β_{1j} will produce two random effects, the random intercept and random slope as:

$$logit\left(P_{ij}\right) = log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \beta_0 + \beta_1 X_{1ij} + U_{0j} + U_{1j} X_{1ij} \tag{3.20}$$

U_{0j} is the random intercept in model 3.20 and U_{1j} is a random slope and are assumed to be level-2 residuals.

From the Equation 3.20 it showed that there were more than one random effects at community level. U_{oj} and U_{1j} were random effects with mean zero, and variance denotes as σ_0^2, σ_1^2 and covariance of σ_{01}^2; Where: β_0 was the average intercept of response variable and β_1 was fixed regression coefficient given explanatory variable X_i. U_0 was random coefficient in the model and $U_0 + U_1 X_{1ij}$ was the random part of the model.

The random effect were correlated in the model and the variances and covariances of level-two random effects of U_{0j}, U_{ij} were denoted by:

$$\begin{aligned} \text{var}\ (U_{0j}) &= \sigma_{00} = \sigma_0^2 \\ \text{var}\ (U_{1j}) &= \sigma_{11} = \sigma_1^2 \\ \text{var}\ (U_{0j}, U_{1j}) &= \sigma_{01}^2 \end{aligned}$$

Now, including more explanatory variables X_1, X_2, X_3 that had random effect and that all the predictor variables had varying slopes and random intercept would yield:

$$\begin{aligned} (\beta_{0j}) &= \beta_0 + U_{0j}, \\ \beta_{1j} = \beta_1 + U_{1j}, ..., \beta_{hj} &= \beta_h + U_{hj} \text{ for h = 1, 2, ..., k,} \end{aligned}$$

then we had:

$\text{Logit}(P_{ij}) = (\beta_0 + U_{0j}) + (\beta_1 + U_{1j}) X_{1ij} + ... + (\beta_h + U_{hj}) X_{hij}$

$= \beta_0 + \Sigma_{h=1}^k \beta_h X_{hij} + U_0 + \Sigma_{h=1}^k U_{hj} X_{hij}$

where, $\beta_0 + \Sigma_{h=1}^k \beta_h X_{hij}$ was fixed part of the model. $U_0 + \Sigma_{h=1}^k U_{hj} X_{hij}$ was the random part of the model and $U_{0j}, U_{1j}, ..., U_{hj}$ were independent between communities and maybe correlated within communities. The components of the vector $U_{0j}, U_{1j}, ..., U_{hj}$ were independently distributed as a multivariate normal distribution with zero mean vector and variances and co-variances Ω given by:

$$\Omega = \begin{bmatrix} \sigma_0^2 & \sigma_{10} & . & . & . & \sigma_{k0} \\ \sigma_0^2 & \sigma_{10} & . & . & . & \sigma_{k1} \\ \sigma_0^2 & \sigma_{10} & \sigma_2^2 & . & . & \sigma_{k2} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \sigma_{0k} & \sigma_{1k} & . & . & . & \sigma_k^2 \end{bmatrix} \tag{3.21}$$

The random slope binary logistic regression model was one of the model to be applied in the analysis but looking at the nature of the hierarchical data, the model was inevitable and not used as it required some approximation to be involved.

3.2.11 Estimation of Coefficients

3.2.11.1 Random effect

In Model 3.17, we had the fixed effects and random effects making it to be mixed-effect model. The random effects part of the model could not be estimated but could be summarized according to their estimated variances and co-variances. The random effect varied across different levels of hierarchy while allowing for correlation with observations at all levels of the model. The model assumed that the community effects were random. It was assumed as:

$$log(P_{ij}) = \log \frac{P_{ij}}{1-P_{ij}} = \beta_0 + \beta_1 X_{ij} + U_j, U_j \sim N(0, \sigma_u^2) \tag{3.22}$$

where: σ_u^2) was the level-2 (community) variance or the between-community variance in the log-odds that $y = 1$ after accounting for x.

Also in the model, the random effects allowed us to examine the role of contextual (community) effects of child anemia. Possible contextual effects were measured by the variance partition coefficient (VPC); this was a variant of intercommunity correlation (ICC) since the outcome was nonlinear. For a dichotomous variable such as presence or absence anemia, VPC was calculated using formula used by (número 442, 2012).

$VPC = V_n/(V_n + \pi^{2/3})$

Vn = community variance VPC represented the percentage of total variance of the effect of anemia in school-aged children attributable to the community level and was also used as a measure of clustering of anemia in communities. A high VPC would reflect a high clustering of anemia effects and a high community effect on individual risk of anemia.

3.2.11.2 Fixed effect

In model 3.17 again, the fixed effects were associated with predictors at any level in the outcome variable. Fixed effect were estimated in the parameter model and were represented as: $\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_k x_{kij}$ which could be estimated directly where;

- β_0 was the log odds that $y = 1$ when $x = 0$ and $u = 0$.
- β_1 was an effect on log-odds of 1-unit increase in x for individuals in same group (same value of u).
- β_1 was often referred to as cluster - specific or unit specific effect of x.
- $\exp\beta_1$ was an odds ratio, comparing odds for individuals spaced 1-unit apart on x but in the same group.

3.2.12 The Variance Components Model

Variance component two-level model was used for a dichotomous outcome variable to check for normality assumptions in level-2 units (communities); and specify the probability distribution for group-dependent probabilities P_j in $Y_{ij} = P_j + \varepsilon_{ij}$ without taking further explanatory variables into account. We focused on the model that specified the transformed probabilities $f(P_j)$ to have a normal distribution. This was expressed, for a general link function $f(P_j)$, by the formula

$$\log \frac{P_{ij}}{1 - P_{ij}} = \beta_{0j} + \upsilon_{0j} \tag{3.23}$$

where β_0 had the community average of the transformed probabilities and υ_{0j} the random deviation from this average for group j. The variance component model could reveal the fixed part of Model 3.18 in the analysis and could check for the existence of community variations. If variations are zero will mean no existence of community variations.

3.2.12.1 Proportional change in variance (PCV)

PCV was calculated with reference to the null model to check for relative contribution of factors to explain variation of anemia in school-aged children.

$$PCV = \left(\frac{\sigma_u^2 - \sigma_{u_1}^2}{\sigma_u^2} \right) * 100 \tag{3.24}$$

where: σ_u^2 was the between community variance in the null model. $\sigma_{u_1}^2$ was the between community variance in Model 3.19 (Merlo et al., 2016).

3.2.12.2 Testing of level-2 (residual) variance

Testing of level two variance or the between-group variance, used log-odds that $y = 1$ after accounting for x. A Wald test using σ_u^2/se would be applied to check for community differences.

3.2.13 Parameter Estimation using the Likelihood Function

The Likelihood function reflects information about the parameters contained in the model. For the two-level logistic Bernoulli responses, the random effect were assumed to be multivariate normal and independent across the community. The marginal likelihood function that was given by:

$$l(\beta, \Omega) = \Pi_i f \Pi_i [(\pi_{ij})^{y_{ij}} (1 - \pi_{ij})^{(1-y_{ij})}] \tag{3.25}$$

where; Ω was variance co-variance matrix.
The likelihood contributed from the *ith* subject and in the *jth* group were as Bernoulli:

$$\text{Bernoulli}(p_{ij}) = p_{ij}^{yij} (1 - p_{ij})^{1-y_{ij}} \tag{3.26}$$

From 3.26; p_{ij} represents the probability of the event for subject i in j community and that the covariate vector were x_{ij} and y_{ij} that indicated having anemia, $(y_{ij} = 1)$ and no anemia $(y_{ij} = 0)$. In multilevel logistic regression we had:

$$p_{ij} = \frac{e^{\beta_0+\beta_1 x_{1ij}+\beta_2 x_{2ij}+\cdots+\beta_k x_{kij}+U_{0j}}}{1+e^{\beta_0+\beta_1 x_{1ij}+\beta_2 x_{2ij}+\cdots+\beta_k x_{kij}+U_{0j}}} \tag{3.27}$$

where $\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_k x_{kij}$ was fixed part of the model and U_{0j} was the random part of the model and $U_0 \sim N(0, \sigma_u^2)$.

In the study the binary logistic regression and multilevel logistic regressions were employed to get estimates of the the fixed effects and their standard errors, variance covariance matrices and their correlation coefficients using data from MDHS 2015 - 16 and MNS 2015 - 16 on school-aged children 5-14 years old. The response variable of the study was anemia. Firstly, ordinary logistic regressions was fitted using the fixed effects characteristics as described in Table 3.2, and their co-variance estimated. We then fitted the mixed effect multilevel logistic model on the co-variates. The random variable was the community areas. The co-variance matrix was then estimated. The estimated results from both logistic regression model and those from multilevel logistic regression were compared as presented in the results section.

CHAPTER 4

RESULTS

The present data set has a two-level hierarchical structure with 800 school-aged children (SAC) 5-14 years nested within 105 communities. In this study we examined effect of anemia in SAC with respect to their demographic, socio-economic and approximate factors. The analysis was carried out in three parts; the first part, we presented effect of anemia in SAC by using descriptive analysis, logistic regression and finally by using multilevel logistic regression model.

4.1. **Bivariate Analysis between Response and Predictor Variables**

The paragraph reports the association between the response variable and each predictor variables. The bivariate analysis, based on Pearson's chi-square statistics provided the relationship between the dependent variable and the independent variables in the study. The high values of chi-square of the independent variable in the study indicated that there was strong association between each of the given independent variables and the dependent variable while keeping effect of other factors constant. In order to make any decision in the bivariate analysis, it was based on chi-square value and p-value of 0.05 significant level.

The descriptive statistic that summarized the association between predictors and response variable has been presented in Table 3. The results from the table showed higher percentage of anemia in school-aged children of 5-10 years (23%) while for school-aged children of 11-14 years (12%). The effect of anemia among SAC 5-14 years also differed by the type of place of residence. Accordingly, higher numbers of anemic children (21%) resided in rural areas and small number of anemic school-aged children (9%) resided in urban areas.

From Table 3 it also showed that there was a bivariate association between anemia status in school-aged children 5-14 years and other independent variables that anemia was strongly associated with; Inflammation (mild 28% and severe 38%) respectively, with p-value at 0.05. Likewise for wealth quintile; We observed that there was a statistically significant association between level of anemia and school-aged children from poorest households (P-value = 0.015) and having higher percentage of being anemic from poorest families (25%) while least anemic were from richest household (10%). Other independent variables like: Malaria in the last two weeks, if received zinc and if either mother is alive or father is alive, sex of the head of household and finally if the child received vitamin A, were not significant. The wealth index in this study is a composite measure of cumulative living standards of a household used as a proxy for social economic status.

The drawback of using descriptive statistics approach is that it ignores the possibility that a collection of variables could be weakly associated with the outcome Table 4.1.

Table 3: Cross Tabulation of Anemia Status versus Explanatory Variables

Variable	Anemia Status		Total	DF	CHI	P-Value
	Anemic n (%)	Not Anemic n (%)				
Total	154(19)	649(81)	800			
Residence						
Urban	9(9)	87(91)	96	1	6.84	0.009
Rural	145(21)	559(79)	704			
Age of a Child (Years)						
5 to 10	123(23)	411(77)	534	1	14.8	0.001
11 to 14	31(12)	235(88)	266			
Sex of a Child						
Girl	84(55)	328(51)	412	1	0.65	0.42
Boy	70(45)	316(49)	386			
Education of a Child						
Class 0-4	148(96)	565(88)	713	1	9.15	0.002
Class 5-8	6(4)	79(12)	85			
Water Source						
Improved	119(77)	540(84)	659	1	3.74	0.053
Unimproved	35(23)	4(16)	39			
Malaria last 2 weeks						
Yes	18(28)	46(72)	64	1	3.56	0.059
No	134(18)	593(82)	727			
Inflammation						
Moderate	63(13)	432(87)	495			
Mild	43(28)	108(72)	151	2	45.3	0.001
Severe	42(38)	70(63)	112			
Wealth Index						
Poorest	38(25)	115(75)	153			
Poor	34(23)	117(77)	151			
Middle	35(18)	158(82)	193	4	12.4	0.015
Richer	34(20)	134(80)	168			
Richest	13(10)	122(90)	135			
If received ZINC						
Yes	149(19)	616(81)	765	1	0.591	0.442
No	5(14)	30(86)	35			
Mother Alive						
Yes	148(19)	610(80)	758	1	0.6	0.439
No	6(15)	35(85)	41			
Father Alive						
Yes	145(20)	593(80)	738	2	2.532	0.282
No	8(14)	51(86)	59			
Head Sex						
Male	111(19)	465(81)	576	1	0.964	0.001
Female	43(19)	181(81)	224			
IF received VitA						
Yes	5(33)	10(67)	15	1	2.04	0.15
No	137(19)	595(81)	732			

4.2. **Results of Multilevel Models**

Multilevel analysis allowed for more accurate estimation of regression coefficients and standard errors due to non-independence and quantification of between-cluster variation (ICC). Variables in this case were measured at different levels of hierarchy. It also allowed for correct inferences about community-level variables to be made. Additionally, the magnitude of the association between variables and outcome varied between communities which sometimes could not be handled by traditional regression techniques.

4.2.1 Variance Component Model

The variance component model predicted the probability of anemia status in school-aged children. The results from Table 4 revealed the information of the fixed effect that estimated the log-odds of anemia among children aged 5-14 years in Malawi with $\beta_0 = -1.631$. The β_0 expressed the overall proportion $\left(\frac{e^{-}\beta}{1+e^{-}\beta}\right) = 0.164$ of effect of anemia in SAC between 5-14 years in Malawi without accounting for other sources of variation. The between community variance U_0 using log odds of being anemic was estimated as $\sigma_u^2 = 0.593$. which showed that there was non-zero community variation and that the community variations contributed to effect of anemia in school-aged children 5-14 years. The intracommunity correlation coefficient (ICC) in the variance component model, (Table 4) = 0.153, meant that 15 percent of the total variability in the effect on anemia in SAC 5-14 years old was significantly related to community level, whereas the remaining 85 percent was related to within community difference. The systematic differences from ICC gave concept of applying multilevel analysis in the data.

Table 4: Estimates for variance component model

Fixed effects	**Estimate**	**S.error**	**Z-value**	**P-Value**
β_0	-1.631	0.140	-11.650	0.000
Log Likelihood	-382.850			
Random effect				
σ_u^2	0.593	0.232	2.559	0.005
ICC	0.153	0.051		
Chi-Square	17.18			0.001

4.2.2 Estimate of Cross Level Interaction

The Cross level interaction in Table 5 used estimates in the covariance matrix to detect the covariance between the slope and intercept. The result of the covariance matrix in Table 5 showed that there was a negative relationship of $U_1, U_2 = -0.0066$. These results showed the existence of clustering in the data set implying that multilevel modelling was to be used in the analysis. Likewise, for the residual estimate of level-1 and the intercept estimate of level 2 were also greater than 0 showing the existence of the variations in communities.

Table 5: Cross level Interaction

		Estimate	**Std. Error**	**Z-value**	**P-value**
Residuals	Level 1	0.1129	0.0690	1.64	0.000
Intercept	Level 2	0.1241	0.0049	25.08	0.000
Slope	Level 2	0.1572	0.0066	23.70	0.000
Cov(I,S)	Level 2	-0.0066	0.0048	-1.38	0.000

4.2.3 Multicollinearity in Data Set

Using the variance Inflation factor (VIF) where

$$\text{VIF} = < 1\text{: no multicollinearity}$$
$$\text{VIF = between 1 to 5: moderate multicollinearity}$$
$$\text{VIF} = > 5\text{: High multicollinearity}$$

From Table 6, the $\frac{1}{VIF}$ for each independent variable showed that multicollinearity was not existing since all variables were less than 1 i.e. were falling beteween 0 and 1 meaning there was no association among independent variables of $\beta_1, \beta_2, \beta_3, \cdot, \beta_k$. All variables were found to be less than one and were to be included in modeling effects on anemia in school-aged children in Malawi.

Table 6: Checking for Existence of multicollinearity in Data Set

Variable	**VIF**	**1/VIF**
Weight of Child	1.09	0.920832
WealthIndex		
Poorer	1.62	0.616067
Middle	1.73	0.57949
Richer	1.67	0.598543
Richest	1.59	0.628896
Inflammention		
Mild	1.09	0.921043
Severe	1.09	0.915546
Mean VIF	**1.41**	

4.2.4 Logistic Regression verses Multilevel Logistic Regression

Table 7 provides the odds ratio and confidence intervals for the logistic regression analysis and multilevel logistic regression analysis of effect on anemia in school aged children 5-14 years. The estimates were provided side by side to facilitate comparisons between single-level logistic regression and multilevel logistic regression. The vari-

ables were presented according to hierarchical rationale starting with community related variables, control variables, health related variables, then household variables and the outcome variable was anemia.

From Table 7 where, logistic regression produced almost similar results to multilevel logistic regression. These results were similar to other studies conducted by (Kawo, Asfaw, & Yohannes, 2018) on applying multilevel analysis of determinants of anemia prevalence among children aged 6-59 months in Ethiopia where Classical and Bayesian approaches were applied. The estimates of the odds ratio are almost similar for all variables with very minor differences. The multilevel analysis had made it possible to quantify the contribution of community-level effects on anemia in school-aged children.

Community-level variables: Since the results of logistic regression and multilevel logistic regression are almost similar, the odds of school-aged children 5-14 years living in rural areas were more associated with effect on anemia (OR = 1.70, 95% confidence interval [CI]: 0.77-3.72) in logistic regression analysis and (OR = 1.78, CI: 0.69 - 4.61) using the multilevel logistic regression analysis.

Control Variables: Likewise, the results from control variables in Table 7 from logistic regression analysis and multilevel logistics regression analysis gave almost similar results. School-aged children 11-14 years had odds estimate of (0.54, CI: 0.34 - 0.84) in logistic regression analysis while in multilevel logistic regression was estimated at (OR = 0.55, CI: 0.34 - 0.89) meaning that effect of anemia in both logistic regression analysis and multilevel logistic regression analysis was less in SAC 11-14 years.

Health Related variables: Checking for variables of health related factors in Table 7, The effect on anemia in school-aged children was two times and four times higher for children with mild to severe inflammation with the (OR = 2.44 to 4.25, CI: 1.49 - 3.99 to CI: 2.48 - 7.28) in multilevel logistic regression analysis. This gives a slight difference with the results of logistic regression analysis where the effect on anemia in school-aged children with mild inflammation had (OR = 2.35, CI: 1.45 - 3.69) and for those children with severe inflammation had (OR = 3.63, CI: 2.25 - 5.88). For other health related variables like if the child received vitamin A, if the child had malaria in

the past two weeks and source of drinking water had both similar results with very small variations in both logistic regression and multilevel logistic regression analysis.

Household Related variables: For school-aged children 5-14 years who were living in richest families were less associated with anemia in multilevel logistic regression analysis likewise in logistic regression analysis as compared to those living from poor families, middle families and richer families with (OR = 0.36; 95%CI = 0.15 - 0.87) verses (OR = 0.79, CI: 0.42 - 1.49) in poorest families, (OR = 0.71, CI: 0.34 - 1.35) in middle families, and (OR = 0.92, CI: 0.47- 1.77) in richer families (Table 7).

Table 7: Logistic Regression verses Multilevel Logistic Analysis

Variable	n	Logistic regression		Multilevel Logistic	
		OR	95%CI	OR	95%CI
Community Related Variables					
Residence					
Urban (ref)	96	-			
Rural	704	1.695	0.774 - 3.717	1.781	0.688 - 4.608
Control variables					
Age of child					
5 to 10 (ref)	534	-			
11 to 14	266	0.536	0.343 - 0.838	0.552	0.344 - 0.885
Sex of the child					
Girl (ref)	412	-			
Boy	386	0.936	0.643 - 1.363	0.891	0.594 - 1.338
Head sex					
Male (ref)	576	-			
Female	224	0.876	0.567 - 1.354	0.944	0.573 - 1.556
Health-related variables					
Water Source					
Improved (ref)	659	-			
Un improved	39	1.083	0.677 - 1.731	1.042	0.584 - 1.857
Malaria					
Yes (ref)	64	-			
No	727	0.662	0.358 - 1.224	0.644	0.330 - 1.260
Inflammation					
None (ref)	495	-			
Mild	151	2.351	1.450- 3.687	2.438	1.491 - 3.987
Severe	112	3.637	2.249 - 5.883	4.250	2.479 - 7.284
VitA					
Yes (ref)	15	-			
No	732	0.512	0.157 - 1.676	0.481	0.132 - 1.754
Household related Variables					
Wealth Index					
Poor (ref)	151	-			
Poorer	153	0.784	0.445 - 1.384	0.791	0.421 - 1.488
Middle	193	0.695	0.398 - 1.212	0.708	0.373 - 1.346
Richer	168	0.802	0.455 - 1.414	0.921	0.480 - 1.767
Richest	135	0.406	0.189 - 0.874	0.361	0.151 - 0.867
Mother Alive					
No (ref)	758	-			
Yes	41	1.191	0.464 - 3.058	1.314	0.473 - 3.651
Father Alive					
No (ref)	738	-			
Yes	59	1.477	0.645 - 3.384	1.970	0.790 - 4.911
Don't Know	59	9.581	0.448 - 187.940	15.285	0.600 - 389.100
Log Likelihood	-355.63			-348.65	

4.2.5 Random Intercept Model

Table 8 identified the random intercept model which was the multilevel model with fixed effects and random effects. The analysis revealed that correlates of anemia varied among communities or districts. The deviance based chi-square test for the random effects in random intercept model was χ^2= 13.95 (d.f. = 18, p-value = 0.0001). This indicated that the random intercept model gave a better fit as compared to various component model in Table 4 with $\chi^2 = 17.18$ and p-value of 0.0001. The fixed part of the table showed variables like; wealth index, age of the child and if experience inflammation which were statistically significant on effects of anemia in SAC 5-14 years. From the results in Table 8, controlling for community differences in the effect of anemia in SAC would result in the odds of decreasing by a factor of $e^{-}0.595$ = 0.552 for each year increase in age group of 11-14 years.

The intraclass correlation coefficient (ICC) is a measure of variation of effect of anemia in SAC among communities. The ICC of 0.153 about 15% of the variation in effect of anemia in school-aged children is attributable to variation within communities and only about 85% variations were due to level two effects or between communities. The intraclass correlation coefficient is statistically significant at 5 percent level of significance. The random intercept in (Table 8) with chi-square probability of 0.0001 indicates that correlates of anemia in SAC differs from community to community taking into account all covariates measured.

The log odds of residing in rural areas, having no malaria in the past week, if mother was alive, if father was alive and if having mild or severe inflammation were having more than once chance in contributing to effect of anemia in SAC 5-14 years without affecting the random effects and covariates unchanged or without affecting community or district variations. The variance component of the random intercept is significant at 0.0001 suggesting that there remains some variation in the effect of anemia in school-aged children which are not accounted for by the variables in the model. These estimates can be justified by estimating an alternative model that contains random coefficient model.

Table 8: Estimate of Random Intercept Model

Variable	Coef.	Std.Er.	Z-value	P-value	OR	[95% Conf. Upper	Interval] Lower
Fixed effect							
Residence							
Urban (ref)							
Rural	0.577	0.485	1.19	0.234	1.781	-0.373	1.528
Malaria							
Yes (ref)							
No	-0.439	0.342	-1.28	0.199	1.552	-1.110	0.231
ChildSex							
Girl (ref)							
Boy	-0.115	0.207	-0.56	0.578	0.644	-0.521	0.291
Inflammation							
Moderate							
Mild	0.891	0.251	3.55	0.001	2.438	0.399	1.383
Severe	1.447	0.275	5.26	0.001	4.250	0.908	1.986
Wealth Index							
Poor							
Poorer	-0.234	0.322	-0.73	0.468	0.791	-0.866	0.398
Middle	-0.345	0.328	-1.05	0.292	0.708	-0.987	0.297
Richer	-0.082	0.332	-0.25	0.805	0.921	-0.733	0.569
Richest	-1.018	0.446	-2.28	0.023	0.361	-1.893	-0.143
HeadSex							
Male (ref)							
Female	-0.058	0.255	-0.23	0.821	0.944	-0.557	0.442
Mother Alive							
No (ref)							
Yes	0.273	0.522	0.52	0.601	1.314	-0.749	1.295
Father Alive							
No (ref)							
Yes	0.678	0.466	1.45	0.146	1.970	-0.235	1.592
Don't Know	2.727	1.652	1.65	0.099	15.286	-0.510	5.964
Age of Child							
5-10(ref)							
11-14	0.595	0.241	-2.46	0.014	0.554	-1.068	-0.122
Vitamin A							
Yes (ref)							
No	0.732	0.660	-1.11	0.267	0.481	-2.027	0.562
Water source							
Improved (ref)							
Unimproved	0.041	0.295	1.14	0.890	1.153	-0.537	0.619
β_0	-1.880	1.123	-1.67	0.094		-4.082	0.321
Random effect							
$\sigma_0^2 = var(U_0 j)$	0.595	0.249				0.262	1.351
ICC(ρ)	0.153						

4.3. **Estimates of the Random Coefficients Model**

We generated a model so that the effect of level-1 covariates differed in each community. This was done by adding the random coefficient to individual-level covariates of the model from Table 8. In the random intercept model we allowed the intercept only to vary across communities or districts by fixing explanatory covariates (Table 9). The relationship between explanatory and dependent variables differed between communities in this study in many ways; School-aged children were nested in communities or districts which allowed one to estimate odds as a function of lower-level variables like (e.g. child age) while higher level variable was taken as community which resulted to cross level interactions. These results were similar with a study conducted by (Sommet and Morselli (2017) on a study where pupils were nested in different classrooms.

A model that depended on community through individual characteristics was fitted to allow the probability of effect on anemia in school-aged children 5-14 years old. this was done by allowing the model intercept to vary randomly across communities in the random intercept model. We assumed, however, that the effects of individual characteristics such as; place of residence and water source were the same in each community meaning that the coefficient of all explanatory variables were fixed across communities and treated as random. The random coefficient model then allowed both the intercept and coefficients to vary randomly across communities or districts (Table 9).

The random coefficient model estimated the intercepts and varying significantly at 5 percent significant level. This meant that there was considerable variation in the covariates and the variables differed significantly across communities. The community variance component for the variance of intercept in the random coefficient model of 0.072 (Table 9) and district variance of 0.721 (Appendix A: Table 13) which were larger relative to their standard errors of 6.831 and 6.830 respectively, explained the community-level variances which were unaccounted for in the model (Table 9).

Effect on anemia was higher using estimate of random coefficient in SAC 5-14 years with mild to severe inflammation and those children with lower age group of 5-10 years. School-aged children (SAC) of 5-10 years were one times higher on anemia effects as compared to SAC of 11-14 years (OR = 0.933, 95%CI = -0.125 to -0.013). Children

with mild inflammation (OR = 1.132; 95%CI = 0.055 - 0.193) and severe inflammation (OR = 1.253, 95%CI = 0.147- 0.304) were on higher risks of effects on anemia. The intra-community correlation coefficient of 0.073 (7.3 percent) showed variations in communities that were due to random factors of level-two which are still unexplained, while 92.7% were due to fixed effects. This meant that parameters had their own estimates at each community or district (Table 9).

Table 9: Estimate of Random Coefficient Model

Variable	Coef.	Std.Er.	Z-value	P-value	OR	[95% Conf. Upper	Interval] Lower
Fixed effect							
Residence							
Urban (ref)							
Rural	0.055	0.054	1.02	0.310	1.056	-0.051	0.160
Malaria							
Yes (ref)							
No	-0.061	0.050	-1.23	0.217	0.941	-0.158	0.036
ChildSex							
Girl (ref)							
Boy	-0.016	0.027	-0.58	0.563	0.985	-0.068	0.037
Inflammation							
Moderate							
Mild	0.124	0.035	3.53	0.001	1.132	0.055	0.193
Severe	0.225	0.040	5.65	0.001	1.253	0.147	0.304
Wealth Index							
Poor							
Poorer	-0.032	0.045	-0.70	0.482	0.969	-0.120	0.057
Middle	-0.049	0.044	-1.11	0.266	0.952	-0.136	0.038
Richer	-0.016	0.046	-0.34	0.732	0.984	-0.105	0.074
Richest	-1.105	0.054	-1.96	0.050	1.111	-0.210	-0.0001
HeadSex							
Male (ref)							
Female	0.007	0.032	-0.21	0.837	0.993	-0.070	0.057
Mother Alive							
No (ref)							
Yes	0.033	0.062	0.54	0.589	1.034	-0.088	0.154
Father Alive							
No (ref)							
Yes	0.080	0.053	1.50	0.133	1.083	-0.024	0.184
Don't Know	0.419	0.273	1.53	0.125	1.520	-0.116	0.953
Age of Child							
5-10(ref)							
11-14	-0.069	0.029	-2.40	0.016	0.933	-0.125	-0.013
Vitamin A							
Yes (ref)							
No	-0.120	0.099	-1.21	0.226	0.887	-0.314	0.074
Water source							
Improved (ref)							
Unimproved	0.012	0.040	1.29	0.773	1.012	-0.067	0.090
β_0	0.225	0.148	1.52	0.129		-0.066	0.515
Random effect							
$\sigma_0^2 = var(U_0 j)$	0.102	0.019				0.071	0.146
$\sigma_{12}^2 = cov(U_1 j, U_2 j)$	0.363	0.010				0.344	0.382
Community variance	0.072	6.831				1.83e	2.84e
District variance	0.721	6.830				1.87e	2.79e
ICC(ρ)	0.073						

4.4. **Model Comparison**

In order to analyze data which is in the hierarchical form requires the technique of the best model to choose that can explain the variability of the data. From Table 10, based on the chi-square; The variance component chi-square ($\chi^2 = 17.18$), random intercept ($\chi^2 = 13.95$) and random coefficient chi-square ($\chi^2 = 14.15$) with p-value of 0.001 showed that all the models were significant allowing multilevel models to be applied in analyzing of hierarchical data that helped to check the impact of the within community variation and between community variation in SAC 5-14 years old.

From Table 10, it has shown that the random intercept and the random coefficient models were better models for the data as compared to variance component model with less AIC of 730.33 and 733.31 as compared to variance component model AIC = 769.69. Variations in SAC were accounted in all the models described in the study.

Table 10: Model Comparison

Fitted Model	**Variance Component**	**Random Intercept**	**Rondom Coefficient**
-2*log likelihood	-382.85	-348.66	-346.27
Chi-square	17.18	13.95	14.15
Degree of freedom	1	1	1
P-value	0.001	0.001	0.001
AIC	769.69	733.31	730.53
BIC	779.06	817.57	819.47
ICC	0.150	0.153	0.073

CHAPTER 5

DISCUSSION & CONCLUSION

5.1. **Discussion**

The aim of this study was to model effects on anemia in school-aged children 5-14 years old in Malawi by applying multilevel logistic regression analysis. Anemia in the study was determined by individual-level, community-level and district-level factors. All these were supported by the observed heterogeneity in odds on anemia between communities and between districts. The variables included in the study for individual-level and community-level analysis were place of residence (rural, urban), malaria in the past two weeks, child sex, wealth index (poor, poorer, middle, richer, richest), head sex, if mother is alive , if father is alive, if received vitamin A, inflammation (moderate, mild, severe) and type of water source (improved and unimproved). Variables; Wealth index (richest), inflammation (mild and severe) and age of child were found to be significant in both individual-level and community level models. These results were in agreement with findings from several studies conducted in Ethiopia and the world on under-five children by (Sommet & Morselli, 2017) and (Tezera, Sahile, Yilma, Misganaw, & Mulu, 2018).

A bivariate analysis between response variable and each of the predictor variable was fitted into the data to check for association between response variable and predictor variables, while a multilevel model was employed to estimate the quantification of between-community variation (ICC). Since variables were measured at different levels of the hierarchy, it allowed for correct inferences about community-level variables to be made and at the same time checking for the magnitude of the association between variables and outcome that varied between communities, which was something that could not easily be handled by simple logistic regression.

The random intercept was fitted to allow intercepts to vary while retaining common slopes across communities, and random coefficient model was fitted to allow both the intercepts and slopes to vary across communities. Lastly, model diagnostic was tested to check for the better fit model by either using BIC or AIC.

The 2015-16 Malawi Demographic Health Survey (MDHS) and Micronutrient Survey (MNS) was based on two-stage stratified sampling. The first stage was selecting a sample frame from the Malawi Population and Housing Census (PHC). The second stage was selection of households from sampled communities from a sample frame. Based on the result of this study, less proportion on anemia cases were observed among school-aged children with moderate and severe inflammation. As a result school-aged children who had mild and severe anemia were more than 100% more likely to experience anemia as compared to school-aged children with moderate inflammation. These results are linked to a study conducted by (Syed et al., 2016) on determinants of anemia among school-aged children in Mexico, the United States and Colombia.

The logistic regression analysis and multilevel logistic regression analysis in this study showed that age had a significant effect on anemia for the SAC 5-14 years at 5% level of significance. The odds of SAC 11-14 years were 0.6 times less likely to develop anemia as compared to SAC 5-10 years old. The odds of multilevel regression analysis considering heterogeneity for SAC who resided in urban areas were 2 times less likely to become anemic as compared to school-aged children residing from rural areas. Likewise, the odds for school-aged children who experienced no malaria in past two weeks decreased by 34% as compared to those who experienced malaria in the past two weeks. Applying for multilevel regression analysis in the data, we found that there were community-level clustering and district-level clustering of ICC = 15% and ICC = 12% respectively (Table 4). With these results therefore, anemia interventions need not to be community-specific, otherwise effects on anemia may decrease in one area living other areas unattended. Identifying a successful program would include a focus on overall, not focusing only on specific groups (individuals) but covering variations between communities or districts.

The results from this study will not be compared more directly with previous studies because this study is one of the few anemia studies conducted in Malawi and one of the first study to look into effects on anemia in school-aged children of 5-14 years using the multilevel modelling. The previous studies interpreted of high prevalence of anemia in an area/community while this study separated the contribution of individual characteristics to effect on anemia from the contribution of the community which is the real clustering effects. A similar study conducted in Mixico and Colombia on determinants of anemia in SAC by (Syed et al., 2016) showed similar results of evidence of clustering but at household level.

With a newly available hierarchical regression techniques, it is possible to separate these individual effects from contextual effects and therefore, to give accurate measure between community or district variation (Vugutsa Luvai et al. (2017)). More multilevel studies are needed to determine whether relatively clustering effects targeting the age group of school-aged children 5-14 years applies to other countries. Fewer data are available on anemia in growing children of age 5-14 years (Mehta, Sachdeva, & Tripathi, 2021).

In order to look for a model that would fit the data to be used in the study the variance component model was applied using level-2 model to check if grouping variable at level-2 significantly affected the intercept of the dependent variable at level-1. A significant ICC from the result proved that the level-2 clustering was significant with non zero ICC and therefore multilevel modeling was applied.

It is important to note that this was the first study to use multilevel logistic regression to estimate the community variation and compare the variance component model, random intercept model and random coefficient model that would better fit the hierarchical data to predict the effect on anemia in SAC 5-14 years. From the results the better fit model to the hierarchical anemia data were both the random coefficient model and random effect model since their variation did not differ much with less AIC of 730.53 and 733.31 as compared variance component model with 769.69. Otherwise if we choose among the two, then we may conclude that the better model using AIC is the random coefficient model AIC = 730.53. (Table 10).

More studies are therefore needed to shed further light on the usefulness of using multilevel analysis on anemia in SAC 5-14 years in identifying between community effects.

5.2. **Strength and Limitation**

The data used in this study was obtained from a study that was conducted at a national level with huge sample size. The study used weighted data and proper statistical analysis was applied considering the hierarchical nature of the MDHS and MNS data. The unobserved heterogeneity of this study was accounted by using the random parameter approaches.

The limitations were that the MDHS and MNS were based on respondents self-report that might have the possibility of recall bias. The cross sectional nature of the study of MDHS and MNS data would not permit to investigate on the cause and effect relationships to be established between explanatory variables and anemia like eating habits, parasite infections, urinary schistosomiasis, urinary iodine and red blood cell folate deficiency that might have made it difficult to identify independent effects of the considered variables. Modeling community effects on anemia in school-aged children using multilevel logistic regression modeling was the first study in Malawi which brought a challenge to produce trends on anemia. Therefore, a prospective study should be conducted to address such effects and also to focus on more specific and relevant variables that would be able to generate more useful information.

5.3. **Conclusion**

This study revealed that effect on anemia among SAC 5-14 years in Malawi was still a public health problem where 19% of the SAC were found to be anemic. Therefore, policy makers and stakeholders should pay attention to all significant factors mentioned in the analysis of the study. Inflammation (mild and severe), age of the child, wealth index, education of the child, drinking water source and head sex. Community or District variation should be the main issue of consideration when intervention are to take place.

On methodological aspects, the multilevel logistic regression analysis was the right model to be applied to hierarchical data used in the analysis as compared to single level logistic regression analysis. The multilevel regression analysis in the study could take

into account the quantification of the magnitude between community variations. The multilevel analysis was also used in the study to handle multiple levels of clustering.

In conclusion it is better to design appropriate national strategy for prevention on anemia in SAC by considering the community characteristics. Multilevel logistic regression examined the effects of the explanatory variables at different levels simultaneously. It produced more accurate estimates of regression coefficients, standard errors, and significant test as compared with single-level logistic regression. The effects on anemia in SAC were found by factors evolving at community-levels (level-2). There was a weak negative correlations between individual and community variations.

The results also showed that ignoring the hierarchical nature of the data could result in over estimating the significance of some of the variables included in the model. Using the standard logistic regression would not explain any variation of within and between cluster in the analysis. The only appropriate approach to analyze anemia data from this study was therefore, based on nested sources of variability, where units at lower-level (level-1), i.e. school-aged children 5-14 years who are nested within units at higher-level (level-2) community-level. The hierarchical nature of the data used in the study allows that multilevel modeling should be a natural choice to use in this analysis. The multilevel random coefficient was better as compared to random intercept and the variance component in fitting the data.

5.4. **Recommendation**

As there was variation and differences in effect on anemia in SAC 5-14 years across communities and districts, it is recommended that interventions should cover all the communities and districts. It is also recommended that multilevel models are appropriate methods that investigates effects on anemia in SAC when data is hierarchical (clustered data) just because it takes into account its variations among communities and districts as opposed to ordinary regression methods.
Therefore, this study is important in that: It firstly provide an alternative model that can preferably represent the current data set to model the effect on anemia in SAC. Secondly it provides information about variations across communities and districts and finally it points that further studies should be conducted to incorporate spatial variations

on effect on anemia in SAC by utilization of other models such as Spatial Models and Geo-additive models to investigate spatial variations of effect on anemia in SAC in communities and districts.

REFERENCES

Austin, P. C., & Merlo, J. (2017). Intermediate and advanced topics in multilevel logistic regression analysis. *Statistics in Medicine*, *36*(20), 3257–3277. doi: 10 .1002/sim.7336

Barth, J. H., Luvai, A., Jassam, N., Mbagaya, W., Kilpatrick, E. S., Narayanan, D., & Spoors, S. (2018). Comparison of method-related reference intervals for thyroid hormones : studies from a prospective reference population and a literature review. , *55*(1), 107–112. doi: 10.1177/0004563217691549

Bates, M. S., I., & Sarkinfada, F. (2007). *Anemia: A useful indicator of neglected disease burden and control*. PLos Medicine. doi: org/10.1371/

Berhie, K. A., & Gebresilassie, H. G. (2016). Logistic regression analysis on the determinants of stillbirth in Ethiopia. *Maternal Health, Neonatology and Perinatology*, *2*(1), 1–10. Retrieved from `http://dx.doi.org/10.1186/s40748-016-0038-5` doi: 10.1186/s40748-016-0038-5

Calis, J. C., Phiri, K. S., Faragher, E. B., Brabin, B. J., Bates, I., Cuevas, L. E., ... van Hensbroek, M. B. (2016, sep). Severe anemia in Malawian children. *Malawi medical journal : the journal of Medical Association of Malawi*, *28*(3), 99–107.

Curini, L., Franzese, R., & Steenbergen, M. (2020). Multilevel Analysis. *The SAGE Handbook of Research Methods in Political Science and International Relations*, 679–700. doi: 10.4135/9781526486387.n39

Custodio, E., Descalzo, M., Roche, A., Sanchez, J. I., Molina, L., Lwanga, M., ... Baylin, A. (2008). Nutritional status and its correlates in equatorial guinean preschool children. (29). (doi:10.1177/156482650802900106.)

Dayton, C. M. (1992). *Logistic regression analysis.* Retrieved from `https://www.researchgate.net/publication/268416984`

De Leeuw, J., & Meijer, E. (2008). *Handbook of multilevel analysis* (No. February 2015). doi: 10.1007/978-0-387-73186-5

Dey, S., & Raheem, E. (2016). Multilevel multinomial logistic regression model for identifying factors associated with anemia in children 6–59 months in north-eastern states of India. *Cogent Mathematics*, *3*(1), 1159798. doi: 10.1080/23311835.2016.1159798

Franke, T. M., Ho, T., & Christie, C. A. (2012). The Chi-Square Test: Often Used and More Often Misinterpreted. *American Journal of Evaluation*, *33*(3), 448–458. doi: 10.1177/1098214011426594

Gebremeskel, M. G., Mulugeta, A., Bekele, A., Lemma, L., Gebremichael, M., Gebremedhin, H., ... Shushay, S. (2020). Individual and community level factors associated with anemia among children 6—59 months of age in Ethiopia: A further analysis of 2016 Ethiopia demographic and health survey. *PLoS ONE*, *15*(11 November), 1–17. Retrieved from `http://dx.doi.org/10.1371/journal.pone.0241720` doi: 10.1371/journal.pone.0241720

Guo, G., & Zhao, H. (2000). *Multilevel Modeling for Binary Data* (Vol. 26; Tech. Rep.).

Gutema, B., Adissu, W., Asress, Y., & Gedefaw, L. (2014). Anemia and associated factors among school-age children in Filtu Town, Somali region, Southeast Ethiopia. *BMC Hematology*, *14*(1), 1–6. doi: 10.1186/2052-1839-14-13

Hossain, M. G., Kamruzzaman, & Wadood, A. (2018). Two-Level Logistic Regression Analysis of Factors Influencing Anemia Among Nonpregnant Married Women of Reproductive Age in Bangladesh. , 11–19. doi: 10.1007/978-981-10-6104-2_2

Howie. (2008). Lecture 7 Logistic Regression with Random Intercept. , 1–48. Retrieved from `papers2://publication/uuid/63E4924E-1AA4-446E-8E9F-3564F0CDB6AF`

Hox, J. (1998). Multilevel Modeling: When and Why. (1965), 147–154. doi: 10.1007/978-3-642-72087-1_17

Janjua, N. Z., Khan, M. I., & Clemens, J. D. (2006, 12). Estimates of intraclass correlation coefficient and design effect for surveys and cluster randomized trials on injection use in pakistan and developing countries. *Tropical Medicine and International Health*, *11*, 1832-1840. doi: 10.1111/j.1365-3156.2006.01736.x

Kawo, K. N., Asfaw, Z. G., & Yohannes, N. (2018). Multilevel Analysis of Determinants of Anemia Prevalence among Children Aged 6-59 Months in Ethiopia: Classical and Bayesian Approaches. *Anemia*, *2018*(June). doi: 10.1155/2018/3087354

Kazembe, L. N., & Ngwira, A. (2015). Analysis of severity of childhood anemia in Malawi: A Bayesian ordered caegry modelsanzanian trauma patients' prehospital experience:a qualitative interview-based study. *BMJ Open 2015*, *5*(4). (doi:10.1136/bmjopen-2014-006921)

Kumar Chowdhury, S., Ahmed, S., Ara Hossain, I., Yasmin, R., Faruquee, M., & Salim, A. (2019). Status of Iron Deficiency Anemia among Bangladeshi Chil-

dren: Urban and Rural Settings. *Acta Scientific Paediatrics*, *2*(8), 08–12. doi: 10.31080/aspe.2019.02.0103

Li, B., Lingsma, H. F., Steyerberg, E. W., & Lesaffre, E. (2011). Logistic random effects regression models: A comparison of statistical packages for binary and ordinal outcomes. *BMC Medical Research Methodology*, *11*, 1–11. doi: 10.1186/1471-2288-11-77

Liang, Y., & Carriere, K. C. (2013). A comparison of statistical methods for analyzing discrete hierarchical data: A case study of family data on alcohol abuse. *Open Journal of Statistics*, *03*, 1-6. doi: 10.4236/ojs.2013.34a001

Lipi, N., Alam, M. S., & Hossain, S. S. (2021, 7). A generalized estimating equations approach for modeling spatially clustered data. *Austrian Journal of Statistics*, *50*, 36-52. doi: 10.17713/ajs.v50i4.1097

Mehta, G., Sachdeva, M., & Tripathi, R. (2021). Prevalence of Anemia in Children of Rural Population of Northern State of India. *Ars Pharmaceutica (Internet)*, *62*(2), 182–189. doi: 10.30827/ars.v62i2.17762

Merlo, J., Wagner, P., Ghith, N., & Leckie, G. (2016). An original stepwise multilevel logistic regression analysis of discriminatory accuracy: The case of neighbourhoods and health. *PLoS ONE*, *11*(4), 1–31. Retrieved from `http://dx.doi.org/10.1371/journal.pone.0153778` doi: 10.1371/journal.pone.0153778

Messelu, Y., & Trueha, K. (2016). Application of Multilevel Binary Logistic Regressions Analysis in Determining Risk Factors of Diarrheal Morbidity among under Five Children in Ethiopia. *Public Health Research 2016*, *6*(4), 110–118. doi: 10.5923/j.phr.20160604.03

National Statistical Office. (2015). Malawi Demographic and Health Survey 2015-16. *National Statistics Office The DHS Program*, 1–658. Retrieved from `http://dhsprogram.com/pubs/pdf/FR319/FR319.pdf`

Ngnie-Teta, I., Receveur, O., & Kuate-Defo, B. (2007). Risk factors for moderate to severe anemia among children in Benin and Mali: Insights from a multilevel analysis. *Food and Nutrition Bulletin*, *28*(1), 76–89. doi: 10.1177/156482650702800109

Ngwira, A., & Kazembe, L. (2016). Analysis of severity of childhood anemia in Malawi: a Bayesian ordered categories model. *Open Access Medical Statistics*, *6*, 9. Retrieved from `https://www.dovepress.com/`

analysis-of-severity-of-childhood-anemia-in-malawi-a-bayesian-ordered--peer-reviewed-article-OAMS doi: 10.2147/OAMS.S95159

Ntenda Morton., F. N. T., Kun-Yang Chung., & chih Chuang, Y. (2017). Multilevel Analysis of the Effects of Individual and Community - Level Factors on Childhood Anemia. *Oxford University Press*, *12*. (doi:10.1093/tropej/fmx059)

número 442, A. (2012). No Title . , 32.

Prieto-Patron, A., der Horst Zsuzsa V. Hatton, K. V., & Detzel, P. (2018). Association Between Anemia in Children 6 to 23 Months Old and Child Mother Household and Feeding Indicators. *Algorithmic Operations Research*, *10*(1269). Retrieved from www.preprints.org.Nutrients

Reviews, A. (2016). Multilevel Modeling for Binary Data Author (s): Guang Guo and Hongxin Zhao Published by : Annual Reviews Stable URL : http://www.jstor.org/stable/223452. *Annual Review of Sociology*, *26*(2000), 441–462.

Sanagou, M., Wolfe, R., Forbes, A., & Reid, C. M. (2012). Hospital-level associations with 30-day patient mortality after cardiac surgery: A tutorial on the application and interpretation of marginal and multilevel logistic regression. *BMC Medical Research Methodology*, *12*. doi: 10.1186/1471-2288-12-28

Sanku, D. (2013). Identifying predictors of childhood anemia in north east india. , *31*(4), 462. (Journal of health, population, and nutrition.)

Skrondal, A., & Rabe-Hesketh, S. (2007, dec). Latent variable modelling: A survey. *Scandinavian Journal of Statistics*, *34*(4), 712–745. doi: 10.1111/j.1467-9469.2007.00573.x

Snijders, A. B. R., Tom. (2001). Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling.
((London; Sage, 2001))

Sommet, N., & Morselli, D. (2017). Keep calm and learn multilevel logistic modeling: A simplified three-step procedure using stata, R, Mplus, and SPSS. *International Review of Social Psychology*, *30*(1), 203–218. doi: 10.5334/irsp.90

Syed, S., Addo, O. Y., De La Cruz-Góngora, V., Ashour, F. A., Ziegler, T. R., & Suchdev, P. S. (2016). Determinants of anemia among school-aged children in Mexico, the United States and Colombia. *Nutrients*, *8*(7), 1–15. doi: 10.3390/nu8070387

Tezera, R., Sahile, Z., Yilma, D., Misganaw, E., & Mulu, E. (2018). Prevalence of anemia among school-age children in Ethiopia: A systematic review and meta-analysis. *Systematic Reviews*, *7*(1), 1–7. doi: 10.1186/s13643-018-0741-6

Thompson, D. M., Fernald, D. H., & Mold, J. W. (2012). Intraclass correlation coefficients typical of cluster-randomized studies: Estimates from the robert wood johnson prescription for health projects. *Annals of Family Medicine*, *10*(3), 235–240. doi: 10.1370/afm.1347

Vugutsa Luvai, L., Mohamed Mohamed El-Sayed, A., Goldstein, H., Rasbash, J., Example, A. S., Aa, M. A., ... Livingston, M. (2017). *Multi-level Models Powerpoint* (Vol. 4; Tech. Rep. No. 1).

WHO. (2008). *World prevalence of anemia 1993-2005: Who global database on anemia*. Switzerland: Geneva.

APPENDICES

Appendix 1: Tables

Table 11: Multilevel Regression of District Estimates

Variable	OR	STDerr	Z	Pvalue	95 %CI
Residence					
Urban (ref)					
Rural	2.065	0.873	1.72	0.086	0.90-4.73
Age of child					
5 to 10 (ref)					
11 to 14	0.550	0.130	-2.53	0.012	0.46-0.88
Child sex					
Girl (ref)					
Boy	0.879	0.176	-0.64	0.521	0.59 - 1.30
Head sex					
Male (ref)					
Female	0.737	0.180	-1.25	0.211	0.46-1.19
Water Source					
Improved (ref)					
Unimproved	1.093	0.292	0.33	0.740	0.65-1.84
Malaria					
Yes (ref)					
No	0.731	0.241	-0.95	0.340	0.38-1.39
Inflammation					
None (ref)					
Mild	2.321	0.560	3.49	0.001	1.45-3.73
Severe	4.151	1.082	5.46	0.001	2.49-6.92
VitA					
Yes (ref)					
No	0.543	0.338	-0.98	0.326	0.16-1.84
Wealth Index					
Poor (ref)					
Poorer	0.727	0.224	-1.04	0.300	0.40-1.33
Middle	0.625	0.20	-1.50	0.133	0.34-1.15
Richer	0.806	0.256	-0.68	0.496	0.43-1.50
Richest	0.338	0.143	-2.56	0.011	0.15-0.76
Mother Alive					
Yes (ref)					
No	1.317	0.664	0.55	0.585	0.49-3.54
Father Alive					
Yes (ref)					
No	1.753	0.776	1.27	0.204	0.74-4.17
Don't Know	7.627	11.859	1.31	0.191	0.36-160.60
β	0.154	0.164	-1.76	0.078	0.02-1.24

Table 12: Multilevel Regression of Community Estimates

Variable	OR	STDerr	Z	Pvalue	95 %CI
Residence					
Urban (ref)					
Rural	1.781	0.864	1.19	0.234	0.69-4.61
Age of child					
5 to 10 (ref)					
11 to 14	0.552	0.133	-2.46	0.014	0.34-0.89
Child sex					
Girl (ref)					
Boy	0.891	0.185	-0.56	0.578	0.59 - 1.34
Head sex					
Male (ref)					
Female	0.994	0.251	-1.23	0.821	0.457-1.56
Water Source					
Improved (ref)					
Unimproved	1.042	0.307	0.14	0.890	0.58-1.86
Malaria					
Yes (ref)					
No	0.644	0.221	-1.28	0.199	0.33-1.26
Inflammation					
None (ref)					
Mild	2.438	0.612	3.55	0.001	1.49-3.99
Severe	4.250	1.168	5.26	0.001	2.48-7.28
VitA					
Yes (ref)					
No	0.481	0.318	-1.11	0.267	0.13-1.75
Wealth Index					
Poor (ref)					
Poorer	0.791	0.255	-0.73	0.468	0.42-1.50
Middle	0.718	0.232	-1.05	0.292	0.37-1.35
Richer	0.921	0.306	-0.25	0.805	0.48-1.77
Richest	0.361	0.161	-2.28	0.023	0.15-0.87
Mother Alive					
Yes (ref)					
No	1.314	0.685	0.52	0.601	0.47-3.65
Father Alive					
Yes (ref)					
No	1.970	0.918	1.45	0.146	0.79-4.91
Don't Know	15.285	25.244	1.65	0.099	0.60-389.10
β	0.153	0.171	-1.67	0.094	0.02-1.38

Table 13: Logistic Regression Analysis

Variable	OR	STDerr	Z	Pvalue	95 %CI
Residence					
Urban (ref)					
Rural	1.695	0.679	1.32	0.187	0.77-3.72
Age of child					
5 to 10 (ref)					
11 to 14	0.536	0.122	-2.73	0.006	0.34-0.84
Child sex					
Girl (ref)					
Boy	0.936	0.179	-0.35	0.729	0.64 - 1.36
Head sex					
Male (ref)					
Female	0.876	0.195	-1.60	0.551	0.57-1.35
Water Source					
Improved (ref)					
Unimproved	1.083	0.259	0.33	0.739	0.68-1.73
Malaria					
Yes (ref)					
No	0.662	0.208	-1.32	0.188	0.36-1.22
Inflammation					
None (ref)					
Mild	2.351	0.540	3.72	0.001	1.45-3.69
Severe	3.637	0.892	5.26	0.001	2.25-5.88
VitA					
Yes (ref)					
No	0.512	0.310	-1.11	0.269	0.16-1.68
Wealth Index					
Poor (ref)					
Poorer	0.784	0.227	-0.84	0.402	0.45-1.38
Middle	0.695	0.197	-1.28	0.199	0.40-1.21
Richer	0.802	0.232	-0.76	0.446	0.46-1.44
Richest	0.406	0.159	-2.30	0.021	0.19-0.87
Mother Alive					
Yes (ref)					
No	1.19	0.573	0.36	0.717	0.46-3.06
Father Alive					
Yes (ref)					
No	1.478	0.625	0.92	0.357	0.65-3.38
Don't Know	9.581	14.549	1.49	0.137	0.45-187.94
β	0.267	0.267	-1.32	0.187	0.04-1.90

Table 14: Correlation Matrix in Logistic Analysis

Correlation matrix of coefficients of logistic model																		
Covariance matrix	Zinc	EA	Malaria	Vit A	Age 11-14yr	Inflam 1	Inflam 2	Child Sex	Mother Alive	Father Alive	Don't Know	Poorer	Middle	Richer	Richest	Head Sex	Individual Variance	Cluster variance
Zinc	1.00																	
EA	-0.08	1.00																
Malaria	-0.05	0.02	1.00															
Vit.A	-0.06	0.03	0.00	1.00														
Age	-0.03	-0.02	0.01	0.04	1.00													
Inflamention 1	0.01	-0.01	0.00	0.01	0.17	1.00												
Inflamention 2	0.09	-0.02	0.09	-0.06	0.11	0.35	1.00											
ChildSex	-0.01	-0.01	0.11	-0.10	0.00	0.02	0.08	1.00										
MotherAlive	-0.02	-0.05	-0.01	0.03	0.02	0.00	0.04	-0.07	1.00									
FatherAlive	-0.01	0.01	0.04	0.03	0.07	0.05	0.06	0.01	-0.19	1.00								
Don't Know	0.01	-0.05	0.00	-0.01	0.07	0.08	0.09	0.07	-0.06	0.27	1.00							
Poorer	0.08	-0.03	-0.06	-0.04	0.02	0.00	-0.11	-0.04	-0.03	0.07	-0.04	1.00						
Middle	0.02	0.06	-0.01	-0.04	0.02	0.01	0.03	-0.06	0.02	0.05	0.00	0.49	1.00					
Richer	0.02	0.08	0.00	-0.02	-0.03	-0.01	-0.04	-0.07	0.02	0.09	0.00	0.51	0.52	1.00				
Richest	-0.02	0.34	-0.01	-0.06	0.02	0.01	0.00	-0.06	0.01	0.03	-0.10	0.38	0.42	0.43	1.00			
Sex of Head	0.03	0.00	0.08	-0.04	0.01	0.03	-0.05	-0.04	-0.01	0.18	0.00	0.17	0.20	0.22	0.18	1.00		
Variance	-0.22	-0.36	-0.32	-0.56	-0.12	-0.13	-0.14	-0.03	-0.36	-0.37	-0.08	-0.14	-0.19	-0.22	-0.23	-0.18	1.00	
Correlation matrix of coefficients of multilevel model																		
Zinc	1.00																	
EA	-0.06	1.00																
Malaria	-0.05	0.02	1.00															
Vit.A	-0.07	0.03	0.01	1.00														
Age	-0.02	-0.02	0.01	0.05	1.00													
Inflamention 1	0.01	-0.01	0.01	0.01	0.18	1.00												
Inflamention 2	0.09	-0.01	0.08	-0.06	0.13	0.37	1.00											
ChildSex	-0.01	-0.02	0.10	-0.11	-0.01	0.01	0.05	1.00										
MotherAlive	-0.02	-0.06	-0.01	0.03	0.02	0.03	0.09	-0.07	1.00									
FatherAlive	0.01	0.00	0.01	0.02	0.06	0.07	0.09	0.01	-0.14	1.00								
Don't Know	0.01	-0.05	-0.01	-0.03	0.07	0.09	0.11	0.08	-0.04	0.29	1.00							
Poorer	0.08	-0.03	-0.08	-0.04	0.03	0.00	-0.09	-0.04	0.02	0.06	-0.02	1.00						
Middle	0.02	0.06	-0.02	-0.03	0.03	0.05	0.06	-0.07	0.08	0.04	0.01	0.50	1.00					
Richer	0.03	0.08	-0.05	-0.05	-0.04	-0.01	-0.01	-0.07	0.04	0.11	0.02	0.50	0.53	1.00				
Richest	-0.01	0.28	-0.02	-0.06	0.02	-0.01	-0.03	-0.06	0.03	0.00	-0.11	0.39	0.44	0.45	1.00			
Sex of Head	0.03	0.01	0.06	-0.07	0.02	0.04	0.00	-0.04	0.02	0.18	0.01	0.17	0.22	0.22	0.18	1.00		
Individual variance	-0.21	-0.38	-0.29	-0.54	-0.13	-0.15	-0.18	0.00	-0.38	-0.38	-0.09	-0.15	-0.23	-0.23	-0.21	-0.19	1.00	
Cluster variance	0.01	0.03	0.00	-0.03	0.02	0.10	0.22	-0.05	0.09	0.16	0.11	0.05	0.05	0.12	-0.06	0.10	-0.19	1.00

Table 15: Correlation Matrix in multilevel Analysis

Covariance matrix	Zinc	EA	Malaria	Vit A	Age 11-14yr	Inflam 1	Inflam 2	Child Sex	Mother Alive	Father Alive	Don't Know	Poorer	Middle	Richer	Richest	Head Sex	Individual Variance	Cluster variance
Correlation matrix of coefficients of weighted logistic model																		
Zinc	1.00																	
EA	-0.11	1.00																
Malaria	0.02	0.14	1.00															
Vit.A	-0.15	0.05	-0.01	1.00														
Age	-0.08	0.01	-0.12	0.10	1.00													
Inflamention 1	0.00	0.13	0.02	-0.08	0.12	1.00												
Inflamention 2	0.19	0.04	0.05	-0.14	0.09	0.35	1.00											
ChildSex	0.07	-0.04	0.13	-0.16	-0.13	0.03	0.14	1.00										
MotherAlive	0.01	-0.13	0.02	0.06	0.00	-0.09	-0.10	-0.12	1.00									
FatherAlive	0.03	0.01	-0.01	0.07	0.18	0.08	0.06	-0.10	-0.24	1.00								
Don't Know	0.01	-0.07	-0.03	-0.01	0.08	0.06	0.07	0.03	-0.08	0.23	1.00							
Poorer	0.16	-0.06	-0.06	-0.07	-0.03	0.00	-0.02	-0.13	0.01	0.04	-0.03	1.00						
Middle	0.17	-0.06	-0.03	-0.12	0.00	-0.01	0.07	-0.14	0.04	0.01	-0.01	0.58	1.00					
Richer	0.15	0.14	0.01	-0.01	-0.02	0.00	-0.04	-0.20	0.08	0.04	-0.02	0.56	0.56	1.00				
Richest	0.16	0.44	0.12	-0.09	-0.06	0.07	0.02	-0.01	0.04	-0.04	-0.12	0.38	0.38	0.45	1.00			
Sex of Head	0.09	0.11	0.16	-0.03	-0.14	0.04	-0.08	-0.11	0.05	0.12	-0.03	0.29	0.23	0.26	0.24	1.00		
Variance	-0.31	-0.42	-0.43	-0.42	-0.11	-0.14	-0.12	0.05	-0.28	-0.38	-0.03	-0.19	-0.18	-0.33	-0.38	-0.31	1.00	
Correlation matrix of coefficients of weighted multilevel model																		
Zinc	1.00																	
EA	-0.18	1.00																
Malaria	0.06	0.27	1.00															
Vit.A	-0.22	-0.02	-0.07	1.00														
Age	-0.07	0.11	-0.12	-0.21	1.00													
Inflamention 1	-0.10	0.09	-0.07	0.00	0.21	1.00												
Inflamention 2	0.34	0.11	0.14	-0.24	0.06	0.56	1.00											
ChildSex	0.22	0.17	0.30	-0.14	-0.35	-0.10	0.15	1.00										
MotherAlive	0.01	-0.32	-0.04	0.18	-0.10	-0.24	-0.13	-0.31	1.00									
FatherAlive	0.00	-0.04	0.09	-0.27	0.26	0.03	0.26	-0.03	-0.02	1.00								
Don't Know	0.02	-0.09	-0.13	-0.01	0.04	0.12	0.09	0.01	-0.05	0.30	1.00							
Poorer	0.29	-0.14	0.12	-0.07	-0.11	-0.06	0.16	-0.05	0.24	0.23	0.07	1.00						
Middle	0.08	-0.29	-0.05	-0.11	-0.06	-0.24	0.06	-0.22	0.34	0.09	0.00	0.49	1.00					
Richer	0.19	0.17	0.30	-0.21	0.13	-0.12	0.16	-0.10	0.17	0.23	-0.02	0.48	0.41	1.00				
Richest	0.10	0.43	0.32	-0.18	0.04	-0.10	0.20	0.31	-0.08	0.03	-0.16	0.33	-0.02	0.41	1.00			
Sex of Head	0.04	0.12	0.19	-0.22	0.09	-0.19	0.14	-0.05	0.16	0.46	0.00	0.46	0.28	0.34	0.35	1.00		
Variance	-0.24	-0.34	-0.53	-0.19	0.00	0.03	-0.33	-0.08	-0.36	-0.44	-0.05	-0.45	-0.14	-0.47	-0.37	-0.47	1.00	
Cluster variance	0.16	-0.03	0.26	-0.12	0.00	0.08	0.23	0.01	0.23	0.34	0.09	0.27	0.11	0.32	0.05	0.19	-0.46	1.00

Table 16: Checking Multicollinearity in Level-2 Model

Variable	**VIF**	**1/VIF**
Weight of Child	1.09	0.920832
WealthIndex		
Poorer	1.62	0.616067
Middle	1.73	0.57949
Richer	1.67	0.598543
Richest	1.59	0.628896
Inflammention		
Mild	1.09	0.921043
Severe	1.09	0.915546
Mean VIF	**1.41**	

Appendix 2: STATA DO FILE

```
use ”anemia.dta”, clear
recode hv201 (10/31=1) (41 = 1) (51= 1) (nonmissing = 2), gen(watersource)

lab def watersource 1 ”improved” 2 ”unimproved”
lab val watersource watersource

lab def childSex 1 ”girl” 2 ”boy”
lab val childSex childSex
lab def sac_agecat 1 ”5-10” 2 ”11-14”
lab val sac_agecat sac_agecat
lab def mtype 1 ”Urban” 2 ”Rural”
lab val mtype mtype

replace malaria = 2 if missing(malaria)
lab def malaria 1 ”Yes” 2 ”No”
lab val malaria malaria

//Dealing with missing values

replace inflame = 0 if missing(inflame)
lab def inflame 0 ”No inflamation” 1 ”Mild” 2 ”Severe”

lab val inflame inflame

replace m311 = 2 if missing(m311) //Vitamin A

lab def m311 1 ”Yes” 2 ”No”
lab val m311 m311

replace hv107 = 0 if missing(hv107) //Those who never attended school
```

recode hv107 (0/4=1 ”0-4”) (5/8=2 ”5-8”), pre(school_cat) label(school_cat) rename school_cathv107 school_cat

ORDINARY LOGISTIC REGRESSION

logistic anemia i.mtype i.malaria i.childSex i.wealthIndex i.head_sex i.momAlive i.dadAlive i.sac_agecat i.m311 i.inflame i.watersource,

*ESTIMATING OF MULTILEVEL LOGISTIC REGRESSION

meqrlogit anemia i.mtype i.malaria i.childSex i.wealthIndex i.head_sex i.momAlive i.dadAlive i.sac_agecat i.m311 i.inflame i.watersource, || mcluster:,or

estat icc

estat ic

*ESTIMATE FOR RANDOM INTERCEPT

melogit anemia i.mtype i.malaria i.childSex i.wealthIndex i.head_sex i.momAlive i.dadAlive i.sac_agecat i.m311 i.inflame i.watersource, || mcluster:

estat icc

estat ic

*ESTIMATE OF RANDOM COEFFICIENT

xtmixed anemia i.mtype i.malaria i.childSex i.wealthIndex i.head_sex i.momAlive i.dadAlive i.sac_agecat i.m311 i.inflame i.watersource, || mcluster:

estat icc

estat ic

==

RESEARCH

Modelling Community Effects on Anemia in School-Aged Children in Malawi using Multilevel Regression Model

Glory Gondwe Mshali[1,2] and Jupiter Simbeye[2*]

Sample of title note

[*]Correspondence:
jsimbeye@unima.ac.mw

[2]Department of Mathematical Sciences, University of Malawi, Chancellor College, Chirunga, 280 Zomba, Malawi

Full list of author information is available at the end of the article

[†]Senior author

Abstract

Background: Anemia is a widely spread public health problem and affects individuals at all levels. However there is evidence to show that both individuals and community level effect play an important role in the effect on anemia. Therefore, the aim of this study was to model community effect on anemia in school-aged children 5-14 years (SAC) in Malawi by using multilevel regression models.

Methods: Cross sectional data of 800 school-aged children 5-14 years from the Malawi Micronutrient Survey (MNS) 2015-16 and Malawi Demographic Health Survey (MDHS) 2015-16 was used for the analysis. The Variance Component Model, Null Model, Random Intercept model, Fixed Effect, random effect and Random Coefficients Model were used to find the better fit model and measure the effect within and between communities.

Results: Community differential analysis indicated that Multilevel logistic regression better suited the hierarchical clustered data with higher values of log likelihood estimates of -348.65 as compared to logistic regression model with values of -355.65. From the estimate of variance component, the intracommunity correlation of 0.153 showed that 15% of the variation were due to within community effects while 85% were related to between community differences. Further, the better model that would explain the within and between community variations in effect on anemia in school-aged children was the random coefficient model with less BIC and AIC of 730.33.

Conclusion: The study showed that the variation in effect and impact on anemia among school-aged children 5-14 years in Malawi was mainly linked to between community variations. Policies, interventions and programmes that aim at addressing this health problem should factor-in the community and location differences as a measure of tackling the community effects on anemia in school aged children.

Keywords: Cross sectional data; Multilevel modelling; Hierarchical data; Community variations

1 Background

Anemia refers to a condition in which the number of red blood cells or their oxygen-carrying capacity is insufficient to meet physiological needs; it continues to be an important public health problem worldwide, especially in developing countries[1]. Iron deficiency (ID) is the most common cause of anemia, which is responsible for around 25%-50% of all the cases of anemia worldwide [2] [3] [4]. The prevalence of anemia among school-aged children is 25.4%, three to four times more prevalent in non-industralized regions than industralized ones [5] [6] [7]. Anemia results in low resistance to diseases and increased susceptibility to infection, poor cognitive development, physical development and school performance [2]. It has also economic impact by increasing impairment of lives and disability, reduction in intellectual capacity and loss of productivity due to increased morbidity from infectious diseases. Children and adolescents have increased demand for iron due to their rapid growth and development. In these age groups of 5-14 years, Iron deficiency and anemia are common nutritional problems. [2]

In Malawi, anemia prevalence among school-aged children (SAC) and preschool children were estimated at 22% and 30 % respectively. In urban areas, among SAC it was 15.9% whilst in rural areas was 22.2%. [8]. Most anemia studies were focused on pregnant women and pre-school children in Malawi. Even though the Malawi Micronutrient study showed the prevalence of anemia in school-aged children at national-level, unfortunately, most of the studies conducted used single-level analysis technique and assumed that there was no community-level effects beyond the characteristics of individuals ([9] [1]) necessitating a study on the impact of community-level effects on anemia in school-aged children (5 -14 years) to be studied. The results from this study of locality variation in effect on anemia in school-aged children was found to be of interest as it concurs with several authors in the world ([4], [10]) . Hierarchical approach of analyzing clustered data guaranted correct estimation of parameters and standard errors [11]. Further, it will inform designing more effective interventions on anemia in communities.

Usage of multilevel regression in analyzing hierarchical data from MDHS and MNS (data collected in cluster form) that tends to investigate individual-level and community-level effects on anemia in school-aged children is rare in literature. In this paper we develop a multilevel regression model to determine the individual-level and community-level effects on anemia in school aged children 5 to 14 years in Malawi. The hierarchical approach was applied by considering the nature of data, and the main concern was on how this multilevel regression modeling will be applied in hierarchical data to investigate the community effects on anemia in school-aged children, check for both the variation between locations and also checking for correlates of these variations in communities [12] [13]. The correlates will be the risk factors that have association on anemia at the individual level and also vary between communities [14]. Lastly investigating the impact of community variance component on coefficients of the estimated model of the effect on anemia in school-aged children 5-14 years old.

2 Materials and methods

This study used data from the 2015-16 Malawi Demographic Health Survey (MDHS) and the Malawi Micronutrient Survey of 2015-16. These surveys were conducted by National Statistical Office from 19 October, 2015 to 18 February, 2016 in joint collaboration with the Ministry of Health (MoH), Centre for Disease Control and Prevention (CDC), the Community Health Services Unit (CHSU) and International Care Foundation (ICF). The survey was based on a nationally representative sample that provided estimates at the national and regional levels and for urban and rural areas with key indicator estimates at the district level. The survey included 850 (MDHS) clusters and 26,361 (MDHS) households of which the micronutrient sample allocation of clusters and households were 105 and 2,262 respectively.

Variables Included in the Study

The variables considered in the study were at national level. They covered; demographic characteristics, health characteristics, social characteristics, community characteristics and geographic characteristics. These variables were considered/classified as dependent and explanatory variables.

Dependent Variables

The dependent variable for the study was anemia. It was dichotomous; coded as (1)-if the child had anemia and (0)-if the child had no anemia.

$$Y_{ij} = \begin{cases} 1 & \text{for children having anemia} \\ 0 & \text{for normal(not anemic) children} \end{cases} \quad (1)$$

Explanatory Variables

In this study the explanatory variables were; demographic i.e. (age, sex, education); economic i.e. (wealth index); health i.e. (malaria, inflammation, zinc); and socio i.e. (source of drinking water) were expected to have impact on anemia in school-aged children (SAC) and were classified as individual level variables and community level variables.

2.1 Statistical Methods

The multilevel logistic regression was used to model community effects on anemia in school-aged children of 5-14 years old. Multilevel model for measuring community effects, investigating of correlates on anemia and the impact of community variance component on the coefficients of the estimated model on anemia in school-aged children was developed. The response variable of the study was anemia. Ordinary logistic regression was the obvious model of choice when one thinks of modeling a binary outcome. Considering the nature of data used in this study, the cluster sampling and data obtained from multistage-clustered samples, a multilevel regression model was to be

applied. The use of single-level statistical models was no longer valid and reasonable. Consequently, when using the single-level (logistic regression) most of the factors would appear significant which would result in giving wrong policy implications for Malawi. In order to draw appropriate inferences and conclusions from multistage stratified clustered survey data, application of the modeling techniques as multilevel logistic regression modeling was the best technique to be used.

2.1.1 ***Binary Logistic Regression Model***

The binary logistic regression model was used to investigate effects of predictors on anemia before going in to multilevel modeling. The focus was on using two-level hierarchical data (individuals and communities) in estimating the effect on anemia (binary outcome) in children aged 5-14 years.

Let π_{ij} be the proportion of anemia in school-aged child $i(i = 1, 2, 3, ..., i)$ in community $j(j = 1, 2, 3, ..., j)$ then,
$\Pr(Y_{ij} = 1) = \pi_{ij}$ and $\Pr(Y_{ij} = 0) = 1 - \pi_{ij}$, where $Y_{ij} \sim Bernoulli(\pi_{ij})$

The logistic model is defined as follows; $X_{nx}(k + 1)$ denote the single level binary logistic regression data matrix of k predictor variables of the school-aged children given as:

$$\mathrm{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & . & . & . & X_{1k} \\ 1 & X_{21} & X_{22} & . & . & . & X_{2k} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & X_{n1} & X_{n2} & . & . & . & X_{nk} \end{bmatrix} \sim nx(k+1), \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_k \end{bmatrix} \sim (k+1)xl$$

where X - is the design matrix
β - is the vector of unknown coefficients of the covariates and intercept.

2.1.2 ***Intraclass correlation coefficient (ICC)***

Intraclass correlation coefficient (ICC) represented the proportion of the total variance that was attributable to between-group differences; and it provided an assessment of whether or not significant between-group variation existed in the model [15] [16]. The ICC was estimated by using the mathematical formula:

$$ICC = \rho = \log \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \tag{2}$$

where σ_u^2 was the between-community variance and σ_e^2 was the within-community variance.

2.1.3 *Random Intercept Binary Logistic Regression Model*

When data is from different communities, a varying - intercept model can be interpreted as a model with different intercept within each community [17]. In this case, the intercept model considered only the random effect of SAC meaning that the communities differed with respect to the SAC who were anemic, but not explaining differences between communities. Suppose there is a random intercept model expressed as:

$$\text{logit}(\pi_{ij}) = \beta_{0j} + \beta_1 X_{1ij} + \beta_2 X_{2ij} + ... + \beta_k X_{kij} = \beta_{0j} + \Sigma_{h-1}^{k} \beta_h X_{hij} \tag{3}$$

Where, $\text{logit}(\pi_{ij})$ does not include level-1 of SAC having anemia. β_{0j} varies randomly and explained by average intercept of β_0 and community dependent deviations of u_j. Replacing $\beta_{0j} = \beta_0 + u_j$ in equation (3) will get:

$$\text{logit}(\pi_{ij}) = \beta_0 + \sum_{i=1}^{J} \beta_i X_{ij} + U_j \tag{4}$$

Where, β_i is the unit difference between X_i values of individuals in the same community which is associated with the log-odds with the difference of β_i. u_j denoted randomness in communities and is assumed that they are mutually independent and normally distributed with mean zero and variance of σ_0^2.

2.1.4 *The Random Coefficients Logistic Regression Model*

In random coefficient model, both the intercept and slopes differ across the communities. Consider k, the explanatory variables $X_1, \cdots, X_k$ and values of $X_h (h = 1, \cdots, k$ which can be indicated as X_{hij} for $h = 1, \cdots, k; i = 1, \cdots, n$ and $j = 1, \cdots, N$. Some of these variables could be level-1 variables where the effects on anemia probability may not be the same for all the school-aged children in a given locality. The effects on anemia probability may depend on the individuality of individual school-aged children but on the same time on their localities and this could be donated as P_{ij}. The outcome variable was expressed as the sum of effects on anemia probability which was the (expected value of the outcome variable) and the residual term e_{ij}. where,

$$y_{ij} = p_{ij} + e_{ij} \tag{5}$$

The residual e_{ij} are assumed to have mean zero and variance σ_e^2. The logistic regression models with random coefficients expresses the log-odds of logit P_{ij}, sum of a linear function of the explanatory variables with randomly varying coefficients. That is:

$$logit\,(P_{ij}) = log\left(\frac{P_{ij}}{1 - P_{ij}}\right) = \beta_{0j} + \beta_{1j} X_{1ij} +, \cdot, + \beta_{kj} X_{kj} \tag{6}$$

let $\beta_{0j} = \beta_0 + U_{0j}$ and $\beta_{hj} = \beta_h + U_{hj}$ for $h = 1, \cdots, k$

$$logit\left(P_{ij}\right) = log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \beta_0 + \sum_{h=1}^{k}\beta_h X_{hij} + U_{0j} + \sum_{h=1}^{k} U_{hj}X_{hij} \tag{7}$$

P_{ij} can be solved as:

$$\frac{e^{\beta_0+\sum_{h=1}^{k}\beta_h X_{hij}+U_{0j}+\sum_{h=1}^{k}U_{hi}X_{hij}}}{1+e^{\beta_0+\sum_{h=1}^{k}\beta_h X_{hij}+U_{0j}+\sum_{h=1}^{k}U_{hj}X_{hij}}} \tag{8}$$

Therefore, a unit difference between the X_h values of two school-aged children 5-14 years in the same community is associated with a difference of β_h in their log-odds, or equivalently, a ratio of exp (β_h) in their odds. The solved P_{ij} do not include level one residual because it is the equation for probability of P_{ij} rather than outcome Y_{ij}. The fixed part of the model $\beta_0 + \sum_{h=1}^{k}\beta_h X_{hij}$.

Estimation of Coefficients

2.1.5 *Random effect*

In this Model 9, there is fixed effects and random effects making it to be mixed-effect model. The random effects part of the model can not be estimated but can be summarized according to their estimated variances and co-variances. The random effect varied across different levels of hierarchy while allowing for correlation with observations at all levels of the model. The model assumed that the community effects were random.

And it was assumed as:

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 X_{ij} + u_j \tag{9}$$

and $u_j \sim N(0, \sigma_u^2)$

where σ_u^2 is the level-2 community variance or the between-community variance in the log-odds that $y = 1$ after accounting for x.

The random effects in the model allowed to examine the role of contextual community effects on child anemia. Possible contextual effects were measured by the variance partition coefficient (VPC); this was a variant of intraclass correlation coefficient (ICC) when the outcome was nonlinear. For a dichotomous variable such as presence or absence on anemia, VPC was calculated using formula used by [18].

$$VPC = V_n/(V_n + \pi^{2/3})$$

Vn = community variance, and *VPC* represented the percentage of total variance of the effect on anemia in school-aged children attributable to the community level and was also used as a measure of clustering on anemia in communities. A high *VPC* would reflect a high clustering on anemia effects and a high community effect on individual risk on anemia.

2.1.6 ***Fixed effect***

In model 9 again, the fixed effects were associated with predictors at any level in the outcome variable. Fixed effect were estimated in the parameter model and were represented as: $\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_k x_{kij}$ which could be estimated directly where;

β_0 was the log odds that $y = 1$ when $x = 0$ and $u = 0$.

β_1 was an effect on log-odds of 1-unit increase in x for individuals in same group (same value of u).

β_1 was often referred to as cluster - specific or unit specific effect of x.

expβ_1 was an odds ratio, comparing odds for individuals spaced 1-unit apart on x but in the same group.

2.1.7 ***The Variance Components Model***

Variance component two-level model was used for a dichotomous outcome variable to check for normality assumptions in level-2 units (communities); and specify the probability distribution for group-dependent probabilities π_j in $Y_{ij} = \pi_j + \varepsilon_{ij}$ without taking further explanatory variables into account. The focus was on the model that specified the transformed probabilities $f(\pi_j)$ to have a normal distribution. This was expressed, for a general link function $f(\pi_j)$, by the formula

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_{0j} + U_{0j} \tag{10}$$

where β_0 had the community average of the transformed probabilities and U_{0j} the random deviation from this average for group j. The variance component model could reveal the fixed part of Model 10 in the analysis and could test for existence of community variations;

2.1.8 *Proportional change in variance (PCV)*

PCV was calculated with reference to the null model to check for relative contribution of factors to explain variation of anemia in school-aged children.

$$PCV = \left(\frac{\sigma_u^2 - \sigma_{u_1}^2}{\sigma_u^2}\right) * 100 \tag{11}$$

where: σ_u^2 was the between community variance in the null model. $\sigma_{u_1}^2$ was the between community variance in Model 11 [19].

2.1.9 *Testing of level-2 (residual) variance*

Testing of level two variance or the between-group variance was from log-odds that $y = 1$ after accounting for x. A Wald test can be used to test for community differences and can be rejected if $\sigma_u^2 = 0$

2.2 Parameter Estimation using the Likelihood Function

The Likelihood function reflects information about the parameters contained in the model. For the two-level logistic Bernoulli responses, the random effect were assumed to be multivariate normal and independent across the community. The marginal likelihood function that was given by:

$$l(\beta, \Omega) = \Pi_i f \Pi_i [(\pi_{ij})^{y_{ij}} (1 - \pi_{ij})^{(1-y_{ij})}] \tag{12}$$

where;
Ω was variance co-variance matrix.

The likelihood contributed from the *ith* subject and in the jth group were as Bernoulli:

$$\text{Bernoulli}(p_{ij}) = p_{ij}^{yij} (1 - p_{ij})^{1-y_{ij}} \tag{13}$$

where;
p_{ij} represents the probability of the event for subject i in j community and that the co-variate vector were x_{ij} and y_{ij} that indicated having anemia, $(y_{ij} = 1)$ and no anemia $(y_{ij} = 0)$. In multilevel logistic regression we had:

$$p_{ij} = \frac{e^{\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_k x_{kij} + U_{0j}}}{1 + e^{\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_k x_{kij} + U_{0j}}} \tag{14}$$

where $\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_k x_{kij}$ was fixed part of the model and U_{0j} was the random part of the model and $U_0 \sim N(0, \sigma_u^2)$.

3 RESULTS

The results have been analyzed into two parts; the descriptive analysis and multilevel regression analysis from data which was in hierarchical in nature.

3.0.1 Descriptive Analysis

Descriptive statistics are a set of data that are described in form of figures that can represent the entire sample. A total of 105 clusters were included in the sample with 2,262 households. Eight hundred (800) school-aged children 5-14 years old were randomly sampled to participate in the study.

Bivariate Analysis Between a Response Variable and Explanatory Variable

The bivariate analysis provided association between the response variable and the explanatory variables. The chi-square explained the association between all the explanatory variables and the response variable. The high values in the chi-square showed how strong the association was while keeping other factors constant. The decision was made on chi-square and p-value of 0.05.

The descriptive statistic that summarized the association between explanatory variables and response variable has been presented in Table 1. The results from the table has shown row percentage on anemia in school-aged children 5-14 years old. From the Table there was a bivariate association between anemia status in school-aged children and independent variables which showed that anemia was strongly associated with place of residence (Urban, Rural), age of the child, inflammation, wealth quintile and weight of the child. Other independent variables like: Malaria in the last two weeks , if received Zinc and if either mother is alive or father is alive, Sex of the head of household and finally if the child received vitamin A, were not significant.

The effects on anemia in school-aged children in 5-14 years old was high for children with moderate inflammation and severe inflammation being (28%, 38%) respectively. The school-aged children 5-14 years from the poorest quintile had higher percentage on anemia effects with 25% with p-value of 0.02 and less effects were from richest families 10%. The proportion on anemia among school-aged children of 5 to 14 years as observed in (Table 1) differed with their age groups. The higher proportion on anemia was observed in school-aged children of 5 to 10 years (23%) as compared to school aged children of 11 to 14 years 12%. The effects on anemia also differed by place of residence. The higher number of school-aged children with anemia 21% resided in rural areas while a small number of anemic children resided in urban areas 9%. Other independent variables like: Malaria in the last two weeks, if received zinc, if mother is alive or father is alive, sex of the head of a household and finally if the child received vitamin A, were not significant.

The drawback of using descriptive statistics approach was that it ignored the possibility that a collection of variables that could be weakly associated with the outcome (Table 1).

Table 1 Cross Tabulation of Anemia Status versus Explanatory Variables

Variables	Anemia Status				Total	DF	CHI	P-Value
	Anemic		Not Anemic					
	N	%	N	%				
Total	154	19	646	81	800			
Place of Residence								
Urban	9	9	87	91	96	1	6.84	0.009
Rural	145	21	559	79	704			
Age of a Child								
5 to 10 Years	123	23	411	77	534	1	14.8	0.000
11 to 14 Years	31	12	235	88	266			
Sex of a Child								
Girl	84	55	328	51	412	1	0.65	0.420
Boy	70	45	316	49	386			
Education of a Child								
0 to 4	148	96	565	88	713	1	9.15	0.002
5 to 8	6	4	79	12	85			
Water source								
Improved	119	77	540	84	659	1	3.74	0.053
Unimproved	35	23	4	16	39			
Malaria in the last Two weeks								
Yes	18	28	46	72	64	1	3.56	0.059
No	134	18	593	82	727			
Inflamation								
None	63	13	432	87	495			
Mild	43	28	108	72	151	2	45.3	0.000
Severe	42	38	70	63	112			
Wealth Index								
Poorest	38	25	115	75	153			
Poor	34	23	117	77	151			
Middle	35	18	158	82	193	4	12.4	0.015
Richer	34	20	134	80	168			
Richest	13	10	122	90	135			
If Received ZINC								
Yes	149	19	616	81	765	1	0.591	0.442
No	5	14	30	86	35			
Mother Alive								
Yes	148	19	610	80	758	1	0.6	0.439
No	6	15	35	85	41			
Father Alive								
Yes	145	20	593	80	738	2	2.532	0.282
No	8	14	51	86	59			
Head Sex								
Male	111	19	465	81	576	1	0.01	0.964
Female	43	19	181	81	224			
Received VIT. A	5	33	10	67	15			
Not Receive VIT. A	137	19	595	81	732	1	2.04	0.15

3.1 Results of Multilevel Regression Model

From Table 2 the variance component for the community level at level-2, effects on anemia in school-aged children was significant 0.001 which showed that there was variability/clustering at level-2 that allowed the application of multilevel regression modeling. Using the intracommunity correlation (ICC) from Table 2 which was greater than zero (0.085) indicated that multilevel modeling should be applied in the clustered data used in the study. If ICC is equal to zero, then multilevel modeling should not be applied which was not the case with the this clustered data used for this study.

Table 2 Estimating of Multilevel Regression Model by Using Random Effect Model

Anemia	Estimate	Std Error	Z	P-value	95 % Confidence Interval	
Cluster variance	0.013	0.004	10.46	0.001	0.006	0.025
Residual Variance	0.142	0.006			0.128	0.158
Intraclass Correlation						
Level	**ICC**	**Std Error**				
mcluster	0.085	0.027			0.045	0.154

3.2 Model Diagnostics

3.2.1 Variance Component Model

The variance component model predicted the probability on anemia status in school-aged children. The results from Table 3 revealed the information of the fixed effect that estimated the log-odds on anemia among children aged 5-14 years in Malawi with $\beta_0 = -1.631$. The β_0 expressed the overall proportion $\left(\frac{e^{-\beta}}{1+e^{-\beta}}\right) = 0.164$ of effect of anemia in SAC between 5-14 years in Malawi without accounting for other sources of variation. The between community variance U_0 using log odds of being anemic was estimated as $\sigma_u^2 = 0.593$. which showed that there was non-zero community variation and that the community variations contributed to effect on anemia in school-aged children 5-14 years. The intracommunity correlation coefficient (ICC) in the variance component model (Table 3) = 0.153, meant that 15 percent of the total variability in the effect on anemia in SAC 5-14 years old was significantly related to community level, whereas the remaining 85 percent was related to within community difference. The systematic differences from ICC gave concept of applying multilevel analysis in the data.

Table 3 Estimates for Variance Component Model in School-Aged Children Data

Fixed effects	Estimate	S.error	Z-value	P-Value
β_0	-1.631	0.140	-11.650	0.001
Log Likelihood	-382.850			
Random effect				
σ_u^2	0.593	0.232	2.559	0.005
ICC	0.153	0.051		
Chi-Square	17.18			0.001

3.2.2 *Cross level interaction*

The Cross level interaction in Table 4 used estimates in the covariance matrix to detect the covariance between the slope and intercept. The result of the covariance matrix in Table 4 showed that there was a negative relationship of $U_1, U_2 = -0.0066$. These results showed the existence of clustering in the data set implying that multilevel modelling was to be used in the analysis. Likewise, for the residual estimate of level-1 and the intercept estimate of level 2 were also greater than 0 showing the existence of the variations in communities.

Table 4 Estimates of Cross Level Interaction

		Estimate	Std. Error	Z-value	P-value
Residuals	Level 1	0.1129	0.0690	1.64	0.001
Intercept	Level 2	0.1241	0.0049	25.08	0.001
Slope	Level 2	0.1572	0.0066	23.70	0.001
Cov(I,S)	Level 2	-0.0066	0.0048	-1.38	0.001

3.3 Multicollinearity in Data Set

All variables were checked if they were correlated for multicollinearity. If multicollinearity exists were dropped by Using the variance Inflation factor (VIF) where:

$$\text{VIF} = < 1\text{: no multicollinearity}$$
$$\text{VIF = between 1 to 5: moderate multicollinearity}$$
$$\text{VIF} = > 5\text{: High multicollinearity}$$

From Table 5, the $\frac{1}{VIF}$ for each independent variable showed that multicollinearity was not existing since all variables: weight of the child, wealth index,and inflammation were less than 1, falling between 0 and 1. These results showed that there was no association among independent variables of $\beta_1, \beta_2, \beta_3, \cdot, \beta_k$. These variables with no multicollinearity were included to model effect on anemia in school aged children by applying multilevel modeling.

Table 5 Checking for Existence of multicollinearity in Data Set

Variable	VIF	1/VIF
Weight of Child	1.09	0.920832
WealthIndex		
Poorer	1.62	0.616067
Middle	1.73	0.57949
Richer	1.67	0.598543
Richest	1.59	0.628896
Inflammention		
Mild	1.09	0.921043
Severe	1.09	0.915546
Mean VIF	**1.41**	

3.3.1 The Multilevel Logistic Regression versus Ordinary Logistic Regression

From Table 6 where logistic regression produced almost similar results to multilevel logistic regression. These results were similar to other studies conducted by [2]. The estimates of the odds ratio are almost similar for all variables with minor differences, despite that Logistic regression overestimates the effects on anemia in school-aged children. The multilevel analysis had made it possible to quantify the contribution of community-level effects on anemia in school-aged children.

From community-level effects, the odds of school-aged children 5-14 years living in rural areas were more associated with effects on anemia using both models; the ordinary logistic regression and multilevel logistic regression with (OR = 1.70, 95% confidence interval [CI: 0.77-3.71] and (OR = 1.78, CI: 0.69 - 4.61). For control variables in the Table 6 and for school-aged children 11-14 years old had odds estimate of (0.54, CI: 0.34 - 0.84) in logistic regression analysis while in multilevel logistic regression was estimated at (OR = 0.55, CI:0.34 - 0.89) explaining the effect on anemia in both multilevel and ordinary logistic regression were almost the same with few variability as compared to children 5-10 years old.

On health related variables and household related variables; Checking for health related variables in Table 6, effects on anemia were two times and four times higher for children with mild to severe inflammation (OR = 2.44 to 4.25, CI: 1.49 - 3.99 to CI: 2.48 - 7.28) in multilevel logistic regression analysis. This gives a slight difference with the results of logistic regression analysis where the effect of anemia in school-aged children with mild inflammation had (OR = 2.35, CI: 1.45 - 3.69), and for children with severe inflammation had (OR = 3.63, CI: 2.25 - 5.88). For other health related variables like; if child received vitamin A, malaria in the past two weeks and source of drinking water gave similar results with minor variations in both ordinary logistic and multilevel regression. For household variables; school aged children from the richest household were less associated with anemia using either ordinary logistic regression or multilevel logistic regression (OR = 0.41, CI = 0.19 - 0.87) verses (OR = 0.36, CI = 0.15 - 0.87) as compared to those living in poorer families, middle and richer families with: (OR = 0.78, CI = 0.45 - 1.38) verses (OR = 0.79, CI: 0.42 - 1.49) in poorer families; (OR = 0.70, CI: 0.40 - 1.21) verses (OR = 0.71, CI = 0.34 - 1.35) in middle families; and (OR = 0.80, CI: 0.46 - 1.41) verses (OR = 0.92, CI = 0.47 - 1.77) in richer families

0.9

Table 6 Logistic Regression verses Multilevel Logistic Analysis

Variable	n	Logistic regression		Multilevel Logistic	
		OR	95%CI	OR	95%CI
Community Related Variables					
Residence					
Urban (ref)	96	-			
Rural	704	1.695	0.774 - 3.717	1.781	0.688 - 4.608
Control variables					
Age of child					
5 to 10 (ref)	534	-			
11 to 14	266	0.536	0.343 - 0.838	0.552	0.344 - 0.885
Sex of the child					
Girl (ref)	412	-			
Boy	386	0.936	0.643 - 1.363	0.891	0.594 - 1.338
Head sex					
Male (ref)	576	-			
Female	224	0.876	0.567 - 1.354	0.821	0.573 - 1.556
Health-related variables					
Water Source					
Improved (ref)	659	-			
Un improved	39	1.083	0.677 - 1.731	1.042	0.584 - 1.857
Malaria					
Yes (ref)	64	-			
No	727	0.662	0.358 - 1.224	0.644	0.330 - 1.260
Inflammation					
None (ref)	495	-			
Mild	151	2.351	1.450- 3.687	2.438	1.491 - 3.987
Severe	112	3.637	2.249 - 5.883	4.250	2.479 - 7.284
VitA					
Yes (ref)	15	-			
No	732	0.512	0.157 - 1.676	0.481	0.132 - 1.754
Household related Variables					
Wealth Index					
Poor (ref)	151	-			
Poorer	153	0.784	0.445 - 1.384	0.791	0.421 - 1.488
Middle	193	0.695	0.398 - 1.212	0.708	0.34 - 1.346
Richer	168	0.802	0.455 - 1.414	0.921	0.47 - 1.767
Richest	135	0.406	0.188 - 0.874	0.361	0.15 - 0.867
Mother Alive					
No (ref)	758	-			
Yes	41	1.191	0.464 - 3.058	1.314	0.473 - 3.651
Father Alive					
No (ref)	738	-			
Yes	59	1.478	0.645 - 3.384	1.970	0.790 - 4.911
Don't Know	59	9.581	0.448 - 187.940	15.285	0.600 - 389.100
Log likelihood	-355.63			-348.65	

3.3.2 Random Intercept Model

Table 7 identified the random intercept model which was the multilevel model with fixed effects and random effects. The analysis revealed that correlates of anemia varied among communities. The deviance based chi-square test for the random effects in random intercept model was χ^2= 13.95 (d.f. = 18, p-value = 0.0001). This indicated that the random intercept model gave a better fit as compared to variance component model in Table 7 with $\chi^2 = 17.18$ and p-value of 0.0001. The fixed part of the table showed variables like; wealth index, age of the child and if experience inflammation which were statistically significant on effects on anemia in SAC 5-14 years. From the results in Table 7, controlling for community differences in the effect on anemia in SAC would result in the odds of decreasing by a factor of $e^{-}0.5948309$ = 0.552 for each year increase in age group of 11-14 years.

The intraclass correlation coefficient (ICC) is a measure on variation of effect on anemia in SAC among communities. The ICC of 0.153 about 15% of the variation in effect on anemia in school-aged children was attributable to variation within communities and only about 85% of the variations were due to level two effects or between communities. The intraclass correlation coefficient was statistically significant at 5 percent level of significance. The random intercept in (Table 7) with chi-square probability of 0.0001 indicated that correlates on anemia in SAC differed from community to community taking into account all covariates measured.

The log odds of residing in rural areas, having no malaria in the past week, if mother was alive, if father was alive and if having mild or severe inflammation were having more than once chance in contributing to effect on anemia in SAC 5-14 years without affecting the random effects and covariates unchanged or without affecting community variations. The variance component of the random intercept was significant at 0.0001 suggesting that there remains some variation in the effect on anemia in school-aged children which were not accounted for by the variables in the model. These estimates can be justified by estimating an alternative model that contains random coefficient model.

Table 7 Estimate of Random Intercept Model

Variable	Coef.	Std.Er.	Z-value	P-value	OR	[95% Conf. Upper	Interval] Lower
Fixed effect							
Residence							
Urban (ref)							
Rural	0.577	0.485	1.19	0.234	1.781	-0.374	1.528
Malaria							
Yes (ref)							
No	-0.439	0.342	-1.28	0.199	1.552	-1.110	0.231
ChildSex							
Girl (ref)							
Boy	-0.115	0.207	-0.56	0.578	0.644	-0.521	0.291
Inflammention							
Moderate							
Mild	0.891	0.251	3.55	0.001	2.438	0.399	1.383
Severe	1.447	0.275	5.26	0.001	4.250	0.908	1.986
Wealth Index							
Poor							
Poorer	-0.234	0.322	-0.73	0.468	0.791	-0.866	0.398
Middle	-0.345	0.328	-1.05	0.292	0.708	-0.987	0.297
Richer	-0.082	0.332	-0.25	0.805	0.921	-0.733	0.569
Richest	-1.018	0.446	-2.28	0.023	0.361	-1.893	-0.143
HeadSex							
Male (ref)							
Female	-0.058	0.255	-0.23	0.821	0.944	-0.557	0.442
Mother Alive							
No (ref)							
Yes	0.273	0.522	0.52	0.601	1.314	-0.749	1.295
Father Alive							
No (ref)							
Yes	0.678	0.466	1.45	0.146	1.970	-0.235	1.592
Don't Know	2.727	1.652	1.65	0.099	15.286	-0.510	5.964
Age of Child							
5-10(ref)							
11-14	0.595	0.241	-2.46	0.014	0.554	-1.068	-0.122
Vitamin A							
Yes (ref)							
No	0.732	0.660	-1.11	0.267	0.481	-2.027	0.562
Water source							
Improved (ref)							
Unimproved	0.041	0.295	1.14	0.890	1.153	-0.537	0.619
β_0	-1.880	1.123	-1.67	0.094		-4.082	0.321
Random effect							
$\sigma_0^2 = var(U_0 j)$	0.595	0.249				0.262	1.351
ICC(ρ)	0.153						

3.3.3 **The Random Coefficients Model**

From Table 8 The random coefficient model sometimes called mixed effect model was used to check if there was an improvement in correlation over the random intercept model. This was done by adding a random coefficients in the level-1 model. The intercepts for random coefficient varied significantly at 5 percent significant level which implied that there were considerable variations in explanatory variables included in the model like: age of the child, inflammation and wealth index. These variables differed across communities. The variance component of 0.595 was found to be larger than their standard error of 0.249 showing the existence of variation between communities.

By extending the level-1 model an independent variable was added to random effect so that the model includes a random intercept and a random coefficient. By adding of level-1 predictors made ICC to increase and was as $\hat{\rho} = 0.153$ meaning that 15 percent of the total variability in the effect on anemia in school aged children of 5 to 14 years old was attributable to the random factor and community in the random coefficient of the multilevel logistic regression model (Table 8).

From the results of (Table 7) and (Table 8)we can conclude that using the random coefficient would explain the community variation better than the model with fixed coefficients.

Table 8 Estimate of Random Coefficient Model

Variable	Coef.	Std.Er.	Z-value	P-value	OR	[95% Conf. Upper	Interval] Lower
Fixed effect							
Residence							
Urban (ref)							
Rural	0.055	0.055	1.02	0.310	1.056	-0.051	0.160
Malaria							
Yes (ref)							
No	0.061	0.050	-1.23	0.217	0.941	-0.058	0.036
ChildSex							
Girl (ref)							
Boy	0.016	0.027	-0.58	0.563	0.985	-0.068	0.037
Inflammention							
Moderate							
Mild	0.124	0.035	3.53	0.001	1.132	0.055	0.193
Severe	1.225	0.040	5.65	0.001	1.253	0.147	0.304
Wealth Index							
Poor							
Poorer	-0.032	0.045	-0.70	0.482	0.969	-0.120	0.057
Middle	-0.049	0.044	-1.11	0.266	0.952	-0.136	0.038
Richer	-0.016	0.046	-0.34	0.732	0.984	-0.105	0.074
Richest	-1.105	0.054	-1.96	0.050	1.111	-1.210	-0.001
HeadSex							
Male (ref)							
Female	0.007	0.032	-0.21	0.837	0.993	-0.070	0.057
Mother Alive							
No (ref)							
Yes	0.033	0.062	0.54	0.589	1.034	-0.088	0.154
Father Alive							
No (ref)							
Yes	0.080	0.053	1.50	0.133	1.083	-0.024	0.184
Don't Know	0.419	0.273	1.53	0.125	1.520	-0.115	0.953
Age of Child							
5-10(ref)							
11-14	0.069	0.029	-2.40	0.016	1.071	-0.125	-0.013
Vitamin A							
Yes (ref)							
No	0.120	0.099	-1.21	0.226	1.128	-0.314	0.074
Water source							
Improved (ref)							
Unimproved	0.012	0.040	1.29	0.773	1.012	-0.067	0.090
β_0	0.225	0.148	1.52	0.129		-0.066	0.515
Random effect							
$\sigma_0^2 = var(U_0 j)$	0.102	0.019				0.071	0.146
$\sigma_{12}^2 = cov(U_1 j, U_2 j)$	0.363	0.010				0.344	0.382
Community variance	0.072	6.831				1.83e	2.84e
District variance	0.721	6.830				1.87e	2.79e
ICC(ρ)	0.073						

3.3.4 Multilevel Model Comparison

In order to analyze data which is in the hierarchical form requires the technique of the best model to choose that can explain the variability of the data. From Table 9, based on the chi-square random intercept it has shown that the $(\chi^2 = 17.18)$ with p-value of 0.001 was significant for the null model, for the random intercept model $(\chi^2 = 18.36)$ and p-value of 0.001 and for the random coefficient model $(\chi^2 = 11.60)$ and p-value of

0.001 showed that models were significant and allowing multilevel models to be applied in analyzing hierarchical data that helped to check the impact of the within community variations and between community variations on anemia in school-aged children of 5-14 years old.

Comparing AIC and BIC among the null model, the random coefficient model showed less AIC of 730.53 implying that it was a better model as compared to the random intercept model and the null model. The variation of anemia in school-aged children among communities were accounted in all the models described.

Table 9 Model Comparison

Fitted Model	Null Model	Random Intercept	Rondom Coefficient
-2*log likelihood	-382.85	-348.66	-346.27
Chi-square	17.18	13.95	14.15
Degree of freedom	1	1	1
P-value	0.001	0.001	0.001
AIC	769.69	733.31	730.53
BIC	779.06	817.57	819.47
ICC	0.150	0.153	0.073

4 Discussion

The aim of the study was to model effects on anemia in school-aged children 5-14 years old in Malawi by employing multilevel logistic regression analysis. Anemia in the study was determined by individual-level, community-level and district-level factors. All these were supported by the observed heterogeneity in odds on anemia between communities and between districts. The variables included in the study for individual-level and community-level analysis were place of residence (rural, urban), malaria in the past two weeks, child sex, wealth index (poor, poorer, middle, richer, richest), head sex, if mother is alive , if father is alive, if received vitamin A, inflammation (moderate, mild, severe) and type of water source (improved and unimproved). Variables Wealth index (richest), inflammation (mild and severe) and age of child were found to be significant in both individual-level and community level models. These results also seems to agree with findings of several studies conducted in Ethiopia and the world on under-five children by [20] and [21].

A bivariate analysis between response variable and each of the predictor variable was fitted into the data to check for association between response variable and predictor variables, while a multilevel model was employed to estimate the quantification of between-community variation (ICC). Since variables can be measured at different levels of the hierarchy, it allowed for correct inferences about community-level variables to be made and additionally, the magnitude of the association between variables and outcome varied between communities which was something that could not easily be handled by simple logistic regression.

The random intercept was fitted to allow intercepts to vary while retaining common slopes across communities, and random coefficient model was fitted to allow both the intercepts and slopes to vary across communities. Lastly, model diagnostic was tested to check for the better fit model by either using BIC or AIC.

The 2015-16 Malawi Demographic Health Survey (MDHS) and Micronutrient Survey (MNS) was based on two-stage stratified sampling. The first stage was selecting a sample frame from the Malawi Population and Housing Census (PHC). The second stage was selection of households from sampled communities from a sample frame. Based on the result of this study, less proportion of anemia cases were observed among school-aged children with moderate inflammation. As a result school-aged children who had mild and severe anemia were more than 100% more likely to experience anemia as compared to school-aged children with moderate inflammation. These results were linked to a study conducted by [6] on determinants of anemia among school-aged children in Mexico, the United States and Colombia. These studies mentioned were different with this study which aimed at finding the effect on anemia in school-aged children 5-14 years by using multilevel regression models.

The logistic regression analysis and multilevel logistic regression analysis in this study showed that age had a significant effect on anemia for the SAC 5-14 years at 5% level of significance. The odds of SAC 11-14 years were 0.6 times less likely to develop anemia as compared to SAC 5-10 years old. The odds of multilevel regression analysis considering heterogeneity for SAC who resided in urban areas were 2 times less likely to become anemic as compared to school- aged children residing from rural areas. Likewise, the odds for school-aged children who experienced no malaria in past two weeks decreased by 34% as compared to those who experienced malaria in the past two weeks. Applying for multilevel regression analysis in the data, we found that there were community-level clustering and district-level clustering of ICC = 15% and ICC = 12% respectively Table 3. With these results therefore, anemia interventions need not to be community-specific, otherwise effects on anemia may decrease in one area living other areas unattended. Identifying a successful program would include a focus on overall, not focusing only on specific groups (individuals) but covering variations between communities or districts.

The results from this study will not be compared more directly with previous studies because this study is one of the few anemia studies conducted in Malawi and one of the first study to look into effects on anemia in school-aged children of 5-14 years using the multilevel modelling. The previous studies interpreted of high prevalence of anemia in an area/community while this study separated the contribution of individual characteristics to effect on anemia from the contribution of the community which is the real clustering effects. A similar study conducted in Mixico and Colombia on determinants of anemia in SAC by [6] showed similar results of evidence of clustering but at household level.

With a newly available hierarchical regression techniques, it is possible to separate these individual effects from contextual effects and therefore, to give accurate measure between community variation [22]. More multilevel studies are needed to determine

whether relatively clustering effects targeting the age group of school-aged children 5-14 years applies to other countries. Fewer data are available on the prevalence of anemia in growing children of age 5-14 years

In order to look for a model that would fit the data to be used in the study the variance component model was applied using level-2 model to check if grouping variable at level-2 significantly affected the intercept of the dependent variable at level-1. A significant ICC from the result proved that the level-2 clustering was significant with non zero ICC and therefore multilevel modeling was applied.

It is important to note that this was the first study to use multilevel logistic regression to estimate the community variation and compare the variance component model, random intercept model and random coefficient model that would better fit the hierarchical data to predict the effect on anemia in SAC 5-14 years. From the results the best model to fit the hierarchical anemia data were both the random coefficient model and random effect model since their variation did not differ much with less AIC of 730.53 and 733.31 as compared variance component model with 769.69. Otherwise if we choose among the two we say that the best model using AIC is the random coefficient model AIC = 730.53. (Table 9).

More studies are therefore needed to shed further light on the usefulness of using multilevel analysis on anemia in SAC 5-14 years in identifying between community effects.

5 Conclusion

Multilevel binary logistic regression examined the effects of the explanatory variables at the different levels simultaneously. It produced more accurate estimates of regression coefficients, standard errors, and significant test as compared with single-level logistic regression. The effects on anemia in school-aged children were found by factors evolving at individual-level (level-1) and community-levels (level-2). There was a weak negative correlations between individual and community variations. The results showed that ignoring the hierarchical nature of the data could result in over estimating the significance of some of the variables included in the model.

Using the standard logistic regression would not explain any variation of within and between cluster in the analysis. The only appropriate approach to analyze the anemia data from this study was therefore, based on nested sources of variability, where the units at lower-level (level-1) were the individual school-aged children of age 5-14 years old who were nested within units at higher-level (level-2) community-level where variability between community were observed. Having the response variable anemia and being binary it proved that multilevel logistic regression modeling was a natural choice of modeling. The multilevel random coefficient was better as compared to random intercept and the variance component in fitting the data.

6 Recommendation

This study recommended that researchers who want to use the Malawi Demographic Health Survey data should use multilevel models rather than ordinary regression models. This was so because of the nature of the hierarchy of data. The study had suggested that effort should be made by Government to reduce anemia in school-aged children in the communities. The study was very important in that it provided the better fit model in analyzing effect on anemia in school-aged children of 5-14 years old. The study also showed that there were community variations on effect on anemia in SAC. Therefore, further studies should be conducted to incorporate the spatial variations in the effect on anemia by utilizing spatial models and the geo-additive models to the community effects on anemia in school-aged children 5-14 years old in Malawi.

Declaration

Ethics approval and consent to participate

The Malawi Health Research Committee determined that ethical approval was not deemed necessary in this study considering the fact that the study used data from a research study already approved by an ethical research committee. The MDHS and MNS study were ethically approved by Malawi Health Research Committee, Institutional Review Board of ICF Macro, Centre for Disease and Control (CDC) in Atlanta, GA, USA and Prevention IRB. Additionally, informed consent and anonymity do not apply to our study as these were done before and after data collection by the responsible implementing institution.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The dataset used in this study are available from the DHS website https://dhsprogram.com/Data/ upon request from the MEASURE DHS program team.

Acknowledgements

We thank all individuals who contributed in this study

Author details

[1]National Statistical Office, Chimbiya, 333 Zomba, Malawi. [2]Department of Mathematical Sciences, University of Malawi, Chancellor College, Chirunga, 280 Zomba, Malawi.

References

1. Simbauranga, R.H., Kamugisha, E., Hokororo, A., Kidenya, B.R., Makani, J.: Prevalence and factors associated with severe anaemia amongst under-five children hospitalized at Bugando Medical Centre, Mwanza, Tanzania. BMC hematology **15**, 13 (2015). doi:10.1186/s12878-015-0033-5
2. Kawo, K.N., Asfaw, Z.G., Yohannes, N.: Multilevel Analysis of Determinants of Anemia Prevalence among Children Aged 6-59 Months in Ethiopia: Classical and Bayesian Approaches (2018). doi:10.1155/2018/3087354
3. Gr, R.M., Ramachandrappa, H., Rudramurthy, P., Gopalakrishna, D.V., Rao, S.S., Bharath, K.R., Reddy, K., Bharath, K., Reddy, K.R.: RISK FACTORS FOR IRON DEFICIENCY ANEMIA AMONG SCHOOL GOING CHILDREN IN URBAN SOUTH BANGALORE, INDIA. Technical report (2014)
4. Abu-Ouf, N.M., Jan, M.M.: The impact of maternal iron deficiency and iron deficiency anemia on child's health. Saudi medical journal **36**(2), 146–149 (2015). doi:10.15537/smj.2015.2.10289
5. MacLean, C., Newberry, S., Maglione, M., McMahon, M., Ranganath, V., Suttorp, M., Mojica, W., Timmer, M., Alexander, A., McNamara, M., Desai, S.B., Zhou, A., Chen, S., Carter, J., Tringale, C., Valentine, D., Johnsen, B., Grossman, J.: Systematic review: comparative effectiveness of treatments to prevent fractures in men and women with low bone density or osteoporosis. Annals of internal medicine **148**(3), 197–213 (2008). doi:10.7326/0003-4819-148-3-200802050-00198
6. Syed, S., Addo, O.Y., De La Cruz-Góngora, V., Ashour, F.A.S., Ziegler, T.R., Suchdev, P.S.: Determinants of anemia among school-aged children in Mexico, the United States and Colombia. Nutrients **8**(7), 1–15 (2016). doi:10.3390/nu8070387
7. Macdonald, C., Mildon, A., Neequaye, M., Namarika, R., Yiannakis, M.: Anemia - can its widespread prevalence among women in developing countries be impacted ? Anemia in Women : A Global Health Priority (October 2002), 1–42 (2010)
8. National Statistical Office (NSO), Community Health Sciences Unit (CHSU) [Malawi], Centers for Disease Control and Prevention (CDC), Emory, U.: Malawi Micronutrient Survey 2015-16. Atlanta GA, USA NSO, CHSU, CDC Univ Emory (December) (2017)

9. Barth, J.H., Luvai, A., Jassam, N., Mbagaya, W., Kilpatrick, E.S., Narayanan, D., Spoors, S.: Comparison of method-related reference intervals for thyroid hormones : studies from a prospective reference population and a literature review **55**(1), 107–112 (2018). doi:10.1177/0004563217691549
10. Ismail Dragon Legason, A.A.: Prevalence of anemia and associated risk factors among children in north western uganda **17**(10) (2017). Part of Springer Nature.
11. Gebremeskel, M.G., Mulugeta, A., Bekele, A., Lemma, L., Gebremichael, M., Gebremedhin, H., Etsay, B., Tsegay, T., Haileslasie, Y., Kinfe, Y., Gebremeskel, F., Mezgebo, L., Shushay, S.: Individual and community level factors associated with anemia among children 6—59 months of age in Ethiopia: A further analysis of 2016 Ethiopia demographic and health survey. PLoS ONE **15**(11 November), 1–17 (2020). doi:10.1371/journal.pone.0241720
12. Ntenda Morton., F.N.T. Kun-Yang Chung., Chuang, Y.-c.: Multilevel Analysis of the Effects of Individual and Community - Level Factors on Childhood Anemia. Oxford University Press **12** (2017). doi:10.1093/tropej/fmx059
13. Raykov, T., Marcoulides, G.A., Akaeze, H.O.: Comparing Between- and Within-Group Variances in a Two-Level Study: A Latent Variable Modeling Approach to Evaluating Their Relationship. Educational and Psychological Measurement **77**(2), 351–361 (2017). doi:10.1177/0013164416634166
14. Djokic, D.M.R.Z.C.L. D.: Risk factors associated with anemia among serbian school age children 7 - 14 years old: Results of the first national health survey **3** (1999)
15. Richard, L.: Computing Intraclass Correlations (ICC) as Estimates of Interrater Reliability in SPSS. The Winnower (Icc), 1–4 (2015). doi:10.15200/winn.143518.81744
16. Merlo, J., Wagner, P., Austin, P.C., Subramanian, S.V., Leckie, G.: General and specific contextual effects in multilevel regression analyses and their paradoxical relationship: A conceptual tutorial. SSM - Population Health **5**(May), 33–37 (2018). doi:10.1016/j.ssmph.2018.05.006
17. Sanagou, M., Wolfe, R., Forbes, A., Reid, C.M.: Hospital-level associations with 30-day patient mortality after cardiac surgery: A tutorial on the application and interpretation of marginal and multilevel logistic regression. BMC Medical Research Methodology **12** (2012). doi:10.1186/1471-2288-12-28
18. número 442, A.: No Title . , 32 (2012)
19. Merlo, J., Wagner, P., Ghith, N., Leckie, G.: An original stepwise multilevel logistic regression analysis of discriminatory accuracy: The case of neighbourhoods and health. PLoS ONE **11**(4), 1–31 (2016). doi:10.1371/journal.pone.0153778
20. Sommet, N., Morselli, D.: Keep calm and learn multilevel logistic modeling: A simplified three-step procedure using stata, R, Mplus, and SPSS. International Review of Social Psychology **30**(1), 203–218 (2017). doi:10.5334/irsp.90
21. Tezera, R., Sahile, Z., Yilma, D., Misganaw, E., Mulu, E.: Prevalence of anemia among school-age children in Ethiopia: A systematic review and meta-analysis. Systematic Reviews **7**(1), 1–7 (2018). doi:10.1186/s13643-018-0741-6
22. Vugutsa Luvai, L., Mohamed Mohamed El-Sayed, A., Goldstein, H., Rasbash, J., Example, A.S., Aa, M.A., Naing, N.N., Sayers, A., Heron, J., Smith, A.D.A.C., Macdonald-Wallis, C., Gilthorpe, M.S., Steele, F., Tilling, K., Lee, J.Y.L., Green, P.J., Ryan, L.M., Livingston, M.: Multi-level Models Powerpoint. Technical Report 1 (2017). 1709.06288. http://arxiv.org/abs/1709.06288{%}0Ahttp://dx.doi.org/10.1016/S2212-5671(12)00088-3{%}0Apapers3://publication/uuid/DF4927A2-0C76-4B8F-AFB7-F7828DF6B363